

Машинное обучение
часть 1

А.М.Миронов

Московский Государственный Университет
Механико-математический факультет
Кафедра математической теории интеллектуальных систем

Введение

Книга представляет собой введение в основные понятия, методы и алгоритмы машинного обучения, которое находится в настоящее время в состоянии исключительно бурного развития и является теоретической основой для проектирования интеллектуальных систем обработки больших данных.

В первой части книги излагаются элементарные аспекты машинного обучения:

- виды задач и моделей машинного обучения,
- простейшие алгоритмы обучения для линейно разделимых обучающих выборок,
- методы градиентного спуска и его разновидности,
- метод обучения нейронных сетей,
- метод опорных векторов,
- ядерные методы машинного обучения,
- регрессионный анализ,
- метрические и вероятностные модели машинного обучения.

Во второй части будут рассмотрены более глубокие вопросы машинного обучения, в частности, методы прогнозирования индивидуальных последовательностей, сравнительная теория машинного обучения, бустинг, алгоритмы экспоненциального смешивания, агрегирующие алгоритмы, методы теории игр в машинном обучении. В третьей части будут изложены модели и методы автоматного и процессного обучения (automata learning и process mining).

Просьба сообщать автору о всех замеченных недостатках и рекомендациях по улучшению данного текста по адресу amironov66@gmail.com

Оглавление

1	Задачи и модели машинного обучения	4
1.1	Задачи машинного обучения	4
1.1.1	Предмет машинного обучения	4
1.1.2	Обучение функциональным зависимостям	4
1.1.3	Способ описания объектов	5
1.1.4	Пример задачи машинного обучения	6
1.1.5	Виды задач машинного обучения	7
	Задачи классификации	7
	Задачи регрессии	8
	Задачи ранжирования	8
1.2	Модели машинного обучения	8
1.2.1	Понятие модели обучения	8
1.2.2	Примеры предсказательных моделей	9
	Линейная предсказательная модель	9
	Предсказательная модель в виде нейронной сети	9
1.2.3	Алгоритмы обучения	12
1.3	Проблема переобучения	14
1.4	Основные этапы решения задач машинного обучения	15
2	Элементарные методы и алгоритмы машинного обучения	16
2.1	Линейно разделимые выборки	16
2.2	Алгоритм обучения Розенблатта	19
2.2.1	Описание алгоритма обучения Розенблатта	20
2.2.2	Завершаемость алгоритма Розенблатта	21
2.2.3	Модификации алгоритма Розенблатта	23
2.3	Метод градиентного спуска	23
2.3.1	Понятие градиентного спуска	23
2.3.2	Описание метода градиентного спуска	24
2.3.3	Модификации метода градиентного спуска	27
	Метод стохастического градиента	27
	Регуляризация	28

2.4	Метод обратного распространения ошибки для обучения нейронных сетей	29
2.4.1	Идея метода	29
2.4.2	Описание метода	30
2.4.3	Достоинства и недостатки метода	32
2.5	Метод опорных векторов	33
2.5.1	Оптимальность аппроксимирующих функций	33
2.5.2	Построение оптимальной разделяющей гиперплоскости для строго линейно разделимой выборки	35
	Описание задачи	35
	Метод решения оптимизационной задачи	37
	Применение метода	44
2.5.3	Построение оптимальной разделяющей гиперплоскости по зашумленной выборке	47
2.6	Ядерный метод машинного обучения	51
2.6.1	Спрямяющие пространства	51
2.6.2	Примеры ядер	55
2.6.3	Каноническое гильбертово пространство, определяемое ядром	57
2.7	Задача регрессии	60
2.7.1	Линейная регрессия	60
2.8	Метрическая модель обучения	63
2.8.1	Понятие метрики	63
2.8.2	Метод ближайших соседей	64
2.8.3	Метод окна Парзена	65
2.8.4	Метод потенциалов	66
2.8.5	Метод эталонов	67
2.9	Вероятностные модели обучения	68
2.9.1	Дискретная вероятностная модель обучения	68
2.9.2	Оптимальные аппроксимирующие функции	69
2.9.3	Построение АФ по обучающей выборке	72
2.9.4	Непрерывная вероятностная модель обучения	72
2.9.5	EM-алгоритм	76
	Число k компонентов смеси известно	77
	Число компонентов смеси неизвестно	79

Глава 1

Задачи и модели машинного обучения

1.1 Задачи машинного обучения

1.1.1 Предмет машинного обучения

Машинное обучение (Machine Learning, ML) – это раздел теории искусственного интеллекта, предметом которого является поиск методов решения задач путем обучения в процессе решения сходных задач.

Для построения таких методов используются средства алгебры, математической статистики, дискретной математики, теории оптимизации, численных методов, и других разделов математики.

1.1.2 Обучение функциональным зависимостям

Одно из направлений ML связано с задачами следующего вида: имеются

- множество X **объектов**, и
- множество Y **ответов**.

Предполагается, что существует функциональная зависимость

$$f : X \rightarrow Y$$

между объектами и ответами, но она неизвестна. Известна лишь совокупность S пар вида (объект, ответ), называемая **обучающей выборкой (training sample)**:

$$S = \{(x_i, y_{x_i} = f(x_i)) \in X \times Y \mid i = 1, \dots, l\}.$$

Требуется найти приближенный вид этой f путем построения **аппроксимирующей функции (АФ)** $a_S : X \rightarrow Y$ (**decision function**), такой, что

$$\forall x \in X \quad a_S(x) \approx f(x).$$

1.1.3 Способ описания объектов

Объекты в ML могут иметь самую различную природу: это могут быть люди, животные, растения, страны, организации, сайты, столы, стулья, изображения, фильмы, и т.д.

Один из способов описания объектов в форме, пригодной для решения описанных выше задач ML имеет следующий вид:

- задается множество \mathfrak{S} **признаков объектов (features)**,
- каждому признаку $i \in \mathfrak{S}$ сопоставляется множество D_i **значений** этого признака, и
- каждому объекту $x \in X$ и каждому признаку $i \in \mathfrak{S}$ сопоставляется **значение** x^i i -го признака на объекте x .

Множество признаков объектов определяется решаемыми задачами ML. Адекватный выбор этого множества для каждой конкретной задачи ML – нетривиальная проблема, от правильного решения которой существенно зависит успех в решении этой задачи ML.

Каждый признак $i \in \mathfrak{S}$ имеет определенный **тип**. Ниже мы перечисляем некоторые из таких типов и указываем соответствующие им множества значений D_i :

- **бинарный**: $D_i = \{-1, 1\}$ или $\{0, 1\}$,
- **номинальный**: D_i – конечное множество,
- **порядковый**: D_i – конечное линейно упорядоченное множество,
- **количественный**: $D_i = \mathbf{R}$ (множество действительных чисел).

Для удобства мы будем предполагать, что

- множество \mathfrak{S} признаков является конечным, его элементам сопоставлены натуральные числа $1, \dots, n$, которые мы будем отождествлять с соответствующими им признаками, и
- значениями каждого признака являются действительные числа.

Описание объекта $x \in X$ относительно множества $\{1, \dots, n\}$ признаков – это кортеж (x^1, \dots, x^n) значений каждого из признаков $1, \dots, n$. Т.к. по предположению $\forall i = 1, \dots, n \quad x^i \in \mathbf{R}$, то описание объекта x можно рассматривать как элемент векторного пространства \mathbf{R}^n .

Напомним, что \mathbf{R}^n состоит из всех **векторов размерности** n , т.е. последовательностей вида $(\alpha_1, \dots, \alpha_n)$, где $\forall i = 1, \dots, n \quad \alpha_i \in \mathbf{R}$, операции сложения на данных последовательностях и умножения их на действительные числа определяются стандартным образом:

$$\begin{aligned} (\alpha_1, \dots, \alpha_n) + (\alpha'_1, \dots, \alpha'_n) &\stackrel{\text{def}}{=} (\alpha_1 + \alpha'_1, \dots, \alpha_n + \alpha'_n), \\ \alpha(\alpha_1, \dots, \alpha_n) &= (\alpha_1, \dots, \alpha_n)\alpha \stackrel{\text{def}}{=} (\alpha\alpha_1, \dots, \alpha\alpha_n). \end{aligned}$$

1.1.4 Пример задачи машинного обучения

В этом пункте мы приведем пример одной задачи МЛ и связанного с ней описания объектов. Данная задача называется **задачей кредитного скоринга**. Объектами в ней являются заявки на выдачу кредита в некотором банке, а ответами – элементы множества $\{-1, 1\}$. Для каждого объекта x ответ $f(x)$ интерпретируется как решение банка:

- ответ «1»: выдать кредит по заявке x ,
- ответ «-1»: отказать в удовлетворении заявки x .

Решение банка по удовлетворению заявки является правильным, если

- клиент, подавший эту заявку, и получивший кредит, вернет его в должный срок, или
- клиент, которому было отказано в удовлетворении заявки x , в случае получения этого кредита с высокой вероятностью может не вернуть его в должный срок.

Известно множество $S = \{(x_i, y_{x_i}) \mid i = 1, \dots, l\}$ где $\forall i = 1, \dots, l$

- x_i – некоторая заявка на выдачу кредита,
- y_{x_i} – решение банка по этой заявке, которое является правильным.

Требуется обучиться функции a_S , вычисляющей по каждой новой заявке правильное решение. Данную задачу можно решать методами МЛ, в которой S рассматривается как обучающая выборка.

Каждый объект в этой задаче связан с некоторым клиентом, и признаки объекта могут иметь вид различных характеристик этого клиента. Например, могут быть выбраны следующие признаки:

- бинарные: пол, наличие постоянной работы, и т.д.
- номинальные: место проживания, профессия, работодатель и т.д.
- порядковые: образование, должность и т.д.
- количественные: возраст, зарплата, стаж работы, доход семьи, сумма кредита и т.д.

Описанная выше задача МЛ может допускать различные обобщения. Например, можно обучаться не только нахождению правильных решений банка, но и например оценке вероятности дефолта клиента (т.е. возникновения ситуации, когда клиент, получивший кредит, не вернет его в должный срок).

1.1.5 Виды задач машинного обучения

Задачи классификации

Изложенная в предыдущем пункте задача МЛ является примером задач классификации. В задачах данного вида

- множество ответов Y является конечным,
- каждый ответ $y \in Y$ соответствует некоторому классу объектов, и
- задача МЛ заключается в вычислении по каждому объекту соответствующего ему класса.

В описанной в предыдущем пункте задаче рассматривается классификация с двумя классами объектов. Другой пример задач классификации с двумя классами: объектами являются посетители поликлиники, требуется по признакам посетителя дать ответ: болен он, или нет.

Могут быть задачи классификации с числом классов более двух, например

- задача распознавания рукописных букв русского языка, в данном случае $Y = \{a, \dots, я\}$,
- пусть M – некоторое множество болезней, X – множество пациентов поликлиники, для которых требуется диагностировать набор болезней, которыми они болеют, в данном случае классы соответствуют наборам болезней: $Y = \mathcal{P}(M)$.

Задачи регрессии

В задачах регрессии множество ответов Y имеет вид \mathbf{R} или \mathbf{R}^m . Задачи данного типа как правило связаны с прогнозированием (например, курса доллара, или курсов нескольких валют).

Задачи ранжирования

Задачи ранжирования возникают например в системах информационного поиска, или в рекомендательных системах. В данных задачах требуется по каждому объекту $x \in X$ определить его приоритет в выдаче в качестве результата на некоторый запрос. Множество Y ответов здесь является конечным упорядоченным множеством, $\forall x \in X$ значение $f(x)$ представляет собой соответствующий приоритет.

1.2 Модели машинного обучения

1.2.1 Понятие модели обучения

Как правило, для решения задачи построения функции $a_S : X \rightarrow Y$ по обучающей выборке S выбирается некоторая **модель обучения**, состоящая из двух компонентов:

1. Первой компонентой модели обучения является функция

$$a : X \times W \rightarrow Y \quad (1.1)$$

где W – множество, элементы которого называются **параметрами**.

Искомая функция a_S ищется в виде

$$a_S(x) = a(x, w), \quad (1.2)$$

где w – фиксированный параметр. Функцию (1.1) иногда называют **предсказательной моделью (predictive model)**.

2. Другой компонентой модели обучения является **алгоритм обучения**, который представляет собой алгоритм поиска такого значения w , для которого функция a_S , определяемая соотношением (1.2), обладает некоторыми свойствами оптимальности. Более подробно об свойствах оптимальности функции a_S см. в пункте 1.2.3.

1.2.2 Примеры предсказательных моделей

Линейная предсказательная модель

В **линейной предсказательной модели** множество W параметров имеет вид \mathbf{R}^n , где n – число признаков объектов, т.е. каждый параметр w представляет собой вектор действительных чисел $w = (w_1, \dots, w_n)$, и

- в задачах регрессии и ранжирования $Y = \mathbf{R}$, и

$$a(x, w) = \langle x, w \rangle \stackrel{\text{def}}{=} \sum_{i=1}^n x^i w_i$$

(число $\langle x, w \rangle$ называется **скалярным произведением** x и w),

- в задачах классификации $Y = \{-1, 1\}$, и

$$a(x, w) = \text{sign}(\langle x, w \rangle),$$

где sign – **функция знака**, она сопоставляет неотрицательным числам значение 1, а отрицательным – значение -1 .

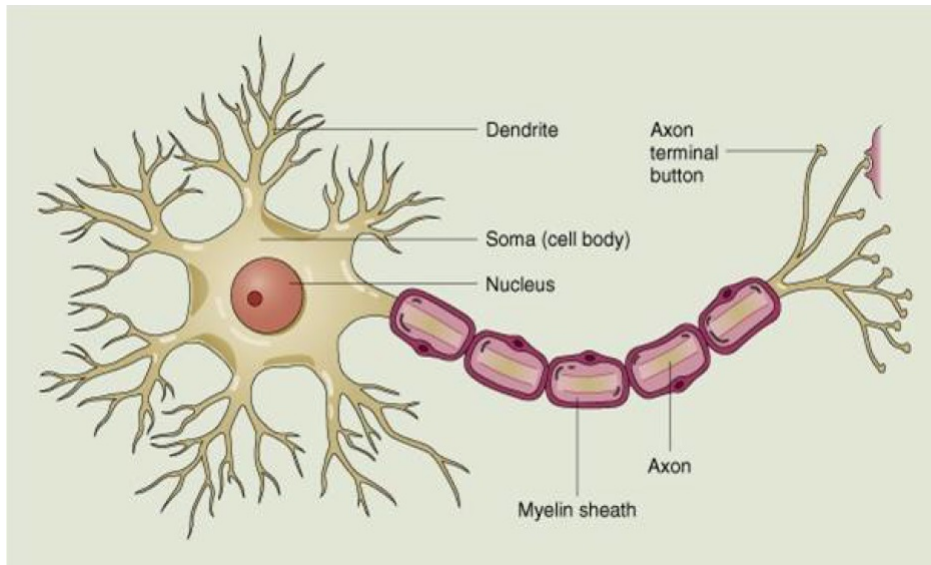
Если какая-либо задача ML не решается в некоторой линейной модели, то для решения этой задачи можно попытаться расширить исходную линейную модель путем добавления новых признаков, получаемых из уже имеющихся признаков. Например, можно добавлять

- комбинации признаков (их произведение, и т.п.),
- функции от признаков, и т.д.

Предсказательная модель в виде нейронной сети

В современных технологиях машинного обучения очень популярны предсказательные модели, основанные на нейронных сетях.

Данные модели используют преобразователи информации, являющиеся аналогами биологических нейронов. Структура биологического нейрона изображена на следующей картинке:



Преобразование информации в нейроне происходит в его центральной части (называемой **телом**), от которой отходят отростки двух типов:

- **дендриты**, по ним поступают входные сигналы, будем считать дендриты занумерованными натуральными числами $1, \dots, n$,
- **аксон**, отросток такого типа ровно один, по нему проходит выходной сигнал.

Схема работы нейрона имеет следующий вид: с каждым из дендритов связано некоторое неотрицательное действительное число, называемое **весовым коэффициентом** (или просто **весом**). Обозначим записью w_1, \dots, w_n список весов, соответствующих каждому из дендритов ($\forall i = 1, \dots, n$ дендриту номер i соответствует вес w_i). Кроме того, с нейроном связано число $w_0 \geq 0$, называемое **порогом возбуждения**.

Сигналы, поступающие в нейрон по дендритам, являются электрическими импульсами различной интенсивности. Обозначим записью x^1, \dots, x^n список интенсивностей сигналов, поступивших в текущий момент по каждому из дендритов ($\forall i = 1, \dots, n$ x^i – интенсивность сигнала, поступившего на нейрон номер i). Данные сигналы вызывают в центральной части нейрона электрический импульс интенсивности $\sum_{i=1}^n x^i w_i$, и

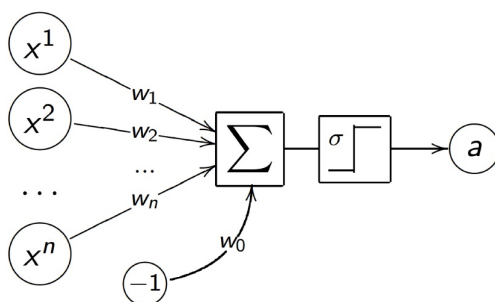
- если интенсивность этого импульса превышает порог возбуждения нейрона w_0 , то нейрон выпускает по аксону выходной сигнал некоторой интенсивности,

- иначе по аксону выходной сигнал не выпускается (мы будем считать, что в этом случае по аксону выпускается выходной сигнал нулевой интенсивности).

Нейрон можно рассматривать как преобразователь числовой информации: на его вход поступает кортеж действительных чисел (x^1, \dots, x^n) , на выход он выдает число a , определяемое соотношением

$$a = \sigma\left(\sum_{i=1}^n x^i w_i - w_0\right)$$

где σ – функция (называемая **функцией активации**), сопоставляющая неотрицательным числам значение 1, и отрицательным числам – значение 0. Функционирование нейрона изображается диаграммой



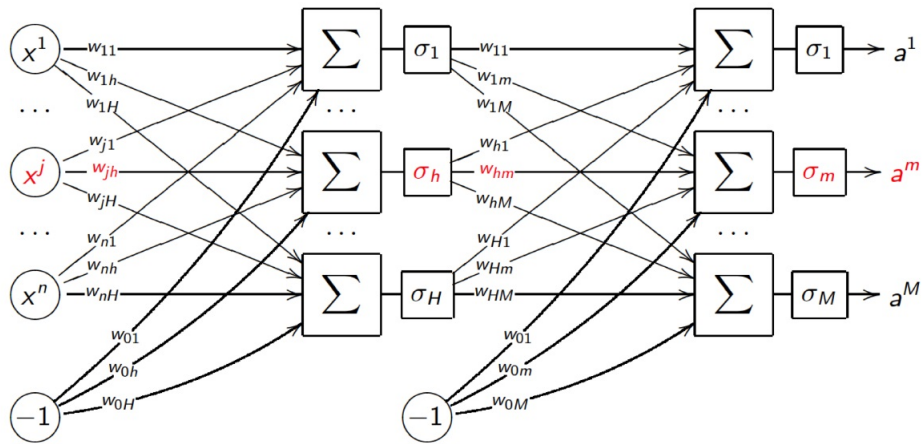
Ниже мы будем называть нейронами не только биологические нейроны, но также и произвольные преобразователи числовой информации, работающие по описанному выше принципу.

Функция активации (σ) в нейронах может иметь не только описанный выше вид, но также и другой вид, например:

- **сигмоида**: $\sigma(x) = \frac{1}{1+e^{-x}}$, или $\text{th}(x)$,
- **ReLU**: $\sigma(x) = 0$ при $x < 0$, и $\sigma(x) = x$ при $x \geq 0$.

Нейроны можно объединять в более сложные преобразователи числовой информации, называемые **многослойными нейронными сетями**. В этих сетях сигналы, выдаваемые на выходах одних нейронов поступают на входы других нейронов. При этом допускается, что один и тот же сигнал с выхода какого-либо нейрона может параллельно подаваться на входы нескольких нейронов.

Приводимая ниже диаграмма является схематическим изображением двуслойной нейронной сети:



Понятие многослойной нейронной сети является основой для методов **глубокого обучения (deep learning)**, которые являются предметом большого количества теоретических и прикладных исследований и коммерческих разработок, но в данном курсе рассматриваться не будут.

1.2.3 Алгоритмы обучения

Алгоритм обучения (learning algorithm) представляет собой алгоритм нахождения по обучающей выборке S такой АФ $a_S : X \rightarrow Y$, которая обладает описываемыми ниже **свойствами оптимальности**. Все эти свойства оптимальности являются детализацией следующего требования: a_S должно как можно лучше приближать исходную неизвестную функцию $f : X \rightarrow Y$ на всем X .

Для точного описания свойств оптимальности алгоритмов обучения используется понятие **функции потерь (loss function)**, которая сопоставляет паре (a_S, x) , где $x \in X$, число $\mathcal{L}(a_S, x)$, выражающее величину ошибки аппроксимации a_S на объекте $x \in X$.

Приведем некоторые примеры функций потерь.

- $\mathcal{L}(a_S, x) = \llbracket a_S(x) \neq f(x) \rrbracket$ (для задач классификации), где для каждого утверждения β запись $\llbracket \beta \rrbracket$ обозначает
 - значение 1, если утверждение β истинно, и
 - значение 0, если утверждение β ложно,
- $\mathcal{L}(a_S, x) = |a_S(x) - f(x)|$ или $(a_S(x) - f(x))^2$ (для задач регрессии).
- Пусть a_S имеет вид $sign(\langle x, w \rangle)$.

Обозначим записью $M_x(w)$ число $\langle x, w \rangle f(x)$ (эта величина называется отступом (margin)). $\mathcal{L}(a_S, x)$ может иметь вид

$$\llbracket M_x(w) < 0 \rrbracket, (1 - M_x(w))^2, e^{-M_x(w)}, \frac{2}{1 + e^{M_x(w)}}, \log_2(1 + e^{-M_x(w)}).$$

Если $S' = \{(x'_i, y'_i) \mid i = 1, \dots, l'\}$ – какая-либо обучающая выборка, соответствующая той же исходной функции $f : X \rightarrow Y$, что и S , то запись $\mathcal{L}(a_S, S')$ обозначает число

$$\frac{1}{l'} \sum_{i=1}^{l'} \mathcal{L}(a_S, x'_i).$$

В описаниях свойств оптимальности алгоритмов обучения используется понятие **функционала эмпирического риска** (называемого ниже просто **риском**) аппроксимации a_S , который определяется как число

$$Q(a_S) = \mathcal{L}(a_S, S).$$

Если a_S имеет вид $a(x, w)$, где $a : X \times W \rightarrow Y$ и $w \in W$ (т.е. риск $Q(a_S)$ является функцией от w), то одно из свойств оптимальности алгоритма обучения по обучающей выборке S имеет следующий вид: значение параметра $w \in W$, определяющее наилучшую аппроксимацию a_S , должно удовлетворять соотношению

$$w = \arg \min_{w \in W} Q(a_S) \tag{1.3}$$

т.е. решение задачи МЛ сводится к оптимизационной задаче: требуется найти такой параметр $w \in W$, который минимизирует риск $Q(a_S)$.

Данная задача лучше всего решается в том случае, когда функция $\mathcal{L}(a(x, w), x)$ является дифференцируемой по w , т.к. в этом случае функция $Q(a_S)$ тоже является дифференцируемой по w , и для ее оптимизации можно применять простые методы: находить минимумы с помощью приравнивания к нулю частных производных, использовать методы градиентного спуска, и т.п.

Если \mathcal{L} разрывна, то для решения задачи оптимизации $Q(a(x, w))$ лучше всего аппроксимировать \mathcal{L} сверху какой-либо непрерывной функцией $\tilde{\mathcal{L}} \geq \mathcal{L}$, и использовать в выражении $Q(a(x, w))$ функцию $\tilde{\mathcal{L}}$ вместо функции \mathcal{L} .

1.3 Проблема переобучения

Переобучение – это чрезмерно точная подгонка АФ a_S под обучающую выборку S , которая дает сильные отклонения значений $a_S(x)$ от правильных значений (т.е. от $f(x)$) для многих объектов x , не входящих в обучающую выборку S .

Причины возникновения переобучения:

- излишние степени свободы в предсказательной модели $a(x, w)$, приводящие к учету при построении a_S различных шумов, неточностей и ошибок в данных,
- неполнота обучающей выборки S .

Переобучение можно обнаружить следующими способами.

1. Скользящий контроль (LOO, leave-one-out).

Пусть задана обучающая выборка $S = \{(x_i, y_{x_i}) \mid i = 1, \dots, l\}$.

$\forall i = 1, \dots, l$ обозначим записью $S - i$ выборку

$$\{(x_i, y_{x_i}) \mid i = 1, \dots, i - 1, i + 1, \dots, l\}.$$

Признаком переобучения является высокое значение выражения

$$\frac{1}{l} \sum_{i=1}^l \mathcal{L}(a_{S-i}, x_i)$$

Данный способ контроля переобучения можно представить в виде одного из условий оптимальности алгоритма обучения: данное условие имеет вид

$$\frac{1}{l} \sum_{i=1}^l \mathcal{L}(a_{S-i}, x_i) \rightarrow \min$$

2. Кросс-проверка (cross-validation).

Делается разбиение выборки на две части S_1 и S_2 , обучение идет по S_1 , а S_2 используется для проверки качества обучения.

Признаком переобучения является высокое значение выражения

$$Q(a_{S_1}, S_2).$$

Данный способ контроля переобучения тоже можно представить в виде одного из условий оптимальности алгоритма обучения: выбирается N различных разбиений обучающей выборки S на две части

$$\left(S_1^{(1)}, S_2^{(1)}\right), \dots, \left(S_1^{(N)}, S_2^{(N)}\right),$$

и одно из условий оптимальности алгоритма обучения имеет вид

$$\sum_{i=1}^N Q(a_{S_1^{(i)}}, S_2^{(i)}) \rightarrow \min$$

1.4 Основные этапы решения задач машинного обучения

Решение каждой задачи машинного обучения состоит из следующих основных этапов:

- адекватное понимание задачи и данных,
- предобработка данных и изобретение признаков и множеств значений каждого из этих признаков,
- построение предсказательной модели $a(x, w)$,
- сведение обучения к оптимизации,
- решение проблем оптимизации и переобучения,
- оценивание качества,
- внедрение и эксплуатация.

Глава 2

Элементарные методы и алгоритмы машинного обучения

2.1 Линейно разделимые выборки

Во многих задачах ML

- множества объектов и ответов имеют вид $X = \mathbf{R}^n$, $Y = \{-1, 1\}$, и
- задача обучения заключается в нахождении по выборке S АФ a_S вида

$$a_S(x) = \text{sign}\left(\sum_{i=1}^n x^i w_i - w_0\right), \quad \text{где } w_1, \dots, w_n, w_0 \in \mathbf{R}, \quad (2.1)$$

которая минимизирует риск $Q(a_S)$.

В некоторых случаях задача нахождения функции a_S вида (2.1) имеет наилучшее решение: можно найти такие значения w_1, \dots, w_n, w_0 , для которых $Q(a_S) = 0$.

Из определения $Q(a_S)$ следует, что равенство $Q(a_S) = 0$ равносильно свойству $\forall (x, y) \in S \ a_S(x) = y$, т.е. если функция a_S имеет вид (2.1), то

$$\forall ((x^1, \dots, x^n), y) \in S \quad \text{sign}\left(\sum_{i=1}^n x^i w_i - w_0\right) = y. \quad (2.2)$$

Выборка S , для которой существуют числа $w_1, \dots, w_n, w_0 \in \mathbf{R}$, удовлетворяющие условию (2.2), называется **линейно разделимой**.

Условие (2.2) можно переписать в виде: $\forall ((x^1, \dots, x^n), y) \in S$

$$\left(\sum_{i=1}^n x^i w_i - w_0\right) y \geq 0. \quad (2.3)$$

Если $\exists w_1, \dots, w_n, w_0 \in \mathbf{R}: \forall ((x^1, \dots, x^n), y) \in S$ неравенство (2.3) строгое, то выборка S называется **строго линейно разделимой**.

Понятие строгой линейной разделимости имеет геометрическую интерпретацию: если выборка $S \subseteq \mathbf{R}^n \times \{-1, 1\}$ строго разделима, т.е. $\exists w_1, \dots, w_n, w_0 \in \mathbf{R}: \forall ((x^1, \dots, x^n), y) \in S$

$$\left(\sum_{i=1}^n x^i w_i - w_0 \right) y > 0, \quad (2.4)$$

то гиперплоскость P , определяемая уравнением $\sum_{i=1}^n x^i w_i - w_0 = 0$, разделяет множества

$$S^+ \stackrel{\text{def}}{=} \{x \in \mathbf{R}^n \mid (x, 1) \in S\} \quad \text{и} \quad S^- \stackrel{\text{def}}{=} \{x \in \mathbf{R}^n \mid (x, -1) \in S\} \quad (2.5)$$

т.е. S^+ и S^- содержатся в разных полупространствах, на которые P делит пространство \mathbf{R}^n . Верно и обратное: если для выборки S можно найти гиперплоскость P , такую, что множества (2.5) содержатся в разных полупространствах, на которые P делит пространство \mathbf{R}^n , то S строго линейно разделима.

Напомним, что $\forall X \subseteq \mathbf{R}^n$

- множество X называется **выпуклым**, если $\forall x, x' \in X$

$$\{\alpha x + (1 - \alpha)x' \mid \alpha \in [0, 1]\} \subseteq X, \quad (2.6)$$

множество в левой части (2.6) называется **отрезком**, соединяющим точки x и x' , и обозначается записью $[xx']$,

- **выпуклой оболочкой** $Conv(X)$ множества X называется наименьшее (по включению) выпуклое множество, содержащее X ,
- $Conv(X)$ совпадает с пересечением совокупности всех выпуклых подмножеств \mathbf{R}^n , содержащих X ,
- если X – конечное множество и имеет вид $\{x_1, \dots, x_n\}$, то

$$Conv(X) = \sum_{i=1}^n \alpha_i x_i, \quad \text{где } \forall i = 1, \dots, n \quad \alpha_i \in \mathbf{R}, \alpha_i \geq 0, \sum_{i=1}^n \alpha_i = 1,$$

и, кроме того, $Conv(X)$ – компактное множество (т.к. оно замкнуто и ограничено).

$\forall x, x' \in \mathbf{R}^n$ запись $|xx'|$ обозначает длину отрезка $[xx']$, которую мы понимаем как евклидову норму $\|x - x'\|$ вектора $x - x'$.

Теорема 1.

Конечная выборка $S \subseteq \mathbf{R}^n \times \{-1, 1\}$ строго линейно делима тогда и только тогда, когда

$$\text{Conv}(S^+) \cap \text{Conv}(S^-) = \emptyset. \quad (2.7)$$

Доказательство.

Пусть выборка S строго линейно делима, т.е. существует гиперплоскость P , разделяющая S^+ и S^- . Обозначим записями L_1, L_2 полупространства, на которые P делит \mathbf{R}^n , и записями $\overset{\circ}{L}_1, \overset{\circ}{L}_2$ – внутренние части этих полупространств, т.е. $\overset{\circ}{L}_i = L_i \setminus P$ ($i = 1, 2$). Строгая делимость множеств S^+ и S^- гиперплоскостью P выражается утверждением

$$S^+ \subseteq \overset{\circ}{L}_1, \quad S^- \subseteq \overset{\circ}{L}_2. \quad (2.8)$$

Нетрудно доказать, что множества $\overset{\circ}{L}_1$ и $\overset{\circ}{L}_2$ выпуклы, поэтому из (2.8) и из определения выпуклой оболочки следует, что

$$\text{Conv}(S^+) \subseteq \overset{\circ}{L}_1, \quad \text{Conv}(S^-) \subseteq \overset{\circ}{L}_2. \quad (2.9)$$

Поскольку $\overset{\circ}{L}_1 \cap \overset{\circ}{L}_2 = \emptyset$, то из (2.9) следует (2.7).

Докажем обратную импликацию. Предположим, что верно (2.7).

Поскольку множество S конечно, то S^+ и S^- тоже конечны. Как было отмечено выше, в этом случае множества $\text{Conv}(S^+)$ и $\text{Conv}(S^-)$ компактны. Следовательно, множество

$$D_S \stackrel{\text{def}}{=} \text{Conv}(S^+) \times \text{Conv}(S^-)$$

тоже компактно.

Обозначим символом ρ функцию вида $D_S \rightarrow \mathbf{R}$, которая сопоставляет паре точек $(x, y) \in D_S$ длину $|xy|$ отрезка, соединяющего эти точки. Функция ρ непрерывна, и ее область определения – компакт, поэтому ρ принимает наименьшее значение в некоторой точке $(x^+, x^-) \in D_S$. Если бы $\rho(x^+, x^-)$ было равно 0, т.е. $x^+ = x^-$, то это бы противоречило предположению (2.7). Следовательно, $\rho(x^+, x^-) > 0$.

Обозначим записями P_{S^+} и P_{S^-} гиперплоскости, проходящие через x^+ и x^- соответственно, и ортогональные отрезку $[x^+, x^-]$. Гиперплоскости P_{S^+} и P_{S^-} параллельны и разбивают \mathbf{R}^n на три множества:

- два полупространства (обозначим их записями L_{S^+} и L_{S^-}), и
- полосу (P_{S^+}, P_{S^-}) между этими полупространствами (не содержащую точек из P_{S^+} и P_{S^-}).

Нетрудно видеть, что

$$L_{S^+} = \{x \in \mathbf{R}^n \mid x = x^+ \text{ или } \widehat{xx^+x^-} \geq \frac{\pi}{2}\}$$

и аналогичное свойство верно для L_{S^-} .

Докажем, что $S^+ \subseteq L_{S^+}$, т.е. $\forall x \in S^+ \setminus \{x^+\} \widehat{xx^+x^-} \geq \frac{\pi}{2}$. Если бы это было неверно, т.е. $\widehat{xx^+x^-} < \frac{\pi}{2}$, то на отрезке $[x^+x]$ была бы точка x' , такая, что $|x'x^-| < |x^+x^-|$. Т.к. $[x^+x] \subseteq \text{Conv}(S^+)$, то $x' \in \text{Conv}(S^+)$. Таким образом, $(x', x^-) \in D_S$ и $\rho(x', x^-) < \rho(x^+, x^-)$. Это противоречит тому, что ρ принимает наименьшее значение на паре (x^+, x^-) .

Аналогично доказывается включение $S^- \subseteq L_{S^-}$.

Из доказанного выше следует, что $(S^+ \cup S^-) \cap (P_{S^+}, P_{S^-}) = \emptyset$.

В качестве гиперплоскости, строго разделяющей S^+ и S^- , можно взять, например, гиперплоскость P , проходящую через любую внутреннюю точку отрезка $[x^+x^-]$ и ортогональную этому отрезку (т.е. P будет параллельна P_{S^+} и P_{S^-}). Поскольку

- $P \subseteq (P_{S^+}, P_{S^-})$, и
- полупространства L_{S^+} и L_{S^-} содержатся в разных полупространствах, на которые P делит \mathbf{R}^n ,

то, следовательно, S^+ и S^- содержатся в разных полупространствах, на которые P делит \mathbf{R}^n . ■

Для задачи классификации в случае строго разделимой выборки S существует алгоритм нахождения функции a_S вида (2.1) со свойством $Q(a_S) = 0$, который называется **алгоритмом обучения Розенблатта** и излагается в следующем параграфе.

2.2 Алгоритм обучения Розенблатта

В этом параграфе рассматривается задача классификации, в которой

- множества объектов и ответов имеют вид $X = \mathbf{R}^n$, $Y = \{-1, 1\}$, и

- задача обучения заключается в нахождении по выборке S АФ a_S вида

$$a_S(x) = \text{sign}\left(\sum_{i=1}^n x^i w_i - w_0\right), \quad \text{где } w_1, \dots, w_n, w_0 \in \mathbf{R}, \quad (2.10)$$

которая минимизирует риск $Q(a_S)$.

2.2.1 Описание алгоритма обучения Розенблатта

Пусть выборка S строго линейно разделима, т.е. $\exists w_1, \dots, w_n, w_0 \in \mathbf{R}$:

$$\forall (x, y) \in S \quad \left(\sum_{i=1}^n x^i w_i - w_0\right)y > 0, \quad \text{где } (x^1, \dots, x^n) = x. \quad (2.11)$$

Нетрудно видеть, что неравенство в (2.11) равносильно неравенству

$$\langle (x^1, \dots, x^n, -1), (w_1, \dots, w_n, w_0) \rangle y > 0,$$

поэтому задача нахождения для строго разделимой выборки S такого вектора (w_1, \dots, w_n, w_0) , который удовлетворяет условию (2.11), сводится к задаче нахождения вектора $w \in \mathbf{R}^{n+1}$, такого, что

$$\forall (x, y) \in S \quad \langle (x, -1), w \rangle y > 0,$$

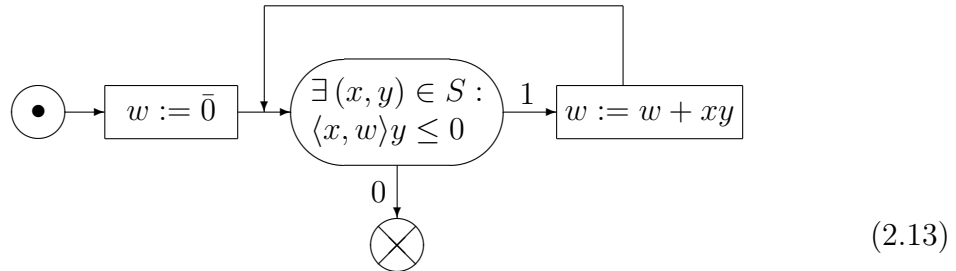
где $(x, -1) \in \mathbf{R}^{n+1}$ получается из x добавлением компоненты, равной -1 .

Таким образом, задача нахождения для строго разделимой выборки АФ a_S со свойством $Q(a_S) = 0$ сводится к следующей задаче: пусть выборка S такова, что существует вектор w , обладающий свойством

$$\forall (x, y) \in S \quad \langle x, w \rangle y > 0. \quad (2.12)$$

Требуется найти хотя бы один вектор w , обладающий свойством (2.12).

Для поиска такого вектора w можно использовать излагаемый ниже алгоритм, называемый **алгоритмом Розенблатта** [5]. Данный алгоритм можно представить в виде следующей блок-схемы:



где $\bar{0}$ – вектор, компоненты которого равны 0, и x, y в правом прямоугольнике – компоненты какой-либо пары $(x, y) \in S$, такой, что $\langle x, w \rangle y \leq 0$.

Нетрудно видеть, что после завершения работы данного алгоритма значение переменной w будет обладать свойством (2.12).

2.2.2 Завершаемость алгоритма Розенблатта

Завершаемость алгоритма Розенблатта обосновывается нижеследующей теоремой, которая называется **теоремой Новикова**.

Теорема 2 (A. Novikoff, 1962) [6], [7].

Пусть конечная выборка $S \subset \mathbf{R}^n \times \{-1, 1\}$ такова, что свойство (2.12) выполняется для некоторого вектора \hat{w} , т.е.

$$\forall (x, y) \in S \quad \langle x, \hat{w} \rangle y > 0.$$

Тогда алгоритм Розенблатта, применяемый к выборке S , завершает свою работу после не более чем $(\sigma/\rho)^2$ циклов, где

$$\sigma = \max_{(x,y) \in S} \|x\|, \quad \rho = \frac{1}{\|\hat{w}\|} \min_{(x,y) \in S} \langle x, \hat{w} \rangle y.$$

Доказательство.

В доказательстве данной теоремы используется **неравенство Коши-Буняковского**, которое верно для произвольной пары a, b векторов из линейного пространства со скалярным произведением $\langle \rangle$, имеет вид

$$\langle a, b \rangle \leq \|a\| \|b\| \tag{2.14}$$

и доказывается следующим образом: $\forall t \in \mathbf{R}$

$$\begin{aligned} 0 &\leq \|at + b\|^2 = \\ &= \langle at + b, at + b \rangle = \\ &= \langle a, a \rangle t^2 + 2\langle a, b \rangle t + \langle b, b \rangle = \\ &= \|a\|^2 t^2 + 2\langle a, b \rangle t + \|b\|^2. \end{aligned}$$

Соотношение

$$\forall t \in \mathbf{R} \quad \|a\|^2 t^2 + 2\langle a, b \rangle t + \|b\|^2 \geq 0 \tag{2.15}$$

равносильно тому, что дискриминант

$$(2\langle a, b \rangle)^2 - 4\|a\|^2 \|b\|^2 \tag{2.16}$$

квадратного трехчлена в (2.15) неположителен, т.е.

$$\langle a, b \rangle^2 - \|a\|^2 \|b\|^2 \leq 0,$$

откуда следует (2.14).

Обозначим записями $w^{(k-1)}$ и $w^{(k)}$ значения переменной w до и после k -го выполнения цикла в (2.13). Т.к. $w^{(k)} = w^{(k-1)} + xy$, то

$$\begin{aligned} \langle w^{(k)}, \hat{w} \rangle &= \langle w^{(k-1)} + xy, \hat{w} \rangle = \langle w^{(k-1)}, \hat{w} \rangle + \langle x, \hat{w} \rangle y \geq \\ &\geq \langle w^{(k-1)}, \hat{w} \rangle + \|\hat{w}\| \rho. \end{aligned} \quad (2.17)$$

Применяя (2.17) k раз, и учитывая $w^{(0)} = \bar{0}$, получаем:

$$\langle w^{(k)}, \hat{w} \rangle \geq \langle w^{(0)}, \hat{w} \rangle + k\|\hat{w}\|\rho = k\|\hat{w}\|\rho. \quad (2.18)$$

Согласно неравенству Коши-Буняковского, $\langle w^{(k)}, \hat{w} \rangle \leq \|w^{(k)}\| \|\hat{w}\|$, поэтому из (2.18) следует, что

$$k\|\hat{w}\|\rho \leq \langle w^{(k)}, \hat{w} \rangle \leq \|w^{(k)}\| \|\hat{w}\|,$$

откуда следует неравенство $k\rho \leq \|w^{(k)}\|$, или

$$k^2\rho^2 \leq \|w^{(k)}\|^2. \quad (2.19)$$

С другой стороны,

$$\|w^{(k)}\|^2 = \langle w^{(k-1)} + xy, w^{(k-1)} + xy \rangle = \|w^{(k-1)}\|^2 + 2\langle x, w^{(k-1)} \rangle y + \|x\|^2, \quad (2.20)$$

и поскольку $\langle x, w^{(k-1)} \rangle y \leq 0$ и $\|x\| \leq \sigma$, то из (2.20) следует, что

$$\|w^{(k)}\|^2 \leq \|w^{(k-1)}\|^2 + \|x\|^2 \leq \|w^{(k-1)}\|^2 + \sigma^2. \quad (2.21)$$

Применяя (2.21) k раз, и учитывая $w^{(0)} = \bar{0}$, получаем:

$$\|w^{(k)}\|^2 \leq k\sigma^2. \quad (2.22)$$

Объединяя (2.19) и (2.22) получаем:

$$k^2\rho^2 \leq \|w^{(k)}\|^2 \leq k\sigma^2,$$

откуда следует неравенство $k^2\rho^2 \leq k\sigma^2$, или $k\rho^2 \leq \sigma^2$, или $k \leq (\sigma/\rho)^2$.

Таким образом, число выполнений цикла в блок-схеме (2.13) не превосходит $(\sigma/\rho)^2$. ■

2.2.3 Модификации алгоритма Розенблатта

Для ускорения процесса нахождения искомого вектора w , удовлетворяющего условию (2.12), приведенный выше алгоритм можно модифицировать:

- в качестве начального значения w в (2.13) можно брать не $\bar{0}$, а любой вектор, рекомендуется брать в качестве начального вектор

$$\left(\frac{\langle x^1, y \rangle}{\|x^1\|^2}, \dots, \frac{\langle x^n, y \rangle}{\|x^n\|^2} \right),$$

где вектора x^1, \dots, x^n, y , определяются следующим образом: если S имеет вид $\{(x_i, y_{x_i}) \mid i = 1, \dots, l\}$, и $\forall i = 1, \dots, l \quad x_i = (x_i^1, \dots, x_i^n)$, то

$$\forall j = 1, \dots, n \quad x^j = (x_1^j, \dots, x_l^j), \quad y = (y_{x_1}, \dots, y_{x_l})$$

(это оптимальный выбор, если $\langle x^i, x^j \rangle$ близко к 0 при $i \neq j$),

- действие $w := w + xy$ можно заменить на $w := w + xy\eta$, где η – фиксированное положительное число.

2.3 Метод градиентного спуска

2.3.1 Понятие градиентного спуска

Как было отмечено в пункте 1.2.3, если задача ML заключается в нахождении по обучающей выборке $S = \{(x_i, y_{x_i}) \mid i = 1, \dots, l\}$ такого параметра w , который минимизирует функцию риска

$$Q(a_S) = \frac{1}{l} \sum_{i=1}^l \mathcal{L}(a(x_i, w), x_i), \quad (2.23)$$

то в том случае, когда функция (2.23) дифференцируема по w , для нахождения искомого параметра w можно использовать **метод градиентного спуска**. Данный метод заключается в итеративном построении последовательных приближений к искомому значению параметра w путем небольших изменений этого параметра. Эти изменения выбираются такими, чтобы на каждой итерации новое значение w как можно сильнее уменьшало бы функцию (2.23).

Метод градиентного спуска не всегда обеспечивает нахождение параметра w , минимизирующего функцию (2.23) (параметр с таким свойством называется **глобальным минимумом**). Иногда с помощью этого метода можно найти лишь **локальный минимум**, т.е. такое значение w , которое невозможно улучшить путем небольших изменений.

2.3.2 Описание метода градиентного спуска

Метод градиентного спуска (МГС) предназначен для поиска минимумов дифференцируемых функций нескольких переменных, и заключается в следующем.

Пусть задана дифференцируемая функция Q от n переменных:

$$Q : W \rightarrow \mathbf{R}, \quad \text{где } W \subseteq \mathbf{R}^n \text{ – открытое множество.}$$

Для нахождения такого $w \in W$, значение функции Q на котором было бы как можно меньшим, производятся следующие действия:

1. выбирается первое приближение $w^{(0)} \in W$ к искомому значению w ,
2. выбирается число $\delta > 0$,
3. среди всех точек, входящих в δ -окрестность точки $w^{(0)}$ выбирается такая точка $w^{(1)} \in W$ (являющаяся следующим приближением к искомому значению w), значение функции Q на которой было бы как можно меньшим, если же значения Q во всех точках δ -окрестности $w^{(0)}$ примерно одинаковы, то работа завершается,
4. предыдущие два действия повторяются, только в качестве $w^{(0)}$ теперь рассматривается точка $w^{(1)}$, в результате выполнения этих действий строится новое приближение $w^{(2)}$, и т.д.

Таким образом, данный алгоритм строит последовательность точек $w^{(0)}, w^{(1)}, \dots, w^{(k)}, w^{(k+1)}, \dots$, в которой переход от каждой точки $w^{(k)}$ к следующей точке $w^{(k+1)}$ производится в соответствии с действием 3.

Рассмотрим более подробно вопрос о том, как следует выбрать точку $w^{(k+1)}$ в действии 3, чтобы значение $Q(w^{(k+1)})$ было бы как можно меньшим. Будем искать $w^{(k+1)}$ в виде

$$w^{(k+1)} = w^{(k)} + \Delta w.$$

Как известно из математического анализа, $\forall \varepsilon > 0 \exists \delta > 0$: значения функции Q в δ -окрестности точки $w^{(k)}$ можно представить (с точностью до ε) линейной формой, т.е. для каждого вектора $\Delta w \in \mathbf{R}^n$, такого, что $\|\Delta w\| \leq \delta$, верно соотношение

$$Q(w^{(k+1)}) = Q(w^{(k)} + \Delta w) \approx_{\varepsilon} Q(w^{(k)}) + a_1 \Delta w_1 + \dots + a_n \Delta w_n, \quad (2.24)$$

где $\Delta w_1, \dots, \Delta w_n$ – компоненты вектора Δw , и $\forall x, y \in \mathbf{R}$ запись $x \approx_{\varepsilon} y$ обозначает утверждение $|x - y| < \varepsilon$.

Вектор (a_1, \dots, a_n) с компонентами из (2.24) обозначается записью $\nabla Q(w^{(k)})$ и называется **градиентом** функции Q в точке $w^{(k)}$, его компоненты a_1, \dots, a_n обозначаются соответственно записями

$$\frac{\partial Q}{\partial w_1}(w^{(k)}), \dots, \frac{\partial Q}{\partial w_n}(w^{(k)}).$$

Соотношение (2.24) можно записать, используя обозначение для скалярного произведения векторов:

$$Q(w^{(k+1)}) = Q(w^{(k)} + \Delta w) \approx_\varepsilon Q(w^{(k)}) + \langle \nabla Q(w^{(k)}), \Delta w \rangle,$$

откуда видно, что выбор точки $w^{(k+1)}$ из δ -окрестности точки $w^{(k)}$, так, чтобы значение $Q(w^{(k+1)})$ (с точностью до ε) было бы как можно меньше, сводится к выбору вектора $\Delta w \in \mathbf{R}^n$, такого, что $\|\Delta w\| \leq \delta$, и значение

$$\langle \nabla Q(w^{(k)}), \Delta w \rangle \tag{2.25}$$

было бы как можно меньше.

Рассмотрим задачу выбора такого вектора Δw в общем виде. Пусть a, b – пара ненулевых векторов из линейного пространства со скалярным произведением $\langle \cdot, \cdot \rangle$. Исследуем вопрос о том, как при фиксированном векторе a выбрать вектор b с ограничением $\|b\| \leq \delta$, чтобы скалярное произведение $\langle a, b \rangle$ принимало наименьшее возможное значение.

Обозначим записью $\cos(a, b)$ число $\frac{\langle a, b \rangle}{\|a\| \|b\|}$. Таким образом,

$$\langle a, b \rangle = \|a\| \|b\| \cos(a, b). \tag{2.26}$$

Согласно неравенству Коши-Буняковского, верны неравенства

$$\langle a, b \rangle \leq \|a\| \|b\| \quad \text{и} \quad \langle -a, b \rangle \leq \|-a\| \|b\|,$$

второе из которых равносильно неравенству $-\|a\| \|b\| \leq \langle a, b \rangle$, поэтому

$$-1 \leq \cos(a, b) \leq 1.$$

Докажем, что $\begin{cases} \cos(a, b) = 1 & \Leftrightarrow \exists t > 0 : b = at, \\ \cos(a, b) = -1 & \Leftrightarrow \exists t > 0 : b = -at. \end{cases}$

- Если $\exists t > 0 : b = at$, $\cos(a, b) = \frac{\langle a, b \rangle}{\|a\| \|b\|} = \frac{\langle a, at \rangle}{\|a\| \|at\|} = \frac{\langle a, a \rangle t}{\|a\| \|a\| t} = 1$.
- Если $\exists t > 0 : b = -at$, то $\cos(a, b) = \frac{\langle a, b \rangle}{\|a\| \|b\|} = \frac{\langle a, -at \rangle}{\|a\| \|-at\|} = \frac{\langle a, a \rangle (-t)}{\|a\| \|a\| t} = -1$.

- Если же два предыдущих случая не имеют место, то $\forall t \in \mathbf{R}$ вектор $at + b$ не является нулевым, поэтому

$$\forall t \in \mathbf{R} \quad \|at + b\| > 0,$$

т.е. квадратный трехчлен в (2.15) принимает только положительные значения, откуда следует, что его дискриминант (2.16) отрицателен, т.е. $\langle a, b \rangle^2 - \|a\|^2 \|b\|^2 < 0$, поэтому $|\cos(a, b)| \neq 1$. ■

Из приведенных рассуждений следует, что (2.26) принимает наименьшее возможное значение в том случае, когда

- $\cos(a, b) = -1$ (т.е. $b = -at$, где $t > 0$), и
- $\|b\|$ принимает наибольшее возможное значение (т.е. $\|b\| = \delta$).

Таким образом, решением задачи является вектор $b = -a \frac{\delta}{\|a\|}$.

Если под a и b понимаются вектора $\nabla Q(w^{(k)})$ и Δw соответственно, то заключаем, что искомый вектор Δw должен иметь вид

$$\Delta w = -\nabla Q(w^{(k)})\eta, \quad (2.27)$$

где η – некоторое небольшое положительное число, называемое **темпом обучения**. Обычно данное число не вычисляется аналитически, а подбирается в процессе работы алгоритма, с использованием некоторых эвристических соображений, среди которых м.б. следующие:

- если η слишком мало, то алгоритм может работать слишком долго (т.к. число итераций в процессе работы будет слишком большим),
- если η слишком велико, то алгоритм может работать неустойчиво, где под **устойчивостью** работы алгоритма понимается сходимость последовательности $w^{(0)}, w^{(1)}, w^{(2)}, \dots$

Как правило, значение η не является фиксированным, а постоянно корректируется в процессе работы алгоритма: сначала оно выбирается небольшим, затем постепенно увеличивается до максимального значения, при котором алгоритм все еще работает устойчиво, в случае возникновения неустойчивости значение η уменьшается, и т.д.

Как было отмечено в описании действия 3, работа алгоритма завершается в том случае когда значения минимизируемой функции Q во всех точках окрестности текущего приближения $w^{(k)}$ к искомому значению примерно (с точностью до ε) одинаковы. Нетрудно видеть, что данная

ситуация эквивалентна тому, что компоненты градиента $\nabla Q(w^{(k)})$ примерно (с точностью до ε) равны нулю. Таким образом, условие завершения работы алгоритма градиентного спуска выражается соотношением

$$\nabla Q(w^{(k)}) \approx_{\varepsilon} (0, \dots, 0).$$

Как было отмечено выше, найденный вектор $w^{(k)}$, на котором завершается работа данного алгоритма, может быть лишь локальным минимумом функции Q . Он будет глобальным минимумом функции Q , если

- удачно выбрано начальное приближение $w^{(0)}$, или
- функция Q является **выпуклой**, т.е. $\forall w, w' \in W, \forall \alpha \in [0, 1]$

$$\begin{aligned} \alpha w + (1 - \alpha)w' &\in W, \\ Q(\alpha w + (1 - \alpha)w') &\leq \alpha Q(w) + (1 - \alpha)Q(w'). \end{aligned}$$

Докажем, что если Q выпукла, и в точке $\hat{w} \in W$ верно соотношение

$$\nabla Q(\hat{w}) = (0, \dots, 0), \quad (2.28)$$

то $Q(\hat{w}) = \min_{w \in W} Q(w)$.

Пусть это неверно, т.е. $\exists w \in W : Q(w) < Q(\hat{w})$, тогда $\forall \alpha \in [0, 1]$

$$\begin{aligned} Q(\hat{w} + \alpha(w - \hat{w})) &= Q(\alpha w + (1 - \alpha)\hat{w}) \leq \\ &\leq \alpha Q(w) + (1 - \alpha)Q(\hat{w}) = Q(\hat{w}) - \alpha(Q(\hat{w}) - Q(w)) \end{aligned} \quad (2.29)$$

Как известно из математического анализа, из (2.28) следует, что

$$Q(\hat{w} + \alpha(w - \hat{w})) - Q(\hat{w}) = o(\alpha),$$

т.е. $\lim_{\alpha \rightarrow 0} \frac{Q(\hat{w} + \alpha(w - \hat{w})) - Q(\hat{w})}{\alpha} = 0$. Но согласно (2.29), этот предел меньше чем

$$-(Q(\hat{w}) - Q(w)) < 0.$$

Таким образом, в рассматриваемой ситуации результат работы МГС (если он достигается) – глобальный минимум Q .

2.3.3 Модификации метода градиентного спуска

Метод стохастического градиента

В том случае, когда обучающая выборка S имеет большой размер, применение МГС может вызвать большие вычислительные сложности, т.к.

на каждой итерации необходимо вычислять градиент $\nabla Q(w^{(k)})$, который зависит от всех элементов обучающей выборки S :

$$\forall i = 1, \dots, n \quad \frac{\partial Q}{\partial w_i}(w^{(k)}) = \frac{1}{l} \sum_{j=1}^l \frac{\partial \mathcal{L}(a(x_j, w^{(k)}), x_j)}{\partial w_i}$$

Для ускорения процесса обучения иногда вместо правила (2.27) используется правило

$$\Delta w = -\nabla \mathcal{L}(a(x_j, w^{(0)}), x_j) \eta, \quad (2.30)$$

где число $j \in \{1, \dots, l\}$ на каждой итерации процесса обучения выбирается случайно. Соответствующий метод обучения (с правилом (2.30) вместо (2.27)) называется **методом стохастического градиента**.

Одной из актуальных проблем является управление выбором j в (2.30) на каждой итерации процесса обучения, так, чтобы сходимость $w^{(k)}$ к оптимальному параметру была бы как можно более быстрой.

Регуляризация

Одной из нежелательных ситуаций во время обучения является чрезмерный рост $\|w^{(k)}\|$. Данная ситуация может возникнуть, например, в следующем случае: предсказательная модель $a : X \times W \rightarrow Y$ имеет вид

$$a(x, w) = \text{sign}(\langle x, w \rangle), \quad \text{где } x \in X \subseteq \mathbf{R}^n,$$

и $\exists u \in \mathbf{R}^n : \forall x \in X \langle x, u \rangle = 0$.

Нетрудно видеть, что в этом случае

$$\forall \gamma \in \mathbf{R} \quad a(x, w + \gamma u) = \text{sign}(\langle x, w + \gamma u \rangle) = \text{sign}(\langle x, w \rangle) = a(x, w)$$

откуда следует, что если минимальное значение риска будет достигаться на \hat{w} , то такое же значение риска будет достигаться на $\hat{w} + \gamma u$ ($\forall \gamma \in \mathbf{R}$), т.е. параметр w , минимизирующий риск, м.б. как угодно большим.

Для борьбы с чрезмерным увеличением $\|w^{(k)}\|$ используется метод, называемый **регуляризацией**. Суть данного метода заключается в модификации минимизируемой функции: она может иметь, например, вид

$$Q(a_S) + \frac{\tau}{2} \|w\|^2,$$

где τ – некоторое положительное число. В этом случае (2.27) заменяется на правило

$$\Delta w = -\nabla Q(w^{(0)}) \eta - \tau w \eta. \quad (2.31)$$

Можно модифицировать не минимизируемую функцию, а функцию потерь: вместо \mathcal{L} рассматривать $\tilde{\mathcal{L}} \stackrel{\text{def}}{=} \mathcal{L} + \frac{\tau}{2} \|w\|^2$, в этом случае (2.27) тоже заменяется на (2.31).

2.4 Метод обратного распространения ошибки для обучения нейронных сетей

2.4.1 Идея метода

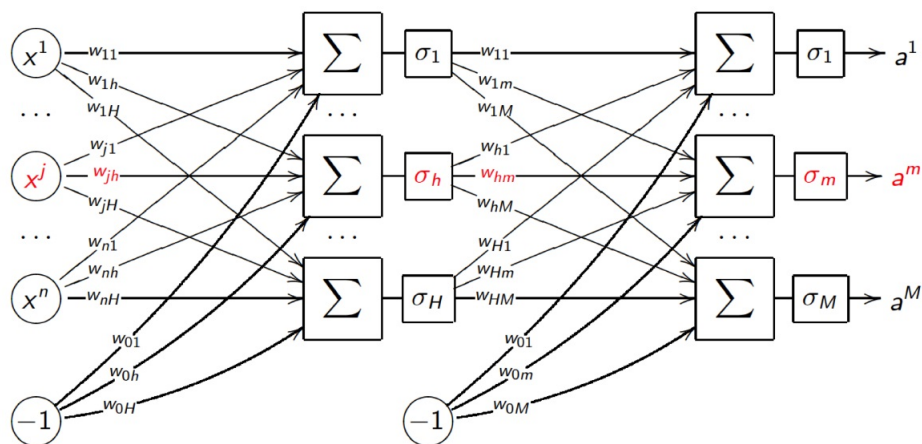
Излагаемый в этом параграфе метод обратного распространения ошибки (error back propagation), или более коротко – метод обратного распространения (МОР), используется при обучении многослойных нейронных сетей (МНС). Данный метод является модификацией метода градиентного спуска. Впервые МОР был описан в 1974 г. А. И. Галушкиным, а также независимо и одновременно Полом Дж. Вербосом.

Идея МОР состоит в распространении сигналов ошибки от выходов МНС к ее входам, в направлении, обратном прямому распространению сигналов в обычном режиме работы. Опишем эту идею более подробно.

Напомним, что компонентами МНС являются нейроны,

- на вход нейрона поступает кортеж чисел вида $(x^1, \dots, x^n) \in \mathbf{R}^n$,
- на выходе нейрон выдает число $a \stackrel{\text{def}}{=} \sigma(\langle x, w \rangle - w_0)$, где σ – функция активации.

Структуру МНС можно представить диаграммой вида



(2.32)

(на данной диаграмме изображена МНС с двумя слоями).

При заданной совокупности w значений весовых коэффициентов w_{ij} эта МНС определяет функцию a_w , отображающую каждый входной вектор $x \in \mathbf{R}^n$ в выходной вектор $a_w(x) \in \mathbf{R}^M$. Если задана обучающая

выборка $S \subseteq \mathbf{R}^n \times \mathbf{R}^M$, то ошибкой данной МНС на паре $(x, y) \in S$ называется число

$$Q(x, y, w) = \frac{1}{2} \|a_w(x) - y\|^2. \quad (2.33)$$

Задача алгоритма МОР заключается в нахождении такой совокупности w весовых коэффициентов данной МНС, которые делают ошибки (2.33) как можно меньше. Алгоритм МОР решает эту задачу путем выполнения нескольких итераций, каждая из которых состоит из двух частей:

- выбор пары $(x, y) \in S$,
- нахождение ошибки (2.33) на выбранной паре (x, y) при текущем наборе весовых коэффициентов w путем вычисления в «прямом направлении» (слева направо) выходов всех нейронов,
- коррекция весовых коэффициентов w_{ij} путем вычисления в «обратном направлении» (сначала корректируются весовые коэффициенты последнего слоя, затем - предпоследнего, и т.д.).

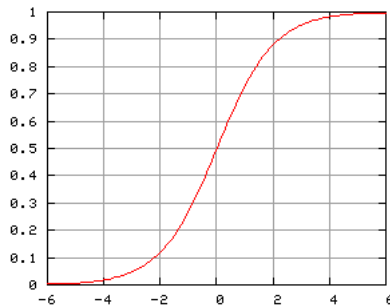
2.4.2 Описание метода

Описание МОР будет изложено на примере двуслойной сети вида (2.32) (для МНС с большим числом слоев метод выглядит аналогично).

Для возможности применения МОР функция активации σ должна быть дифференцируемой. Например, в качестве такой σ может использоваться **сигмоида**:

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \quad (2.34)$$

График этой функции имеет вид



Данная функция стремится к 1 при $x \rightarrow \infty$ и к 0 при $x \rightarrow -\infty$, ее график центрально симметричен относительно точки $(0, 0.5)$.

Ниже будет использоваться легко проверяемое соотношение

$$\sigma'(x) = \sigma(x)(1 - \sigma(x)).$$

Мы будем предполагать, что в рассматриваемой МНС функция активации одинакова для всех входящих в нее нейронов, и имеет вид (2.34).

Алгоритм МОР имеет следующий вид:

1. Инициализация весов МНС небольшими случайными значениями.
2. Делаем итерации (до тех пор пока Q не стабилизируется), каждая итерация заключается в вычислении по текущему набору w весовых коэффициентов нового набора w' , который будет текущим в следующей итерации, и имеет следующий вид:

- случайно выбираем $(x, y) \in S$, $x = (x^1, \dots, x^n)$, $y = (y^1, \dots, y^M)$,
- прямой ход: вычисляем выходы всех нейронов, и

$$Q(x, y, w) := \frac{1}{2} \sum_{m=1}^M (a^m - y^m)^2$$

$$\forall m = 1, \dots, M \quad \frac{\partial Q}{\partial a^m} = a^m - y^m =: \xi^m,$$

- обратный ход (модификация весов в направлении $-\nabla$):

$$w'_{hm} := w_{hm} - \frac{\partial Q}{\partial w_{hm}} \eta, \quad w'_{jh} := w_{jh} - \frac{\partial Q}{\partial w_{jh}} \eta,$$

где $\eta \in (0, 1)$ – подбираемый параметр (темп обучения), и частные производные $\frac{\partial Q}{\partial w_{hm}}$, $\frac{\partial Q}{\partial w_{jh}}$ вычисляются следующим образом: пусть u^1, \dots, u^H – выходы первого слоя, тогда $\forall m = 1, \dots, M$, $\forall h = 1, \dots, H$, $\forall j = 1, \dots, n$

$$a^m = \sigma\left(\sum_{h=1}^H w_{hm} u^h - w_{0m}\right),$$

$$\begin{aligned} \frac{\partial Q}{\partial w_{hm}} &= \frac{\partial Q}{\partial a^m} \frac{\partial a^m}{\partial w_{hm}} = \xi^m \sigma' \left(\sum_{h=1}^H w_{hm} u^h - w_{0m} \right) u^h = \\ &= \xi^m a^m (1 - a^m) u^h, \end{aligned}$$

$$\begin{aligned} \frac{\partial Q}{\partial w_{0m}} &= \frac{\partial Q}{\partial a^m} \frac{\partial a^m}{\partial w_{0m}} = \xi^m \sigma' \left(\sum_{h=1}^H w_{hm} u^h - w_{0m} \right) (-1) = \\ &= -\xi^m a^m (1 - a^m), \end{aligned}$$

$$\begin{aligned}
u^h &= \sigma\left(\sum_{j=1}^n w_{jh}x^j - w_{0h}\right), \\
\frac{\partial Q}{\partial u^h} &= \sum_{m=1}^M \frac{\partial Q}{\partial a^m} \frac{\partial a^m}{\partial u^h} = \sum_{m=1}^M \xi^m \sigma'\left(\sum_{h=1}^H w_{hm}u^h - w_{0m}\right)w_{hm} = \\
&= \sum_{m=1}^M \xi^m a^m(1 - a^m)w_{hm} =: \zeta^h, \\
\frac{\partial Q}{\partial w_{jh}} &= \frac{\partial Q}{\partial u^h} \frac{\partial u^h}{\partial w_{jh}} = \zeta^h \sigma'\left(\sum_{j=1}^n w_{jh}x^j - w_{0h}\right)x^j = \\
&= \zeta^h u^h(1 - u^h)x^j, \\
\frac{\partial Q}{\partial w_{0h}} &= \frac{\partial Q}{\partial u^h} \frac{\partial u^h}{\partial w_{0h}} = \zeta^h \sigma'\left(\sum_{j=1}^n w_{jh}x^j - w_{0h}\right)(-1) = \\
&= -\zeta^h u^h(1 - u^h).
\end{aligned}$$

2.4.3 Достоинства и недостатки метода

Основные достоинства МОР:

- низкая сложность,
- легко реализуется на параллельных архитектурах,
- универсальность (пригоден для любых конфигураций МНС).

Основные недостатки МОР заключаются в следующем.

- Неопределенно долгий процесс обучения. В сложных задачах для обучения сети могут потребоваться дни или даже недели, она может и вообще не обучиться.
- В процессе обучения сети значения весов могут в результате коррекции стать очень большими величинами. Это может привести к тому, что большинство нейронов будут функционировать при очень больших значениях весовых коэффициентов, в области, где производная функции активации очень мала. Так как обратно распространяемая в процессе обучения ошибка пропорциональна этой производной, то процесс обучения может стать парализованным.
- Нет гарантии того, что получаемый в результате обучения локальный минимум является хорошим решением задачи обучения.

Для улучшения сходимости алгоритма обратного распространения можно использовать, например, следующие приемы:

- нормализация входных значений: вектор x в каждой паре $(x, y) \in S$ заменяется на \tilde{x} , определяемый следующим образом:

$$\forall i = 1, \dots, n \quad \tilde{x}^i := \frac{x^i - x_{min}^i}{x_{max}^i - x_{min}^i} \quad \text{или} \quad \tilde{x}^i := \frac{x^i - x_{average}^i}{x_{avsquare}^i}$$

где $x_{average}^i$ – среднее значение, $x_{avsquare}^i$ – среднеквадратическое отклонение ($= \sqrt{\text{дисперсии}}$),

- добавление между слоями МНС промежуточных слоев, реализующих линейные преобразования: если u и v – вектора входов и выходов такого слоя, то реализуемое этим слоем преобразование имеет вид $v = Au + b$, где A и b – матрица и вектор соответствующих размерностей, коэффициенты матрицы A и вектора b тоже обучаются,
- изменение структуры МНС: удаление части нейронов (**dropout**).

2.5 Метод опорных векторов

В этом параграфе излагается наиболее популярный метод машинного обучения – **метод опорных векторов (Support Vector Machines, SVM)**, который был создан в 70-е годы прошлого века сотрудниками Института проблем управления АН СССР В. Н. Вапником и А. Я. Червоненкисом, и впервые опубликован в книге [8] (в которой он назван **методом обобщенного портрета**).

Данный метод предназначен для решения задач классификации и регрессионного анализа.

2.5.1 Оптимальность аппроксимирующих функций

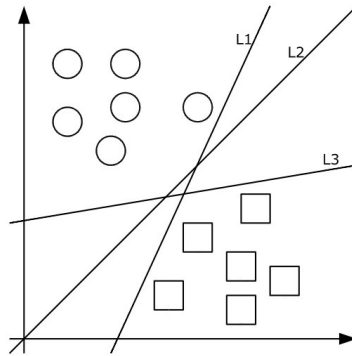
В параграфе 2.2 рассматривалась задача нахождения по строго линейно разделимой выборке $S \subseteq \mathbf{R}^n \times \{-1, 1\}$ АФ a_S вида

$$a_S(x) = \text{sign}\left(\sum_{i=1}^n x^i w_i - w_0\right)$$

такой, что $Q(a_S) = 0$. Как было отмечено в этом параграфе, функция a_S данного вида обладает свойством $Q(a_S) = 0$ тогда и только тогда, когда гиперплоскость P , определяемая уравнением $\sum_{i=1}^n x^i w_i - w_0 = 0$, разделяет

множества S^+ и S^- , т.е. S^+ и S^- содержатся в разных полупространствах, на которые P делит \mathbf{R}^n .

Можно доказать, что задача построения разделяющей гиперплоскости для строго линейно разделимой выборки S имеет бесконечно много решений. Например, несколько различных решений данной задачи изображено на нижеследующем рисунке (в данном случае $n = 2$):



где кружочки обозначают элементы S^+ , а квадратики - элементы S^- .

Встает вопрос о том, можно ли ввести какие-либо меры оптимальности решений данной задачи.

В качестве одной из мер оптимальности функции a_S указанного выше вида можно рассматривать, например, расстояние

$$\rho(S^+ \cup S^-, P) \quad (2.35)$$

между $S^+ \cup S^-$ и P . Напомним, что $\forall A, B \subseteq \mathbf{R}^n$ расстояние $\rho(A, B)$ между A и B определяется как $\inf_{a \in A, b \in B} \|a - b\|$.

Назовем **полосой, разделяющей S^+ и S^-** , и определяемой гиперплоскостью P , часть пространства \mathbf{R}^n , заключенную между гиперплоскостями P_{S^+} и P_{S^-} , которые получаются параллельным переносом гиперплоскости P вдоль вектора нормали к ней

- по направлению к S^+ на расстояние $\rho(P, S^+)$, и
- по направлению к S^- на расстояние $\rho(P, S^-)$,

соответственно. Будем обозначать эту полосу записью $[P_{S^+}, P_{S^-}]$. Нетрудно видеть, что во внутренней части полосы $[P_{S^+}, P_{S^-}]$ точек из S^+ и S^- нет.

Расстояние $\rho(P_{S^+}, P_{S^-})$ назовем **шириной** полосы $[P_{S^+}, P_{S^-}]$. Можно доказать, что (2.35) достигает максимального значения, когда ширина

полосы $[P_{S^+}, P_{S^-}]$ равна $\rho(S^+, S^-)$, и P находится посередине этой полосы.

Назовем гиперплоскость P , находящуюся посередине полосы $[P_{S^+}, P_{S^-}]$ с шириной $\rho(S^+, S^-)$, **оптимальной гиперплоскостью**, разделяющей выборку S . Ниже излагается метод построения такой гиперплоскости.

Кроме того, ниже вводится еще одна мера оптимальности АФ a_S , и излагается алгоритм построения функции a_S , оптимальной относительно этой меры.

2.5.2 Построение оптимальной разделяющей гиперплоскости для строго линейно разделимой выборки

Описание задачи

В этом пункте мы предполагаем, что задана строго линейно разделимая выборка S , и P – какая-либо гиперплоскость, разделяющая S^+ и S^- .

По определению, определенные выше гиперплоскости P_{S^+} и P_{S^-} параллельны, поэтому можно считать, что их уравнения различаются лишь в свободном члене и имеют вид

$$\langle x, v \rangle - a = 0, \quad \langle x, v \rangle - b = 0, \quad \text{где } v \in \mathbf{R}^n, \quad a, b \in \mathbf{R}, \quad a \neq b. \quad (2.36)$$

$\forall \lambda \in \mathbf{R} \setminus \{0\}$ уравнения

$$\langle x, \lambda v \rangle - \lambda a = 0, \quad \langle x, \lambda v \rangle - \lambda b = 0 \quad (2.37)$$

равносильны соответствующим уравнениям в (2.36), т.е. определяют те же самые гиперплоскости P_{S^+} и P_{S^-} . Нетрудно видеть, что если в качестве λ взять число $\frac{2}{a-b}$, то уравнения (2.37) будут равносильны соответствующим уравнениям

$$\langle x, w \rangle - w_0 = 1, \quad \langle x, w \rangle - w_0 = -1, \quad (2.38)$$

где $w = \lambda v$, $w_0 = \frac{a+b}{a-b}$. Таким образом, можно считать, что

- гиперплоскости P_{S^+} и P_{S^-} определяются уравнениями (2.38), и
- точки из S^+ и S^- находятся в непересекающихся полупространствах, определяемых P_{S^+} и P_{S^-} , т.е. удовлетворяют соотношениям

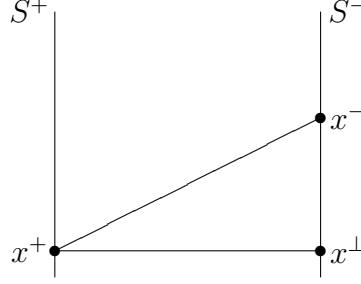
$$\langle x, w \rangle - w_0 \geq 1 \quad \text{и} \quad \langle x, w \rangle - w_0 \leq -1$$

соответственно.

Вычислим ширину ρ полосы $[P_{S^+}, P_{S^-}]$.

Выберем на P_{S^+} и P_{S^-} точки x^+ и x^- соответственно. Пусть x^\perp — основание перпендикуляра, опущенного из точки x^+ на гиперплоскость P_{S^-} .

Искомая ширина ρ равна длине катета $[x^+, x^\perp]$ прямоугольного треугольника с вершинами в точках x^+, x^-, x^\perp :



Эту длину можно вычислить как произведение

- длины гипотенузы $[x^+, x^-]$, т.е. $\|x^+ - x^-\|$, и
- косинуса угла $\varphi = \widehat{x^-x^+x^\perp}$, который выражается через скалярное произведение: $\cos \varphi = \frac{\langle x^+ - x^-, x^+ - x^\perp \rangle}{\|x^+ - x^-\| \|x^+ - x^\perp\|}$,

т.е. $\rho = \frac{\langle x^+ - x^-, x^+ - x^\perp \rangle}{\|x^+ - x^\perp\|}$. Поскольку вектор $x^+ - x^\perp$ ортогонален к P , то он имеет вид μw для некоторого числа μ , поэтому

$$\rho = \frac{\langle x^+ - x^-, \mu w \rangle}{\|\mu w\|} = \frac{\mu}{|\mu|} \frac{\langle x^+ - x^-, w \rangle}{\|w\|} = \sigma \frac{\langle x^+ - x^-, w \rangle}{\|w\|}, \quad (2.39)$$

где $\sigma = 1$ или -1 . Поскольку $x^+ \in P_{S^+}$ и $x^- \in P_{S^-}$, то

$$\langle x^+, w \rangle - w_0 = 1, \quad \langle x^-, w \rangle - w_0 = -1,$$

откуда следует, что

$$\langle x^+ - x^-, w \rangle = \langle x^+, w \rangle - \langle x^-, w \rangle = (w_0 + 1) - (w_0 - 1) = 2, \quad (2.40)$$

Из (2.39) и (2.40) следует, что $\sigma = 1$, и $\rho = \frac{2}{\|w\|}$.

Таким образом, задача поиска оптимальной разделяющей гиперплоскости для S , т.е. такой гиперплоскости P , которая определяет полосу $[S_P^+, S_P^-]$ максимальной ширины, сводится к следующей задаче: найти такие $w \in \mathbf{R}^n$ и $w_0 \in \mathbf{R}$, чтобы

- значение $\|w\|$ было минимально возможным, и

- были выполнены условия $\begin{cases} \forall x \in S^+ & \langle x, w \rangle - w_0 \geq 1, \\ \forall x \in S^- & \langle x, w \rangle - w_0 \leq -1, \end{cases}$ которые можно переписать в виде

$$\forall x \in X_S \quad y_x(\langle x, w \rangle - w_0) - 1 \geq 0 \quad (2.41)$$

где $X_S \stackrel{\text{def}}{=} \{x \in \mathbf{R}^n \mid (x, y_x) \in S\}$.

Если решение (w, w_0) этой задачи найдено, то оптимальная разделяющая гиперплоскость P определяется уравнением $\langle x, w \rangle - w_0 = 0$.

Заметим, что решение данной задачи совпадает с решением задачи

$$\frac{\|w\|^2}{2} \rightarrow \min \quad (2.42)$$

при условиях (2.41). Таким образом, мы свели задачу построения оптимальной разделяющей гиперплоскости для строго линейно разделимой выборки к оптимизационной задаче (2.42) при условиях (2.41).

Метод решения оптимизационной задачи

В этом пункте мы рассмотрим подход к решению общей оптимизационной задачи, частным случаем которой является оптимизационная задача (2.42) при условиях (2.41).

Данная общая задача имеет следующий вид: заданы

- функция $f : \mathbf{R}^n \rightarrow \mathbf{R}$ (называемая **целевой функцией**), которая является **выпуклой**, т.е.

$$\forall x, x' \in \mathbf{R}^n, \forall \alpha \in [0, 1] \quad f(\alpha x + (1 - \alpha)x') \leq \alpha f(x) + (1 - \alpha)f(x'),$$

- множество **условий** вида $g_1(x) \leq 0, \dots, g_m(x) \leq 0$, где g_1, \dots, g_m – выпуклые функции вида $\mathbf{R}^n \rightarrow \mathbf{R}$.

Обозначим записью D_{g_1, \dots, g_m} множество

$$D_{g_1, \dots, g_m} \stackrel{\text{def}}{=} \{x \in \mathbf{R}^n \mid \forall i = 1, \dots, m \quad g_i(x) \leq 0\}. \quad (2.43)$$

Требуется найти аргумент $\hat{x} \in D_{g_1, \dots, g_m}$, на котором достигается минимальное значение функции f , в предположении что минимальность рассматривается относительно множества D_{g_1, \dots, g_m} , т.е. требуется, чтобы

$$f(\hat{x}) = \min_{x \in D_{g_1, \dots, g_m}} f(x).$$

Данную задачу можно решать с помощью **теоремы Каруша-Куна-Таккера**, которую называют основной теоремой выпуклой нелинейной оптимизации. Ниже мы приводим частный случай этой теоремы.

Теорема 3 (W.Karush, 1942, H.Kuhn and A.Tucker, 1951).

Пусть заданы выпуклые функции f, g_1, \dots, g_m вида $\mathbf{R}^n \rightarrow \mathbf{R}$, причем

$$\exists x \in \mathbf{R}^n : \forall i = 1, \dots, m \quad g_i(x) < 0. \quad (2.44)$$

Обозначим символом L функцию (называемую **функцией Лагранжа**)

$$L = f(x) + \sum_{i=1}^m \lambda_i g_i(x), \quad (2.45)$$

где $\lambda_1, \dots, \lambda_m$ – переменные, называемые **множителями Лагранжа**.

$\forall \hat{x} \in D_{g_1, \dots, g_m}$ следующие утверждения эквивалентны:

$$f(\hat{x}) = \min_{x \in D_{g_1, \dots, g_m}} f(x), \quad (2.46)$$

$$\exists \hat{\lambda}_1, \dots, \hat{\lambda}_m \in \mathbf{R} : \begin{cases} \forall i = 1 \dots, m \quad \hat{\lambda}_i \geq 0, \\ \forall i = 1 \dots, m \quad \hat{\lambda}_i g_i(\hat{x}) = 0, \\ L(\hat{x}, \hat{\lambda}_1, \dots, \hat{\lambda}_m) = \min_{x \in \mathbf{R}^n} L(x, \hat{\lambda}_1, \dots, \hat{\lambda}_m). \end{cases} \quad (2.47)$$

Если функции f, g_1, \dots, g_m дифференцируемые, то функция Лагранжа L тоже будет дифференцируемой, и этом случае третье соотношение в (2.47) равносильно соотношению

$$\forall i = 1, \dots, m \quad \frac{\partial L}{\partial x_i}(\hat{x}, \hat{\lambda}_1, \dots, \hat{\lambda}_m) = 0. \quad (2.48)$$

Доказательство.

Сначала докажем, что если для точки $\hat{x} \in D_{g_1, \dots, g_m}$ верно утверждение (2.47), то для нее будет верно утверждение (2.46).

Т.к. $\forall i = 1, \dots, m, \forall x \in D_{g_1, \dots, g_m} \quad \hat{\lambda}_i g_i(x) \leq 0$, то

$$\forall x \in D_{g_1, \dots, g_m} \quad L(x, \hat{\lambda}_1, \dots, \hat{\lambda}_m) = f(x) + \sum_{i=1}^m \hat{\lambda}_i g_i(x) \leq f(x). \quad (2.49)$$

Из второго соотношения в (2.47) следует, что

$$L(\hat{x}, \hat{\lambda}_1, \dots, \hat{\lambda}_m) = f(\hat{x}) + \sum_{i=1}^m \hat{\lambda}_i g_i(\hat{x}) = f(\hat{x}) + \sum_{i=1}^m 0 = f(\hat{x}). \quad (2.50)$$

Из третьего соотношения в (2.47) следует, что

$$\forall x \in \mathbf{R}^n \quad L(\hat{x}, \hat{\lambda}_1, \dots, \hat{\lambda}_m) \leq L(x, \hat{\lambda}_1, \dots, \hat{\lambda}_m). \quad (2.51)$$

Из (2.49) (2.50) и (2.51) следует, что для \hat{x} верно (2.46):

$$\forall x \in D_{g_1, \dots, g_m} \quad f(\hat{x}) = L(\hat{x}, \hat{\lambda}_1, \dots, \hat{\lambda}_m) \leq L(x, \hat{\lambda}_1, \dots, \hat{\lambda}_m) \leq f(x).$$

Теперь докажем, что если для точки $\hat{x} \in D_{g_1, \dots, g_m}$ верно утверждение (2.46), то для нее будет верно утверждение (2.47).

Обозначим символом Λ множество всех векторов $(\lambda_0, \dots, \lambda_m) \in \mathbf{R}^{m+1}$, каждый из которых удовлетворяет следующему условию: $\exists x \in \mathbf{R}^n$:

$$\begin{cases} \lambda_0 > f(x) - f(\hat{x}), \\ \forall i = 1, \dots, m \quad \lambda_i \geq g_i(x). \end{cases}$$

Очевидно что множество Λ непусто.

Нулевой вектор $\bar{0} = (0, \dots, 0)$ не входит в Λ , т.к. иначе $\exists x \in \mathbf{R}^n$:

$$\begin{cases} 0 > f(x) - f(\hat{x}) \quad (\Rightarrow \quad f(x) < f(\hat{x})), \\ \forall i = 1, \dots, m \quad 0 \geq g_i(x) \quad (\Rightarrow \quad x \in D_{g_1, \dots, g_m}), \end{cases}$$

т.е. $\min_{x \in D_{g_1, \dots, g_m}} f(x) < f(\hat{x})$, что противоречит предположению (2.46).

Докажем, что множество Λ выпукло, т.е. $\forall \lambda, \lambda' \in \Lambda, \forall \alpha \in [0, 1]$

$$\lambda^{(\alpha)} \stackrel{\text{def}}{=} \alpha \lambda + (1 - \alpha) \lambda' \in \Lambda. \quad (2.52)$$

Действительно, если λ и λ' имеют вид $(\lambda_0, \dots, \lambda_m)$ и $(\lambda'_0, \dots, \lambda'_m)$ соответственно, и вектора x, x' таковы, что

$$\begin{aligned} \lambda_0 &> f(x) - f(\hat{x}), \quad \forall i = 1, \dots, m \quad \lambda_i \geq g_i(x) \\ \lambda'_0 &> f(x') - f(\hat{x}), \quad \forall i = 1, \dots, m \quad \lambda'_i \geq g_i(x') \end{aligned}$$

то нетрудно проверить (используя выпуклость f, g_1, \dots, g_m), что вектор $x^{(\alpha)} \stackrel{\text{def}}{=} \alpha x + (1 - \alpha)x'$ обладает свойством

$$\begin{cases} \lambda_0^{(\alpha)} > f(x^{(\alpha)}) - f(\hat{x}), \\ \forall i = 1, \dots, m \quad \lambda_i^{(\alpha)} \geq g_i(x^{(\alpha)}). \end{cases} \quad (2.53)$$

Распишем (2.53) во всех деталях:

$$\begin{cases} \lambda_0^{(\alpha)} = \alpha \lambda_0 + (1 - \alpha) \lambda'_0 > \\ > \alpha f(x) - \alpha f(\hat{x}) + (1 - \alpha) f(x') - (1 - \alpha) f(\hat{x}) \geq \\ \geq f(\alpha x + (1 - \alpha)x') - f(\hat{x}) = f(x^{(\alpha)}) - f(\hat{x}), \\ \forall i = 1, \dots, m \quad \lambda_i^{(\alpha)} = \alpha \lambda_i + (1 - \alpha) \lambda'_i \geq \\ \geq \alpha g_i(x) + (1 - \alpha) g_i(x') \geq g_i(\alpha x + (1 - \alpha)x') = g_i(x^{(\alpha)}). \end{cases}$$

Таким образом, (2.52) верно.

Лемма об отделимости.

Множество Λ обладает **свойством отделимости**: $\exists \lambda^* \in \mathbf{R}^{m+1}$:

$$\lambda^* \neq \bar{0}, \forall \lambda \in \Lambda \quad \langle \lambda^*, \lambda \rangle \geq 0 \quad (2.54)$$

(т.е. Λ содержится в одном из полупространств, на которые делит пространство \mathbf{R}^{m+1} гиперплоскость, определяемая уравнением $\langle \lambda^*, x \rangle = 0$).

Доказательство.

Если размерность $Lin(\Lambda)$ линейной оболочки множества Λ (т.е. минимального линейного подпространства пространства \mathbf{R}^{m+1} , содержащего Λ) меньше чем $m + 1$, то в качестве искомого вектора λ^* можно взять любой ненулевой вектор из ортогонального дополнения $Lin(\Lambda)$. В этом случае (2.54) верно по причине того, что

$$\forall \lambda \in \Lambda \quad \langle \lambda^*, \lambda \rangle = 0.$$

Пусть размерность $Lin(\Lambda)$ равна $m + 1$. В этом случае Λ содержит линейно независимое множество из $m + 1$ векторов. Обозначим векторы, входящие в это множество, записями v_1, \dots, v_{m+1} .

Определим $\vec{\Lambda} \stackrel{\text{def}}{=} \{\xi \lambda \mid \xi > 0, \lambda \in \Lambda\}$. Отметим, что $\Lambda \subseteq \vec{\Lambda}$, и $\bar{0} \notin \vec{\Lambda}$.

Множество $\vec{\Lambda}$ выпуклое, т.к. $\forall \xi, \xi' > 0, \forall \lambda, \lambda' \in \vec{\Lambda}, \forall \alpha \in [0, 1]$

$$\alpha(\xi \lambda) + (1 - \alpha)(\xi' \lambda') \in \vec{\Lambda}. \quad (2.55)$$

Действительно, вектор в (2.55) равен $\xi'' \lambda''$, где

- $\xi'' = \alpha \xi + (1 - \alpha) \xi' > 0$, и
- $\lambda'' = \frac{\alpha \xi}{\xi''} \lambda + \frac{(1 - \alpha) \xi'}{\xi''} \lambda' \in [\lambda, \lambda']$, поэтому $\lambda'' \in \Lambda$.

Докажем, что $v = \sum_{i=1}^{m+1} v_i$ – внутренняя точка $\vec{\Lambda}$, т.е. $\exists \varepsilon > 0$: ε -окрестность $U_v^\varepsilon \stackrel{\text{def}}{=} \{v' \in \mathbf{R}^{m+1} \mid \|v' - v\| < \varepsilon\}$ вектора v лежит в $\vec{\Lambda}$.

Обозначим символом V множество, состоящее из всех векторов вида $\sum_{i=1}^{m+1} \alpha_i v_i$, где $\forall i = 1, \dots, m+1 \quad |\alpha_i| \leq \frac{1}{2}$. Для каждого вектора x из единичной сферы $S \stackrel{\text{def}}{=} \{x \in \mathbf{R}^{m+1} \mid \|x\| = 1\}$ обозначим записью ε_x максимальное число, такое, что $\varepsilon_x x \in V$ (такое число существует, т.к. множество V является ограниченным). Поскольку единичная

сфера в \mathbf{R}^{m+1} – компакт, то непрерывная функция $x \mapsto \varepsilon_x$ на S достигает наименьшего значения $\varepsilon > 0$ на некотором элементе S .

Итак, $\forall x \in S \ \varepsilon x \in V$, откуда следует, что $U_0^\varepsilon \subseteq V$, поэтому

$$U_v^\varepsilon = v + U_0^\varepsilon \subseteq v + V = \left\{ \sum_{i=1}^{m+1} \alpha_i v_i \mid \frac{1}{2} \leq \alpha_i \leq \frac{3}{2} \right\} \subseteq \vec{\Lambda}.$$

Следовательно, каждый элемент $U_{-v}^\varepsilon = -v + U_0^\varepsilon$ не лежит в $\vec{\Lambda}$ (иначе некоторый отрезок из $\vec{\Lambda}$ с концами в U_{-v}^ε и U_v^ε содержал бы $\vec{0}$).

Обозначим записью $\vec{\Lambda}^c$ **замыкание** множества $\vec{\Lambda}$, т.е. множество, получаемое добавлением к $\vec{\Lambda}$ всех его предельных точек. Нетрудно доказать, что замыкание любого выпуклого множества является выпуклым множеством. В частности, множество $\vec{\Lambda}^c$ выпукло.

Поскольку $U_{-v}^\varepsilon \cap \vec{\Lambda} = \emptyset$, то $-v$ не является предельной точкой множества $\vec{\Lambda}$, поэтому $-v \notin \vec{\Lambda}^c$.

Определим u_0 как такой вектор, из $\vec{\Lambda}^c$, что

$$\rho(-v, u_0) = \min\{\rho(-v, u) \mid u \in \vec{\Lambda}^c\}.$$

где ρ – евклидово расстояние между векторами. Такой вектор существует. Действительно, обозначим

- символом R расстояние от $-v$ до какого-либо элемента $\vec{\Lambda}^c$, и
- записью B_{-v}^R замкнутый шар с центром в $-v$ радиуса R .

Поскольку множество B_{-v}^R замкнуто и ограничено, то непустое множество $\vec{\Lambda}^c \cap B_{-v}^R$ тоже замкнуто и ограничено, т.е. оно компактно. Нетрудно видеть, что

$$\min\{\rho(-v, u) \mid u \in \vec{\Lambda}^c\} = \min\{\rho(-v, u) \mid u \in \vec{\Lambda}^c \cap B_{-v}^R\}.$$

Существование точки u_0 , на которой функция $u \mapsto \rho(-v, u)$ на компактном множестве $\vec{\Lambda}^c \cap B_{-v}^R$ принимает наименьшее значение, обосновывается теми же методами, которыми обосновывается аналогичное утверждение в теореме из параграфа 2.1.

Также аналогичными методами обосновывается следующее утверждение: гиперплоскость P , проходящая через точку u_0 и ортогональная отрезку $[-v, u_0]$, делит пространство \mathbf{R}^{m+1} на два полупространства, одно из которых (обозначим его L) имеет вид

$$L = \{x \in \mathbf{R}^{m+1} \mid x = u_0 \text{ или } \widehat{-vu_0x} \geq \frac{\pi}{2}\}$$

и $\vec{\Lambda}^c \subseteq L$.

Докажем, что $\bar{0} \in P$. Пусть $\bar{0} \notin P$. Тогда $\bar{0} \neq u_0 \in P$. Т.к. $\bar{0} \in \vec{\Lambda}^c$ ($\bar{0}$ – предельная точка $\vec{\Lambda}$), то $\widehat{-vu_0\bar{0}} \geq \frac{\pi}{2}$. Поскольку $u_1 \stackrel{\text{def}}{=} 2u_0 \in \vec{\Lambda}^c$, то $\widehat{-vu_0u_1} \geq \frac{\pi}{2}$. Однако углы $(-vu_0\bar{0})$ и $(-vu_0u_1)$ являются смежными, сумма их величин равна π , что возможно только если $\widehat{-vu_0\bar{0}} = \frac{\pi}{2}$, поэтому $\bar{0} \in P$.

Вектор, определяемый отрезком $[-v, u_0]$, т.е. $u_0 - (-v) = u_0 + v$, ортогонален P и принадлежит L , поэтому

$$L = \{x \in \mathbf{R}^{m+1} \mid \langle u_0 + v, x \rangle \geq 0\}$$

и в качестве искомого вектора λ^* можно взять $u_0 + v$. ■

Продолжим доказательство теоремы Каруша-Куна-Таккера.

Пусть вектор λ^* , обладающий свойством (2.54), имеет вид $(\lambda_0^*, \dots, \lambda_m^*)$.

1. Докажем, что $\forall i = 0, \dots, m \ \lambda_i^* \geq 0$.

Поскольку Λ содержит вектора

$$\lambda^{(0)} = (1, 0, \dots, 0) \quad \text{и} \quad \lambda^{(i)} = (\varepsilon, 0, \dots, 0, 1, 0, \dots, 0) \quad (\forall i = 1, \dots, m)$$

где $\varepsilon > 0$, и единица в $\lambda^{(i)}$ стоит в позиции номер i (мы считаем, что первая позиция в $\lambda^{(i)}$ имеет номер 0), то из (2.54) следует, что

$$\begin{aligned} \langle \lambda^*, \lambda^{(0)} \rangle &= \lambda_0^* \geq 0 \\ \forall i = 1, \dots, m \quad \langle \lambda^*, \lambda^{(i)} \rangle &= \lambda_0^* \varepsilon + \lambda_i^* \geq 0 \end{aligned}$$

откуда, ввиду произвольности ε , заключаем, что $\lambda_i^* \geq 0$.

2. Докажем, что

$$\forall i \in \{1, \dots, m\} \quad \lambda_i^* g_i(\hat{x}) = 0. \quad (2.56)$$

Т.к. $\forall \varepsilon > 0$ множество Λ содержит вектор

$$\lambda \stackrel{\text{def}}{=} (\varepsilon, 0, \dots, 0, g_i(\hat{x}), 0, \dots, 0),$$

то из (2.54) следует, что $\langle \lambda^*, \lambda \rangle = \lambda_0^* \varepsilon + \lambda_i^* g_i(\hat{x}) \geq 0$, откуда, ввиду произвольности ε , заключаем, что $\lambda_i^* g_i(\hat{x}) \geq 0$.

Таким образом, либо $\lambda_i^* g_i(\hat{x}) = 0$, либо $\lambda_i^* g_i(\hat{x}) > 0$. Однако если бы было верно неравенство $\lambda_i^* g_i(\hat{x}) > 0$, то, учитывая доказанное в предыдущем пункте неравенство $\lambda_i^* \geq 0$, заключаем, что $g_i(\hat{x}) > 0$, что противоречит условию $g_i(\hat{x}) \leq 0$.

3. Докажем, что $\forall x \in \mathbf{R}^n$

$$\lambda_0^* f(\hat{x}) + \sum_{i=1}^m \lambda_i^* g_i(\hat{x}) \leq \lambda_0^* f(x) + \sum_{i=1}^m \lambda_i^* g_i(x). \quad (2.57)$$

Из (2.56) следует, что левая часть (2.57) равна $\lambda_0^* f(\hat{x})$.

Т.к. $\forall \varepsilon > 0$ множество Λ содержит вектор

$$\lambda \stackrel{\text{def}}{=} (f(x) - f(\hat{x}) + \varepsilon, g_1(x), \dots, g_m(x)),$$

то из (2.54) следует, что

$$\langle \lambda^*, \lambda \rangle = \lambda_0^* (f(x) - f(\hat{x}) + \varepsilon) + \sum_{i=1}^m \lambda_i^* g_i(x) \geq 0,$$

откуда, ввиду произвольности ε , следует (2.57).

4. Докажем, что $\lambda_0^* \neq 0$.

Если $\lambda_0^* = 0$, то из $\lambda^* \neq \bar{0}$ следует, что $\exists i \in \{1, \dots, m\} : \lambda_i^* > 0$.

По предположению, для некоторого $x \in \mathbf{R}$ выполнено условие (2.44).

Для этого x верно также (2.57), однако в данном случае левая часть (2.57) равна 0, а правая часть (2.57) меньше 0, что невозможно.

Искомые значения $\hat{\lambda}_1, \dots, \hat{\lambda}_m$ определяются как $\frac{\lambda_1^*}{\lambda_0^*}, \dots, \frac{\lambda_m^*}{\lambda_0^*}$.

Истинность каждого из трех соотношений в (2.47) следует из соответствующих пунктов изложенного выше рассуждения.

Докажем, что если функции f, g_1, \dots, g_m дифференцируемые, то третье соотношение в (2.47) равносильно соотношению (2.48), т.е.

$$L(\hat{x}, \hat{\lambda}) = \min_{x \in \mathbf{R}^n} L(x, \hat{\lambda}) \Leftrightarrow \forall i = 1, \dots, m \quad \frac{\partial L}{\partial x_i}(\hat{x}, \hat{\lambda}) = 0, \quad (2.58)$$

где $\hat{\lambda} = (\hat{\lambda}_1, \dots, \hat{\lambda}_m)$.

Импликация « \Rightarrow » в (2.58) верна потому, что ее правая часть является необходимым условием минимума дифференцируемой функции.

Обоснуем импликацию « \Leftarrow » в (2.58). Из выпуклости f, g_1, \dots, g_m и неотрицательности $\hat{\lambda}_1, \dots, \hat{\lambda}_m$ следует выпуклость функции $x \mapsto L(x, \hat{\lambda})$:

$$\begin{aligned} \forall x, x' \in \mathbf{R}^n, \forall \alpha \in [0, 1] \quad & L(\alpha x + (1 - \alpha)x', \hat{\lambda}) = \\ & = f(\alpha x + (1 - \alpha)x') + \sum_{i=1}^m \hat{\lambda}_i g_i(\alpha x + (1 - \alpha)x') \leq \\ & \leq \alpha f(x) + (1 - \alpha)f(x') + \sum_{i=1}^m \hat{\lambda}_i (\alpha g_i(x) + (1 - \alpha)g_i(x')) = \\ & = \alpha L(x, \hat{\lambda}) + (1 - \alpha)L(x', \hat{\lambda}), \end{aligned}$$

откуда на основании такого же рассуждения, которое приведено в конце пункта 2.3.2, следует, что значение аргумента \hat{x} , удовлетворяющее правой части (2.58), является глобальным минимумом этой функции. ■

Применение метода

Применим изложенный выше метод к оптимизационной задаче (2.42) при условиях (2.41). В данном случае

- целевая функция f имеет вид $\frac{\|w\|^2}{2}$, и
- условия являются линейными неравенствами

$$y_x(\langle x, w \rangle - w_0) - 1 \geq 0, \quad \text{где } x \in X_S. \quad (2.59)$$

Докажем, что целевая функция выпукла. Данная функция является суперпозицией трех функций: функции $w \mapsto \|w\|$, функции $x \mapsto x^2$, и функции $x \mapsto \frac{1}{2}x$. Эти функции выпуклы, т.к.

- выпуклость функции $w \mapsto \|w\|$ следует из неравенства треугольника ($\|a + b\| \leq \|a\| + \|b\|$) для нормы в произвольном векторном пространстве: $\forall w, w' \in \mathbf{R}^n, \forall \alpha \in [0, 1]$

$$\|\alpha w + (1 - \alpha)w'\| \leq \|\alpha w\| + \|(1 - \alpha)w'\| = \alpha\|w\| + (1 - \alpha)\|w'\|,$$

- выпуклость функции $x \mapsto x^2$ обосновывается следующим образом: $\forall x, x' \in \mathbf{R}, \forall \alpha \in [0, 1]$ требуемое неравенство

$$(\alpha x + (1 - \alpha)x')^2 \leq \alpha x^2 + (1 - \alpha)(x')^2$$

после раскрытия скобок, перегруппировки слагаемых и приведения подобных членов преобразуется в эквивалентное неравенство

$$\alpha^2(x - x')^2 \leq \alpha(x - x')^2$$

которое верно потому, что $\alpha \in [0, 1]$,

- функция $x \mapsto \frac{1}{2}x$ выпукла потому, что любая линейная функция является выпуклой.

Нетрудно доказать, что если функции $f : \mathbf{R}^n \rightarrow \mathbf{R}$ и $g : \mathbf{R} \rightarrow \mathbf{R}$ выпуклы и, кроме того, g монотонно неубывающая, то их суперпозиция $(g \circ f)$ тоже выпукла. Действительно, $\forall x, x' \in \mathbf{R}^n, \forall \alpha \in [0, 1]$

$$\begin{aligned} (g \circ f)(\alpha x + (1 - \alpha)x') &= g(f(\alpha x + (1 - \alpha)x')) \leq \\ &\leq g(\alpha f(x) + (1 - \alpha)f(x')) \leq \alpha g(f(x)) + (1 - \alpha)g(f(x')) = \\ &= \alpha(g \circ f)(x) + (1 - \alpha)(g \circ f)(x'). \end{aligned}$$

Поскольку функции $x \mapsto x^2$ и $x \mapsto \frac{1}{2}x : \mathbf{R} \rightarrow \mathbf{R}$ – выпуклы и монотонно неубывающие, то, следовательно их суперпозиция с выпуклой функцией $w \mapsto \|w\|$, т.е. функция $w \mapsto \frac{1}{2}\|w\|^2$ тоже выпукла. ■

Функция Лагранжа для данной задачи имеет вид

$$L = \frac{\|w\|^2}{2} - \sum_{x \in X_S} \lambda_x (y_x(\langle x, w \rangle - w_0) - 1), \quad (2.60)$$

и соотношение (2.48) имеет следующий вид:

$$\forall i = 1, \dots, n \quad \hat{w}_i - \sum_{x \in X_S} \hat{\lambda}_x y_x x_i = 0, \quad 0 - \sum_{x \in X_S} \hat{\lambda}_x y_x (-1) = 0,$$

что можно переписать в виде

$$\hat{w} = \sum_{x \in X_S} \hat{\lambda}_x y_x x, \quad \sum_{x \in X_S} \hat{\lambda}_x y_x = 0. \quad (2.61)$$

Из теоремы 3 следует, что исходная задача сводится к задаче поиска вектора \hat{w} , числа \hat{w}_0 , и набора чисел $\hat{\lambda} = \{\hat{\lambda}_x \geq 0 \mid x \in X_S\}$, удовлетворяющих соотношениям в (2.61) и условию

$$\forall x \in X_S \quad \begin{cases} y_x(\langle x, \hat{w} \rangle - \hat{w}_0) - 1 \geq 0 \\ \hat{\lambda}_x (y_x(\langle x, \hat{w} \rangle - \hat{w}_0) - 1) = 0. \end{cases} \quad (2.62)$$

Теорема 4.

Задача нахождения объектов \hat{w} , \hat{w}_0 , и $\hat{\lambda} = \{\hat{\lambda}_x \geq 0 \mid x \in X_S\}$, удовлетворяющих соотношениям (2.61) и (2.62), сводится к задаче нахождения объектов \hat{w} , \hat{w}_0 , и $\hat{\lambda}$, минимизирующих значения выражения

$$\sum_{x \in X_S} \hat{\lambda}_x (y_x(\langle x, \hat{w} \rangle - \hat{w}_0) - 1) \quad (2.63)$$

при условиях

$$\begin{cases} \hat{w} = \sum_{x \in X_S} \hat{\lambda}_x y_x x, & \sum_{x \in X_S} \hat{\lambda}_x y_x = 0, \\ \forall x \in X_S \quad \begin{cases} \hat{\lambda}_x \geq 0 \\ y_x(\langle x, \hat{w} \rangle - \hat{w}_0) - 1 \geq 0. \end{cases} \end{cases} \quad (2.64)$$

Доказательство.

1. Пусть объекты \hat{w} , \hat{w}_0 , и $\hat{\lambda} = \{\hat{\lambda}_x \geq 0 \mid x \in X_S\}$ удовлетворяют соотношениям (2.61) и (2.62). Тогда при их подстановке вместо соответствующих объектов в (2.63) и (2.64) получаем, что

- значение суммы (2.63) будет равно 0 (т.к., согласно второму равенству в (2.62), каждое слагаемое в этой сумме равно 0), и
- соотношения в (2.64) верны, это следует из (2.61) и (2.62).

С другой стороны, сумма (2.63) при условиях (2.64), не может быть меньше 0, т.к., согласно этим условиям, каждое ее слагаемое является произведением неотрицательных чисел.

Таким образом, объекты \hat{w} , \hat{w}_0 , и $\hat{\lambda} = \{\hat{\lambda}_x \geq 0 \mid x \in X_S\}$ – решение задачи минимизации суммы (2.63) при условиях (2.64).

2. Согласно условиям (2.64), каждое слагаемое в сумме (2.63) при этих условиях неотрицательно, т.е. сумма (2.63) неотрицательна, и

- если минимальное значение этой суммы равно 0, то каждое слагаемое в этой сумме равно 0, т.е. объекты \hat{w} , \hat{w}_0 , и $\hat{\lambda}$, решающие задачу минимизации (2.63) при условиях (2.64), удовлетворяют соотношениям (2.61) и (2.62), и
- если минимальное значение этой суммы больше 0, то тогда решение задачи нахождения \hat{w} , \hat{w}_0 , и $\hat{\lambda} = \{\hat{\lambda}_x \geq 0 \mid x \in X_S\}$, удовлетворяющих соотношениям (2.61) и (2.62), не существует (что по предположению невозможно). ■

Перепишем сумму (2.63) путем раскрытия скобок, перегруппировки слагаемых и использования линейности скалярного произведения:

$$\begin{aligned}
& \sum_{x \in X_S} \hat{\lambda}_x y_x \langle x, \hat{w} \rangle - \sum_{x \in X_S} \hat{\lambda}_x y_x \hat{w}_0 - \sum_{x \in X_S} \hat{\lambda}_x = \\
& = \sum_{x \in X_S} \langle \hat{\lambda}_x y_x x, \hat{w} \rangle - \left(\sum_{x \in X_S} \hat{\lambda}_x y_x \right) \hat{w}_0 - \sum_{x \in X_S} \hat{\lambda}_x = \\
& = \left\langle \sum_{x \in X_S} \hat{\lambda}_x y_x x, \hat{w} \right\rangle - \left(\sum_{x \in X_S} \hat{\lambda}_x y_x \right) \hat{w}_0 - \sum_{x \in X_S} \hat{\lambda}_x.
\end{aligned} \tag{2.65}$$

Из условий (2.64) следует, что (2.65) можно переписать в виде

$$\langle \hat{w}, \hat{w} \rangle - \sum_{x \in X_S} \hat{\lambda}_x = \|\hat{w}\|^2 - \sum_{x \in X_S} \hat{\lambda}_x. \tag{2.66}$$

Выражение (2.66) можно переписать, используя лишь переменные $\hat{\lambda}_x$:

$$\sum_{x, x' \in X_S} \hat{\lambda}_x \hat{\lambda}_{x'} y_x y_{x'} \langle x, x' \rangle - \sum_{x \in X_S} \hat{\lambda}_x. \tag{2.67}$$

Таким образом, исходная задача свелась к задаче нахождения набора

$$\hat{\lambda} = \{\hat{\lambda}_x \mid x \in X_S\}$$

минимизирующего значение выражения (2.67), при условиях

$$\forall x \in X_S \quad \hat{\lambda}_x \geq 0, \quad \sum_{x \in X_S} \hat{\lambda}_x y_x = 0.$$

Такая задача называется **задачей квадратичного программирования (ЗКП)**. Существует много алгоритмов решения этой задачи.

Искомый вектор \hat{w} вычисляется по найденному решению $\hat{\lambda}$ данной ЗКП согласно первому равенству в (2.64). Для вычисления искомого \hat{w}_0 выбирается такая пара $x \in X_S$, что $\hat{\lambda}_x \neq 0$, в этом случае, согласно второму равенству в (2.62), должно быть верно равенство

$$y_x(\langle x, \hat{w} \rangle - \hat{w}_0) - 1 = 0,$$

из которого следует, что

$$\hat{w}_0 = \langle x, \hat{w} \rangle - y_x.$$

Если данная ЗКП имеет не единственное решение, то среди всех этих решений выбирается такое, что число \hat{w}_0 , вычисленное по этому решению, удовлетворяет последнему неравенству в (2.64).

Обоснуем, почему $\exists x \in X_S : \hat{\lambda}_x \neq 0$. Если бы все числа $\hat{\lambda}_x$ были равны 0, то \hat{w} – нулевой вектор, и из последнего неравенства в (2.64) следует, что $\forall x \in X_S \quad -y_x \hat{w}_0 - 1 \geq 0$, или $-y_x \hat{w}_0 \geq 1$. Выборка S предполагается нетривиальной, т.е. y_x м.б. равно как 1, так и -1 , откуда следует, что $\hat{w}_0 \geq 1$ и $-\hat{w}_0 \geq 1$, что невозможно. ■

2.5.3 Построение оптимальной разделяющей гиперплоскости по зашумленной выборке

В некоторых случаях выборка $S \subseteq \mathbf{R}^n \times \{-1, 1\}$ бывает зашумленной, т.е. значения компонентов векторов в S , представляющих объекты, могут немного отличаться от истинных значений. Также в S м.б. и ошибочно размеченные пары, т.е. пары (x, y_x) с неверным значением y_x . Это может привести к тому, что выборка S не будет линейно разделимой, т.е. невозможно найти АФ a_S вида

$$a_S(x) = \text{sign}\left(\sum_{i=1}^n x^i w_i - w_0\right) \quad (2.68)$$

такую, что $Q(a_S) = 0$. Тем не менее, иногда в ситуации зашумленности и линейной неразделимости выборки S все равно имеет смысл искать АФ a_S вида (2.68), допуская при этом небольшие ошибки классификации.

Один из подходов к построению АФ a_S такого вида называется **методом наименьших квадратов**, он будет рассмотрен в следующем параграфе. Данный подход заключается в том, чтобы найти АФ a_S вида (2.68) с наименьшим значением $Q(a_S)$, где

$$Q(a_S) = \sum_{x \in X_S} (a_S(x) - y_x)^2.$$

Другой подход является развитием подхода, изложенного в предыдущем пункте. Он заключается в том, чтобы построить разделяющую полосу максимальной ширины, допуская при этом попадание некоторых объектов не в свое полупространство.

Напомним, что если границами разделяющей полосы являются гиперплоскости, определяемые уравнениями

$$\langle x, w \rangle - w_0 = 1, \quad \langle x, w \rangle - w_0 = -1,$$

то $\forall x \in X_S$ точка x попадает в свое полупространство, если $M_x \geq 1$, где $M_x \stackrel{\text{def}}{=} y_x(\langle x, w \rangle - w_0)$.

Для адекватного определения целевой функции введем набор ξ дополнительных переменных:

$$\xi = \{\xi_x \mid x \in X_S\},$$

где $\forall x \in X_S$ значение переменной ξ_x будем понимать как величину ошибки аппроксимации на объекте x : она должна примерно соответствовать расстоянию от x до своего полупространства, когда x находится не в своем полупространстве (т.е. $M_x < 1$). Например, величину ошибки в этом случае можно определить как $1 - M_x$.

Задача построения оптимальной разделяющей полосы в данной ситуации имеет вид минимизации целевой функции

$$\frac{\|w\|^2}{2} + c \sum_{x \in X_S} \xi_x \quad (\text{где } c - \text{некоторая константа}) \quad (2.69)$$

при условиях

$$\forall x \in X_S \quad \xi_x \geq 1 - M_x, \quad \xi_x \geq 0$$

Целевая функция (2.69) выражает следующие требования:

- разделяющая полоса должна быть пошире (т.е. значение $\|w\|$ должно быть поменьше),
- суммарная ошибка $\sum_{x \in X_S} \xi_x$ должна быть поменьше.

Константа c позволяет регулировать баланс между требованиями

- максимизации ширины разделяющей полосы, и
- минимизации суммарной ошибки.

Обычно ее выбирают методом скользящего контроля:

- сначала задача решается при некотором c ,
- затем из выборки удаляется небольшая доля объектов, имеющих наибольшую величину ошибки (такие объекты будут считаться или чрезмерно зашумленными, или ошибочно размеченными), а c изменяется следующим образом:
 - если ширина полосы получилась слишком маленькая, то c уменьшается (это приведет к увеличению ширины полосы),
 - если слишком много объектов находятся далеко от своего полупространства, то c увеличивается (полоса станет уже), и
- после этого задача решается заново (с новыми S и c).

Возможно, придется проделать несколько таких итераций.

Функция Лагранжа для данной задачи имеет вид

$$L = \frac{\|w\|^2}{2} + c \sum_{x \in X_S} \xi_x - \sum_{x \in X_S} \lambda_x (M_x + \xi_x - 1) - \sum_{x \in X_S} \eta_x \xi_x,$$

и соотношение (2.48) имеет следующий вид:

- $\forall i = 1, \dots, n \quad \frac{\partial L}{\partial w_i}(\hat{w}, \hat{\lambda}, \hat{\xi}) = \hat{w}_i - \sum_{x \in X_S} \hat{\lambda}_x y_x x_i = 0,$
- $\frac{\partial L}{\partial w_0}(\hat{w}, \hat{\lambda}, \hat{\xi}) = \sum_{x \in X_S} \hat{\lambda}_x y_x = 0,$
- $\forall x \in X_S \quad \frac{\partial L}{\partial \xi_x}(\hat{w}, \hat{\lambda}, \hat{\xi}) = c - \hat{\lambda}_x - \hat{\eta}_x = 0,$

что можно переписать в виде

$$\begin{cases} \hat{w} = \sum_{x \in X_S} \hat{\lambda}_x y_x x, \\ \sum_{x \in X_S} \hat{\lambda}_x y_x = 0, \\ \forall x \in X_S \quad \hat{\lambda}_x + \hat{\eta}_x = c. \end{cases} \quad (2.70)$$

Из теоремы Каруша-Куна-Таккера следует, что исходная задача сводится к задаче поиска наборов чисел

$$\hat{\lambda} = \{\hat{\lambda}_x \geq 0 \mid x \in X_S\}, \quad \hat{\eta} = \{\hat{\eta}_x \geq 0 \mid x \in X_S\},$$

а также вектора \hat{w} , числа \hat{w}_0 , и набора чисел $\{\hat{\xi}_x \mid x \in X_S\}$, удовлетворяющих равенствам (2.70) и условию

$$\forall x \in X_S \quad \begin{cases} \hat{\lambda}_x(y_x(\langle x, \hat{w} \rangle - \hat{w}_0) + \hat{\xi}_x - 1) = 0 \\ \hat{\eta}_x \hat{\xi}_x = 0 \\ y_x(\langle x, \hat{w} \rangle - \hat{w}_0) + \hat{\xi}_x - 1 \geq 0 \\ \hat{\xi}_x \geq 0 \end{cases} \quad (2.71)$$

Как и в предыдущем пункте, нахождение искомых наборов $\hat{\lambda}$ и $\hat{\eta}$ сводится к задаче минимизации выражения

$$\sum_{x \in X_S} \hat{\lambda}_x(y_x(\langle x, \hat{w} \rangle - \hat{w}_0) + \hat{\xi}_x - 1) + \sum_{x \in X_S} \hat{\eta}_x \hat{\xi}_x$$

которое, с учетом условий (2.70), можно переписать, используя лишь переменные $\hat{\lambda}_x$ и $\hat{\xi}_x$:

$$\sum_{x, x' \in X_S} \hat{\lambda}_x \hat{\lambda}_{x'} y_x y_{x'} \langle x, x' \rangle - \sum_{x \in X_S} \hat{\lambda}_x + c \sum_{x \in X_S} \hat{\xi}_x. \quad (2.72)$$

Нетрудно установить, что исходная задача, рассматриваемая в настоящем пункте, сводится к задаче минимизации (2.72) с условиями

$$\begin{cases} \sum_{x \in X_S} \hat{\lambda}_x y_x = 0, \\ \forall x \in X_S \quad \begin{cases} 0 \leq \hat{\lambda}_x \leq c, \quad \hat{\xi}_x \geq 0, \\ M_x + \hat{\xi}_x - 1 \geq 0, \end{cases} \end{cases} \quad (2.73)$$

где $M_x \stackrel{\text{def}}{=} y_x(\langle x, \hat{w} \rangle - \hat{w}_0)$, $\hat{w} \stackrel{\text{def}}{=} \sum_{(x, y_x) \in S} \hat{\lambda}_x y_x x$.

Эта задача – тоже ЗКП. В результате ее решения получаются оптимальные $\hat{\lambda}$, $\hat{\xi}$, \hat{w}_0 , по которым можно вычислить оптимальные \hat{w} и $\hat{\eta}$.

Отметим некоторые особенности оптимального решения $\hat{\lambda}$, $\hat{\eta}$, $\hat{\xi}$, \hat{w} , \hat{w}_0 исходной задачи. $\forall x \in X_S$ вектор x относится к одному из трех классов, в зависимости от значения M_x :

- если $M_x > 1$, то x находится в своем полупространстве, но не на его границе (такой вектор называется **периферийным**),

- если $M_x = 1$, то x находится в своем полупространстве, на его границе (такой вектор называется **опорным**),
- если $M_x < 1$, то x находится за пределами своего полупространства (такой вектор называется **выбросом**).

Согласно соотношениям в (2.71) и (2.73), $\forall x \in X_S$

- либо $\hat{\lambda}_x = 0$, $\hat{\eta}_x = c$, либо $\hat{\xi}_x = 1 - M_x$,
- либо $\hat{\lambda}_x = c$, $\hat{\eta}_x = 0$, либо $\hat{\xi}_x = 0$,
- $M_x + \hat{\xi}_x - 1 \geq 0$.

Поэтому

- если x – периферийный, т.е. $M_x > 1$, то невозможно, чтобы было верно равенство $\hat{\xi}_x = 1 - M_x$, поэтому для такого вектора $\hat{\lambda}_x = 0$, $\hat{\eta}_x = c$, откуда следует, что $\hat{\xi}_x = 0$,
- если $0 < \hat{\lambda}_x < c$, то $0 < \hat{\eta}_x < c$, $\hat{\xi}_x = 0$ и $M_x = 1$, т.е. x опорный,
- если x – выброс, т.е. $M_x < 1$, то
 - невозможно, чтобы было верно равенство $\hat{\lambda}_x = 0$, т.к. если бы оно было верно, то тогда было бы верно равенство $\hat{\xi}_x = 0$, поэтому $M_x - 1 \geq 0$, что противоречит неравенству $M_x < 1$,
 - невозможно, чтобы было верно неравенство $0 < \hat{\lambda}_x < c$, т.к. в этом случае $\hat{\xi}_x = 0$, что, как было установлено выше, неверно,
 т.е. остается единственная возможность: $\hat{\lambda}_x = c$, $\hat{\eta}_x = 0$, и $\hat{\xi}_x > 0$.

Читателю предлагается самостоятельно исследовать вопрос: возможно ли, чтобы x был опорный, но $\lambda_x = 0$ или c (если невозможно, то доказать это, а если возможно, то привести соответствующий пример).

2.6 Ядерный метод машинного обучения

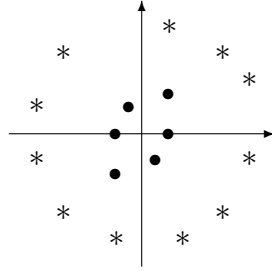
2.6.1 Спрямяющие пространства

В некоторых случаях выборка S линейно неразделима

- не по причинам зашумленности компонентов векторов в S , представляющих объекты, или ошибочной разметки,

- а по причине наличия нетривиальных зависимостей между компонентами векторов в S , представляющих объекты.

Например, рассмотрим выборку $S \subseteq \mathbf{R}^2 \times \{-1, 1\}$, изображенную на картинке (в которой мы представляем вектора из \mathbf{R}^2 точками плоскости)



где черные кружочки изображают объекты из S^+ , а звездочки – объекты из S^- . Такая выборка принципиально линейно неразделима.

В данном случае принадлежность вектора $(x_1, x_2) \in \mathbf{R}^2$ к S^+ или S^- лучше определять по его норме $\|x\| = \sqrt{x_1^2 + x_2^2}$, т.е. искать a_S в виде

$$a_S(x) = \text{sign}(\|x\| - w_0).$$

Поиск АФ, соответствующих таким выборкам $S \subseteq \mathbf{R}^n \times \{-1, 1\}$, для которых невозможно построить АФ a_S в виде

$$a_S(x) = \text{sign}\left(\sum_{i=1}^n x^i w_i - w_0\right), \quad (2.74)$$

может делаться методом **спрямляющих пространств**, который заключается в следующем:

- выбирается подходящее нелинейное отображение

$$\varphi : \mathbf{R}^n \rightarrow \mathbf{R}^N, \quad (2.75)$$

с таким расчетом, чтобы выборка $S_\varphi \stackrel{\text{def}}{=} \{(\varphi(x), y_x) \mid x \in X_S\}$ оказалась бы линейно разделимой,

(\mathbf{R}^N в (2.75) называется **спрямляющим пространством**)

- методом опорных векторов для S_φ ищется АФ a_{S_φ} вида (2.74), и
- искомая АФ a_S определяется как композиция $a_{S_\varphi} \circ \varphi$.

Например, для описанного выше примера в качестве спрямляющего пространства можно взять \mathbf{R}^2 , и $\varphi(x_1, x_2) \stackrel{\text{def}}{=} (x_1^2, x_2^2)$.

В других ситуациях рассматриваются функции $\varphi : \mathbf{R}^n \rightarrow \mathbf{R}^N$ вида

$$\varphi(x_1, \dots, x_n) = (x_1, \dots, x_n, x_1^2, \dots, x_n^2, \dots, x_i x_j, \dots).$$

Нетрудно доказать, что для любой выборки S существует спрямляющее пространство – это м.б., например, пространство \mathbf{R}^N , где N – число элементов в S . Если выбрать в \mathbf{R}^N ортонормированный базис, и определить $\varphi : \mathbf{R}^n \rightarrow \mathbf{R}^N$ как инъективное отображение, сопоставляющее каждому элементу $x \in X_S$ некоторый вектор из этого базиса, то S_φ будет линейно разделима.

Отметим важную особенность функции a_{S_φ} . Если S_φ линейно разделима, то, как было показано выше, a_{S_φ} можно искать в виде

$$\begin{aligned} a_{S_\varphi}(x) &= \text{sign}(\langle \varphi(x), \hat{w} \rangle - \hat{w}_0) = \\ &= \text{sign}(\langle \varphi(x), \sum_{x' \in X_S} \hat{\lambda}_{x'} y_{x'} \varphi(x') \rangle - \hat{w}_0) = \\ &= \text{sign}(\sum_{x' \in X_S} \hat{\lambda}_{x'} y_{x'} \langle \varphi(x), \varphi(x') \rangle - \hat{w}_0), \end{aligned} \quad (2.76)$$

где $\hat{\lambda}$ ищется как решение задачи минимизации выражения

$$\sum_{x, x' \in X_S} \hat{\lambda}_x \hat{\lambda}_{x'} y_x y_{x'} \langle \varphi(x), \varphi(x') \rangle - \sum_{x \in X_S} \hat{\lambda}_x \quad (2.77)$$

при условиях

$$\forall x \in X_S \quad \hat{\lambda}_x \geq 0, \quad \sum_{x \in X_S} \hat{\lambda}_x y_x = 0. \quad (2.78)$$

Из (2.76) и (2.77) следует, что для построения a_{S_φ} не надо знать

- ни размерности спрямляющего пространства,
- ни векторов вида $\varphi(x)$, где $x \in X_S$,

надо лишь знать скалярные произведения вида $\langle \varphi(x), \varphi(x') \rangle$.

Ядерный метод ML заключается в том, что построение АФ a_S делается методом, аналогичным описанному выше, но с заменой всех скалярных произведений $\langle \varphi(x), \varphi(x') \rangle$ на выражения вида $K(x, x')$, где K – функция вида

$$K : \mathbf{R}^n \times \mathbf{R}^n \rightarrow \mathbf{R}.$$

Данная функция называется **ядром**. Она определяется для S методом подбора, и должна обладать следующими свойствами:

- **симметричность:**

$$\forall x, x' \in \mathbf{R}^n \quad K(x, x') = K(x', x), \quad (2.79)$$

- **положительная определенность:**

$$\forall x_1, \dots, x_m \in \mathbf{R}^n, \quad \forall \alpha_1, \dots, \alpha_m \in \mathbf{R} \quad \sum_{i,j=1}^m \alpha_i \alpha_j K(x_i, x_j) \geq 0. \quad (2.80)$$

Построение АФ a_S сводится к задаче минимизации выражения

$$\sum_{x,x' \in S} \hat{\lambda}_x \hat{\lambda}_{x'} y_x y_{x'} K(x, x') - \sum_{x \in X_S} \hat{\lambda}_x, \quad (2.81)$$

при условиях (2.78). Искомая АФ a_{S_φ} определяется следующим образом:

$$\forall x \in \mathbf{R}^n \quad a_{S_\varphi}(x) = \text{sign} \left(\sum_{x' \in X_S} \hat{\lambda}_{x'} y_{x'} K(x, x') - \hat{w}_0 \right),$$

где $\hat{w}_0 \stackrel{\text{def}}{=} \sum_{x' \in X_S} \hat{\lambda}_{x'} y_{x'} K(x'', x') - y_{x''}$, и x'' – произвольный вектор из X_S , такой, что $\hat{\lambda}_{x''} \neq 0$.

В общем случае ядро определяется как произвольная функция вида

$$K : X \times X \rightarrow \mathbf{R} \quad (2.82)$$

(где X – множество, элементами которого являются объекты), являющаяся симметричной и положительно определенной, т.е. верны аналоги соотношений (2.79) и (2.80), в которых аргументами K являются элементы множества X .

Нетрудно доказать, что если K – ядро вида (2.82), то

- $\forall x \in X \quad K(x, x) \geq 0$,
- если $\forall x \in X \quad K(x, x) = 0$, то $\forall x, x' \in X \quad K(x, x') = 0$,
- верен аналог неравенства Коши-Буняковского:

$$\forall x, x' \in X \quad K(x, x') \leq \sqrt{K(x, x)K(x', x')}.$$

2.6.2 Примеры ядер

$K(x, x')$ может иметь, например, следующий вид:

- $(c\langle x, x' \rangle + d)^m$, где $c, d \in \mathbf{R}_{\geq 0}$, $m \geq 0$ (полиномиальное ядро),
(где $\mathbf{R}_{\geq 0}$ – множество неотрицательных действительных чисел),
- $\sigma(c\langle x, x' \rangle + d)$, где σ – сигмоида (сигмоидное ядро),
- $a_0 + \sum_{n \geq 0} a_n \sin(nx) \sin(nx') + \sum_{n \geq 0} b_n \cos(nx) \cos(nx')$, где $\sum_{n \geq 1} a_n^2 + b_n^2 < \infty$
(ядро Фурье),
- $\exp(-\frac{\|x-x'\|^2}{\sigma^2})$ (гауссово ядро),
- $\exp(\frac{\langle x, x' \rangle}{\sigma^2})$ (экспоненциальное ядро),
- $K_1(x, x')K_2(x, x')$, где K_1 и K_2 – ядра,
- $\alpha_1 K_1(x, x') + \alpha_2 K_2(x, x')$, где K_1 и K_2 – ядра, α_1 и $\alpha_2 \in \mathbf{R}_{\geq 0}$,
- $K(\varphi(x), \varphi(x'))$, где K – ядро, φ – произвольная функция.

Приведем еще три примера ядра.

1. В качестве ядра $X \times X \rightarrow \mathbf{R}$ может выступать функция вида

$$(x, x') \mapsto \langle \varphi(x), \varphi(x') \rangle$$

где φ – произвольная функция вида

$$\varphi : X \rightarrow \mathcal{H},$$

и \mathcal{H} – спрямляющее пространство, которое является **гильбертовым пространством**, т.е.

- \mathcal{H} – линейное пространство со скалярным произведением, напомним, что **скалярное произведение** на \mathcal{H} – это функция, сопоставляющая каждой паре $(x, y) \in \mathcal{H} \times \mathcal{H}$ действительное число $\langle x, y \rangle$, и обладающая свойствами: $\forall x, x', y \in \mathcal{H}$
 - $\forall \alpha, \alpha' \in \mathbf{R} \quad \langle \alpha x + \alpha' x', y \rangle = \alpha \langle x, y \rangle + \alpha' \langle x', y \rangle$,
 - $\langle x, y \rangle = \langle y, x \rangle$,
 - $\langle x, x \rangle \geq 0$, причем $\langle x, x \rangle = 0 \Leftrightarrow x = \bar{0}$ (нулевой вектор),

- \mathcal{H} – **полное** метрическое пространство (где метрика ρ на \mathcal{H} определяется нормой: $\rho(x, x') = \|x - x'\|$, а норма определяется скалярным произведением: $\|x\|^2 = \langle x, x \rangle$), т.е. \forall фундаментальная последовательность имеет предел в \mathcal{H} ,
- \mathcal{H} – **сепарабельное** пространство, т.е. \exists счетное подмножество $H \subseteq \mathcal{H}$: $\forall \varepsilon > 0, \forall h \in \mathcal{H} \exists h' \in H : \rho(h, h') < \varepsilon$.

Примеры гильбертовых пространств: \mathbf{R}^n или l_2 , где

$$l_2 = \{(x_1, x_2, \dots) \mid \sum_{i \geq 1} x_i^2 < \infty\},$$

скалярное произведение в l_2 имеет вид

$$\langle (x_1, x_2, \dots), (y_1, y_2, \dots) \rangle = \sum_{i \geq 1} x_i y_i.$$

2. Если K – ядро вида $X \times X \rightarrow \mathbf{R}$, где X – произвольное множество, то функция

$$K' : \mathcal{P}_{fin}(X) \times \mathcal{P}_{fin}(X) \rightarrow \mathbf{R}$$

(где $\mathcal{P}_{fin}(X)$ – множество всех конечных подмножеств X), определяется следующим образом:

$$\forall A, B \in \mathcal{P}_{fin}(X) \quad K'(A, B) = \sum_{a \in A, b \in B} K(a, b)$$

тоже является ядром.

3. Пусть заданы алфавит (т.е. множество символов) A , целое число $n \geq 1$, и действительное число $\lambda \in (0, 1]$.

Обозначим записью n^* множество всех последовательностей

$$\bar{i} = (i_1, \dots, i_k), \quad \text{где } 1 \leq i_1 < \dots < i_k \leq n. \quad (2.83)$$

Если последовательность $\bar{i} \in n^*$ имеет вид (2.83), то

- запись $l(\bar{i})$ обозначает число $i_k - i_1 + 1$, и
- для каждой цепочки $x \in A^n$, если x имеет вид $a_1 \dots a_n$, то запись $x[\bar{i}]$ обозначает цепочку $a_{i_1} \dots a_{i_k}$.

Обозначим символом \mathcal{H} гильбертово пространство функций вида

$$f : A^n \rightarrow \mathbf{R},$$

скалярное произведение на \mathcal{H} определяется следующим образом:

$$\langle f, g \rangle = \sum_{y \in A^n} f(y)g(y).$$

Ядро $K_n : A^n \times A^n \rightarrow \mathbf{R}$ определяется соотношением

$$\forall x, x' \in A^n \quad K_n(x, x') = \langle \varphi(x), \varphi(x') \rangle,$$

где $\varphi : A^n \rightarrow \mathcal{H}$, $\forall x \in A^n \quad \varphi(x) : A^n \rightarrow \mathbf{R}, y \mapsto \sum_{\bar{i} \in n^* : y=x[\bar{i}]} \lambda^{l(\bar{i})}$.

K_n используется в задачах анализа естественно-языковых текстов.

2.6.3 Каноническое гильбертово пространство, определяемое ядром

Каждому ядру $K : X^2 \rightarrow \mathbf{R}$ соответствует **каноническое гильбертово пространство**, определяемое ядром K . Данное пространство обозначается записью \mathcal{H}_K и обладает следующим свойством: $\exists \varphi : X \rightarrow \mathcal{H}_K$:

$$\forall x, x' \in X \quad K(x, x') = \langle \varphi(x), \varphi(x') \rangle. \quad (2.84)$$

\mathcal{H}_K определяется как подпространство линейного пространства \mathbf{R}^X функций вида $X \rightarrow \mathbf{R}$ следующим образом:

- $\forall x \in X$ обозначим записью K_x функцию из \mathbf{R}^X , которая отображает каждый $x' \in X$ в $K(x, x')$,
- обозначим символом \mathcal{F}_X линейное подпространство в \mathbf{R}^X , порожденное всеми функциями вида K_x , т.е.

$$\mathcal{F}_X = \left\{ \sum_{i=1..n} \alpha_i K_{x_i} \mid n \geq 1, \forall i = 1, \dots, n \quad \alpha_i \in \mathbf{R}, x_i \in X \right\},$$

- определим скалярное произведение на \mathcal{F}_X : если f, g из \mathcal{F}_X имеют вид $f = \sum_{i=1..n} \alpha_i K_{x_i}$ и $g = \sum_{j=1..m} \beta_j K_{x'_j}$, то

$$\langle f, g \rangle \stackrel{\text{def}}{=} \sum_{i=1..n, j=1..m} \alpha_i \beta_j K(x_i, x'_j) = \sum_{i=1..n} \alpha_i g(x_i) = \sum_{j=1..m} \beta_j f(x'_j),$$

заметим, что из этого определения следует свойство

$$\forall x, x' \in X \quad \langle K_x, K_{x'} \rangle = K(x, x') = K_x(x') = K_{x'}(x), \quad (2.85)$$

- \mathcal{H}_K определяется как пополнение \mathcal{F}_X в метрике ρ , задаваемой скалярным произведением:

$$\forall f, g \in \mathcal{F}_X \quad \rho(f, g) = \|f - g\| = \sqrt{\langle f - g, f - g \rangle},$$

- скалярное произведение на \mathcal{H}_K определяется следующим образом: $\forall f, g \in \mathcal{H}_K$, если $f = \lim_{n \rightarrow \infty} f_n$ и $g = \lim_{n \rightarrow \infty} g_n$, где $\forall n \geq 1 \ f_n, g_n \in \mathcal{F}_X$, то

$$\langle f, g \rangle \stackrel{\text{def}}{=} \lim_{n \rightarrow \infty} \langle f_n, g_n \rangle$$

(нетрудно доказать, что данный предел существует и не зависит от выбора последовательностей (f_n) и (g_n) , сходящихся к f и g соответственно).

Функция $\varphi : X \rightarrow \mathcal{H}_K$ определяется как отображение $x \mapsto K_x$. Свойство (2.84) следует из определения функции φ и соотношения (2.85).

Нетрудно доказать, что

- $\forall f \in \mathcal{H}_K, \forall x \in X$

$$\begin{aligned} f(x) &= \langle f, K_x \rangle, \\ |f(x)| &\leq \|f\| \sqrt{K(x, x)}, \end{aligned} \tag{2.86}$$

неравенство в (2.86) верно потому, что оно эквивалентно неравенству Коши-Буняковского

$$|\langle f, K_x \rangle| \leq \|f\| \sqrt{\langle K_x, K_x \rangle},$$

- $\forall x \in X$ функционал $\mathcal{H}_K \rightarrow \mathbf{R} : f \mapsto f(x)$ непрерывен, т.е.

$$f = \lim_{n \rightarrow \infty} f_n \quad \Rightarrow \quad f(x) = \lim_{n \rightarrow \infty} f_n(x),$$

это можно обосновать с использованием свойства

$$\forall f, f' \in \mathcal{H}_K, \forall x \in X \quad |f(x) - f'(x)| \leq \|f - f'\| \sqrt{K(x, x)},$$

которое следует из неравенства в (2.86). ■

Теорема 5.

Пусть заданы

- выборка $S \subseteq X \times \mathbf{R}$,

- ядро $K : X^2 \rightarrow \mathbf{R}$,
- функция потерь $c : (X \times \mathbf{R} \times \mathbf{R})^* \rightarrow \mathbf{R}_{\geq 0}$
(например, $c((x_i, y_i, f(x_i))_{i=1..l}) = \frac{1}{l} \sum_{i=1}^l (y_i - f(x_i))^2$), и
- строго возрастающая функция $\Omega : \mathbf{R}_{\geq 0} \rightarrow \mathbf{R}_{\geq 0}$.

Если $\hat{f} \in \mathcal{H}_K$ – решение задачи

$$c((x_i, y_i, f(x_i))_{i=1..l}) + \Omega(\|f\|) \rightarrow \min \quad (2.87)$$

то $\hat{f} = \sum_{i=1}^l \alpha_i K_{x_i}$, где $\alpha_1, \dots, \alpha_l \in \mathbf{R}$.

Доказательство.

Вектор $\hat{f} \in \mathcal{H}_K$ можно представить в виде суммы

$$\hat{f} = \sum_{j=1}^l \alpha_j K_{x_j} + f^*,$$

где f^* – вектор из ортогонального дополнения подпространства в \mathcal{H}_K , порожденного векторами K_{x_1}, \dots, K_{x_l} , т.е.

$$\forall i = 1, \dots, l \quad \langle f^*, K_{x_i} \rangle = 0. \quad (2.88)$$

Согласно равенству в (2.86) и соотношению (2.88), $\forall i = 1, \dots, l$

$$\hat{f}(x_i) = \langle \hat{f}, K_{x_i} \rangle = \left\langle \sum_{j=1}^l \alpha_j K_{x_j}, K_{x_i} \right\rangle + \langle f^*, K_{x_i} \rangle = \sum_{j=1}^l \alpha_j K(x_j, x_i),$$

откуда следует, что $\hat{f}(x_i)$ не зависит от f^* , поэтому первое слагаемое в сумме в (2.87) не зависит от f^* .

Докажем, что $f^* = \bar{0}$ (нулевой вектор). Если $f^* \neq \bar{0}$, то замена f^* на $\bar{0}$ не изменила бы первое слагаемое в сумме в (2.87), а второе бы только уменьшилось, т.к. $\forall f_1, f_2 \in \mathcal{H}_K$ если $\langle f_1, f_2 \rangle = 0$, то

$$\|f_1 + f_2\|^2 = \langle f_1 + f_2, f_1 + f_2 \rangle = \langle f_1, f_1 \rangle + \langle f_2, f_2 \rangle = \|f_1\|^2 + \|f_2\|^2$$

поэтому $\|\hat{f}\|^2 = \left\| \sum_{i=1}^l \alpha_i K_{x_i} \right\|^2 + \|f^*\|^2 > \left\| \sum_{i=1}^l \alpha_i K_{x_i} \right\|^2$, откуда следует неравенство

$$\Omega(\|\hat{f}\|) > \Omega\left(\left\| \sum_{i=1}^l \alpha_i K_{x_i} \right\|\right)$$

которое противоречит предположению о том, что \hat{f} – решение задачи (2.87). ■

2.7 Задача регрессии

В задаче регрессии выборка, как правило, имеет вид

$$S = \{(x_1, y_1), \dots, (x_l, y_l)\} \subseteq \mathbf{R}^n \times \mathbf{R},$$

т.е. ответами являются действительные числа (м.б. также и вектора из действительных чисел). Задача регрессии заключается в том, чтобы по известным значениям y_i некоторой неизвестной функции в заданных точках x_i ($i = 1, \dots, l$) предсказать значения этой функции в других точках.

2.7.1 Линейная регрессия

Пусть задана выборка $S = \{(x_1, y_1), \dots, (x_l, y_l)\} \subseteq \mathbf{R}^n \times \mathbf{R}$.

Требуется найти АФ $a_S(x)$ в виде $\langle x, w \rangle + w_0$, где $w \in \mathbf{R}^n$, $w_0 \in \mathbf{R}$, причем должно быть выполнено условие

$$Q(a_S) = \frac{1}{l} \sum_{i=1}^l \mathcal{L}(a_S, x_i) = \frac{1}{l} \sum_{i=1}^l (a_S(x_i) - y_i)^2 \rightarrow \min.$$

Поскольку множитель $\frac{1}{l}$ не влияет на решение задачи поиска оптимальной АФ a_S , то мы его писать не будем.

Заметим, что

$$a_S(x_i) = \langle x_i, w \rangle + w_0 = \langle \tilde{x}_i, \tilde{w} \rangle,$$

где $\tilde{w} = (w, w_0)$ и $\tilde{x}_i = (x_i, 1)$.

Таким образом, требуется найти $\tilde{w} \in \mathbf{R}^{n+1}$, решающий задачу

$$Q(\tilde{w}) \stackrel{\text{def}}{=} \sum_{i=1}^l (\langle \tilde{x}_i, \tilde{w} \rangle - y_i)^2 \rightarrow \min \quad (2.89)$$

Обозначим символами X и Y матрицу и столбец соответственно

$$X = \begin{pmatrix} \tilde{x}_1 \\ \dots \\ \tilde{x}_l \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ \dots \\ y_l \end{pmatrix}.$$

Если рассматривать \tilde{w} как столбец, то можно переписать (2.89) в виде

$$Q(\tilde{w}) = \|X\tilde{w} - Y\|^2 \rightarrow \min \quad (2.90)$$

$Q(\tilde{w})$ – выпуклая функция, т.к. она является суперпозицией выпуклой функции $\tilde{w} \mapsto \|X\tilde{w} - Y\|$ и выпуклой функции $x \mapsto x^2$, причем вторая

функция возрастает. Выпуклость второй из этих функций была обоснована в пункте 2.5.2, выпуклость первой обосновывается следующим образом: $\forall \alpha \in [0, 1], \forall \tilde{w}, \tilde{w}' \in \mathbf{R}^{n+1}$

$$\begin{aligned} & \|X(\alpha\tilde{w} + (1-\alpha)\tilde{w}') - Y\| = \\ & = \|X(\alpha\tilde{w} + (1-\alpha)\tilde{w}') - \alpha Y - (1-\alpha)Y\| = \\ & = \|\alpha(X\tilde{w} - Y) + (1-\alpha)(X\tilde{w}' - Y)\| \leq \\ & \leq \alpha\|X\tilde{w} - Y\| + (1-\alpha)\|X\tilde{w}' - Y\|. \end{aligned}$$

Поэтому \tilde{w} является решением задачи (2.90) если и только если

$$\forall j = 1, \dots, n+1 \quad \frac{\partial Q}{\partial \tilde{w}_j} = 0. \quad (2.91)$$

Учитывая определение функции Q , перепишем (2.91) в виде

$$\forall j = 1, \dots, n+1 \quad \frac{\partial Q}{\partial \tilde{w}_j} = \sum_{i=1}^l 2(\langle \tilde{x}_i, \tilde{w} \rangle - y_i)(\tilde{x}_i)_j = 0 \quad (2.92)$$

Поскольку $\forall i = 1, \dots, l, \forall j = 1, \dots, n+1$ $(\tilde{x}_i)_j$ есть элемент X_{ij} матрицы X , то соотношения (2.92) можно переписать в виде

$$\forall j = 1, \dots, n+1 \quad \sum_{i=1}^l \langle \tilde{x}_i, \tilde{w} \rangle X_{ij} = \sum_{i=1}^l y_i X_{ij}. \quad (2.93)$$

Обозначим записью X^\top матрицу, транспонированную к X , т.е.

$$\forall j = 1, \dots, n+1, \forall i = 1, \dots, l \quad X_{ji}^\top = X_{ij},$$

и перепишем (2.93) в виде

$$\forall j = 1, \dots, n+1 \quad \sum_{i=1}^l X_{ji}^\top \langle \tilde{x}_i, \tilde{w} \rangle = \sum_{i=1}^l X_{ji}^\top y_i. \quad (2.94)$$

$\forall i = 1, \dots, l$ $\langle \tilde{x}_i, \tilde{w} \rangle$ есть i -й элемент столбца $X\tilde{w}$, обозначим его записью $(X\tilde{w})_i$. Нетрудно видеть, что $\forall j = 1, \dots, n+1$

- левая сумма в (2.94), т.е. $\sum_{i=1}^l X_{ji}^\top (X\tilde{w})_i$, есть j -й элемент столбца $X^\top(X\tilde{w})$, и
- правая сумма в (2.94) есть j -й элемент столбца $X^\top Y$.

Таким образом, столбцы $X^T X \tilde{w}$ и $X^T Y$ совпадают, т.е. искомый вектор \tilde{w} является решением уравнения

$$X^T X \tilde{w} = X^T Y. \quad (2.95)$$

Нетрудно доказать, что данное уравнение всегда имеет решение.

Если матрица $X^T X$ невырождена, то уравнение (2.95) имеет единственное решение

$$\tilde{w} = (X^T X)^{-1} X^T Y,$$

а если она вырождена, то решение уравнения (2.95) м.б. неединственным. Так может произойти, например, если число l пар в обучающей выборке S меньше n . В этом случае вместо задачи (2.90) разумнее решать задачу

$$Q(\tilde{w}) = \|X\tilde{w} - Y\|^2 + c\|\tilde{w}\|^2 \rightarrow \min, \quad (2.96)$$

где c – параметр, выражающий штраф за большую $\|\tilde{w}\|$.

Данная задача называется задачей **гребневой (ridge) регрессии**.

Поскольку минимизируемая функция Q в (2.96) является выпуклой и дифференцируемой, то искомое решение должно обращать в 0 все частные производные функции Q :

$$\forall j = 1, \dots, n+1 \quad \frac{\partial Q}{\partial \tilde{w}_j} = \sum_{i=1}^l 2(\langle \tilde{x}_i, \tilde{w} \rangle - y_i) \tilde{x}_{i,j} + 2c\tilde{w}_j = 0.$$

Это соотношение можно переписать в виде

$$\forall j = 1, \dots, n+1 \quad \sum_{i=1}^l (\langle \tilde{x}_i, \tilde{w} \rangle - y_i) X_{ij} + c\tilde{w}_j = 0. \quad (2.97)$$

Нетрудно доказать, что $\forall j = 1, \dots, n+1$

- первое слагаемое в (2.97), т.е. $\sum_{i=1}^l (\langle \tilde{x}_i, \tilde{w} \rangle - y_i) X_{ij}$, есть j -й элемент столбца $X^T X \tilde{w} - X^T Y$, и
- слагаемое $c\tilde{w}_j$ в (2.97) есть j -й элемент столбца $c\tilde{w}$.

Поэтому искомый вектор \tilde{w} является решением уравнения

$$X^T X \tilde{w} - X^T Y + c\tilde{w} = 0^\downarrow \quad (\text{где } 0^\downarrow \text{ – нулевой столбец порядка } n+1),$$

которое можно переписать в виде

$$(X^T X + cE)\tilde{w} = X^T Y \quad (\text{где } E \text{ – единичная матрица порядка } n+1).$$

Нетрудно доказать, что матрица $X^T X + cE$ невырождена $\forall c > 0$.

Таким образом, решение задачи гребневой регрессии имеет вид

$$\tilde{w} = (X^T X + cE)^{-1} X^T Y.$$

2.8 Метрическая модель обучения

Метрическая модель обучения связана с использованием некоторой меры близости ρ на множестве объектов X , которую называют **метрикой**. Метрика ρ должна быть подобрана так, чтобы зависимость $f : X \rightarrow Y$ между объектами и ответами (которую аппроксимирует АФ a_S) была согласована с ρ , т.е. если объекты x и x' близки по этой метрике, то ответы $f(x)$ и $f(x')$ были бы примерно одинаковы.

Во многих задачах метрический подход существенно проще, чем рассмотренный выше подход, основанный на описании объектов в виде векторов значений признаков. Применение метрической модели более предпочтительно по сравнению с другими моделями в тех случаях, когда объекты имеют сложную структуру (например, это м.б. изображения, временные ряды, структуры белков, и т.п.).

2.8.1 Понятие метрики

Метрикой на множестве X называется функция

$$\rho : X \times X \rightarrow \mathbf{R}_{\geq 0},$$

удовлетворяющая условию: $\forall x, x' \in X \quad \begin{cases} \rho(x, x') = 0 & \Leftrightarrow x = x', \\ \rho(x, x') = \rho(x', x). \end{cases}$

В некоторых случаях ρ может также удовлетворять условию

$$\forall x, x', x'' \in X \quad \rho(x, x'') \leq \rho(x, x') + \rho(x', x'')$$

(которое называется **неравенством треугольника**).

Примеры метрики:

- **евклидова метрика** на $X = \mathbf{R}^n$:

$$\forall x, x' \in \mathbf{R}^n \quad \rho(x, x') = \sqrt{\sum_{i=1}^n (x_i - x'_i)^2},$$

где $x = (x_1, \dots, x_n)$, $x' = (x'_1, \dots, x'_n)$,

- **взвешенная метрика** на $X = \mathbf{R}^n$: предполагается, что заданы веса $w_1, \dots, w_n \in \mathbf{R}_{\geq 0}$ и действительное число $p \geq 1$,

$$\forall x, x' \in \mathbf{R}^n \quad \rho(x, x') = \left(\sum_{i=1}^n w_i |x_i - x'_i|^p \right)^{\frac{1}{p}},$$

где $x = (x_1, \dots, x_n), x' = (x'_1, \dots, x'_n)$,
 (евклидова метрика является частным случаем взвешенной метрики: в ней $p = 2$ и $\forall i = 1, \dots, n \ w_i = 1$),

- метрика на множестве символьных строк C^* , где C – конечное множество, элементы которого называются **символами**: $\forall x, x' \in C^*$ $\rho(x, x')$ равно наименьшему числу элементарных операций, которые надо выполнить чтобы получить x' из x , где под **элементарной операцией** понимается
 - удаление какого-либо символа из строки, или
 - вставка какого-либо символа в произвольное место строки.

2.8.2 Метод ближайших соседей

Пусть задана обучающая выборка $S \subseteq X \times Y$.

Напомним, что X_S обозначает множество объектов, входящих в S :

$$X_S \stackrel{\text{def}}{=} \{x \in X \mid \exists y_x \in Y : (x, y_x) \in S\}.$$

Если на множестве X объектов задана метрика ρ , то $\forall x \in X$ объекты из множества X_S можно упорядочить в соответствии с их близостью к x , т.е. расположить в последовательность

$$x_1, \dots, x_{|S|}, \quad (2.98)$$

удовлетворяющую условию:

$$\rho(x, x_1) \leq \dots \leq \rho(x, x_{|S|}) \quad (2.99)$$

т.е. первым в (2.98) расположен ближайший к x объект, затем – следующий по близости к x объект, и т.д. $\forall i = 1, \dots, l$ объект x_i из последовательности (2.98) называется i -м **ближайшим соседом к x** .

Метод ближайших соседей для построения АФ a_S заключается в том, что выбираются

- натуральное число $k \geq 1$ (число ближайших к x объектов, ответы на которых учитываются при вычислении ответа $a_S(x)$),
- действительные числа $w_1 \geq \dots \geq w_k > 0$, которые имеют смысл весов, определяющих вклад ближайших k соседей объекта x из обучающей выборки S в вычисление ответа для объекта x , например,

$$\forall i = 1, \dots, k \quad w_i = \frac{k+1-i}{k} \quad \text{или} \quad w_i = q^i,$$

где $0 < q < 1$ – заданный параметр,

$\forall x \in X$ значение $a_S(x)$ определяется как такой ответ $y \in Y$, который максимизирует значение выражения

$$\sum_{i=1}^k \llbracket y_{x_i} = y \rrbracket w_i, \quad (2.100)$$

т.е. как такой ответ y , который наиболее характерен среди ответов на k ближайших соседей x из обучающей выборки S .

АФ, построенную в соответствии с приведенным выше определением, будем обозначать записью a_S^k , явно указывая число k ближайших соседей. Оптимальным является такое k , которое минимизирует риск

$$\sum_{(x, y_x) \in S} \llbracket a_{S \setminus \{(x, y_x)\}}^k(x) \neq y_x \rrbracket.$$

2.8.3 Метод окна Парзена

В методе окна Парзена (МОП) используется

- параметр $h > 0$, называемый **шириной окна**, и
- невозрастающая функция $K(r) : \mathbf{R}_{\geq 0} \rightarrow \mathbf{R}_{\geq 0}$, называемая **ядром**.

Примеры ядер, используемых в МОП: $\llbracket r \leq 1 \rrbracket$, $(1 - r^2)\llbracket r \leq 1 \rrbracket$, e^{-r^2} .

АФ a_S , построенная по МОП, определяется почти так же, как в предыдущем пункте, со следующим отличием: вместо выражения (2.100) используется выражение

$$\sum_{i=1}^{|S|} \llbracket y_{x_i} = y \rrbracket K\left(\frac{\rho(x, x_i)}{h}\right).$$

АФ, построенную в соответствии с приведенным выше определением, будем обозначать записью a_S^h , явно указывая ширину окна h . Оптимальным является такое h , которое минимизирует риск

$$\sum_{(x, y_x) \in S} \llbracket a_{S \setminus \{(x, y_x)\}}^h(x) \neq y_x \rrbracket.$$

Ширина окна h может быть не константой, а функцией, зависящей от количества объектов из X_S , находящихся вблизи x .

2.8.4 Метод потенциалов

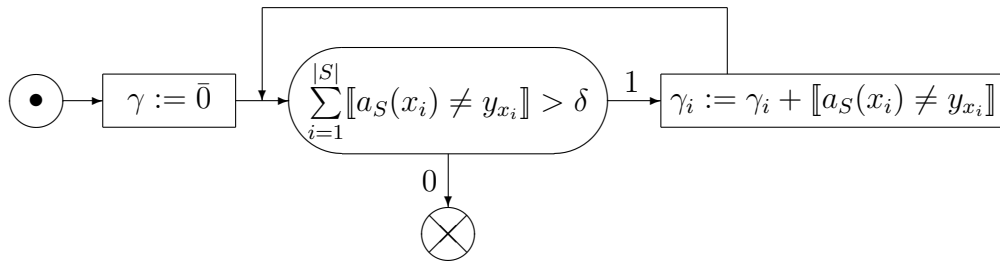
В методе потенциалов используются следующие параметры:

- размеры окон $h_1, \dots, h_{|S|} > 0$,
- порог ошибки $\delta \geq 0$, и
- ядро $K(r)$.

АФ a_S , построенная по методу потенциалов, определяется почти так же, как в пункте 2.8.2, со следующим отличием: вместо выражения (2.100) используется выражение

$$\sum_{i=1}^{|S|} \llbracket y_{x_i} = y \rrbracket \gamma_i K\left(\frac{\rho(x, x_i)}{h_i}\right),$$

в котором $\gamma_i \geq 0$ – веса, настраиваемые по следующему алгоритму:



где

- $\gamma = (\gamma_1, \dots, \gamma_{|S|})$,
- $\bar{0}$ – вектор размерности $|S|$, все компоненты которого равны 0, и
- при каждом выполнении оператора в правом прямоугольнике индекс i выбирается равновероятно из множества $\{1, \dots, |S|\}$.

При $Y = \{-1, +1\}$ можно понимать

- объекты из X_S как положительные и отрицательные заряды,
- коэффициенты γ_i как абсолютные величины этих зарядов,
- $K(r)$ как зависимость потенциала от расстояния до заряда,
- значение $a_S(x)$ как знак потенциала в точке x .

2.8.5 Метод эталонов

В этом пункте рассматривается задача построения по обучающей выборке S АФ a_S по следующему принципу:

- $\forall x \in X$ элементы множества X_S располагаются в последовательность $x_1, \dots, x_{|S|}$, удовлетворяющую условию (2.99), и
- значение $a_S(x)$ определяется как такой ответ $y \in Y$, который максимизирует значение выражения

$$(x \mapsto_S y) = \sum_{i=1}^{|S|} \llbracket y_{x_i} = y \rrbracket w(i, x), \quad (2.101)$$

где $\forall i = 1, \dots, |S|$ $w(i, x)$ – заданное число, выражающее степень важности i -го ближайшего соседа x (т.е. x_i) для вычисления $a_S(x)$.

Значение (2.101) можно интерпретировать как

- меру уверенности в том, что АФ a_S отображает x именно в y , или
- меру близости $a_S(x)$ к y .

$\forall x \in X_S$ сопоставим объекту x число $M_S(x)$, называемое **типичностью** объекта x , и определяемое следующим образом:

$$M_S(x) = (x \mapsto_S y_x) - \max_{y \in Y \setminus y_x} (x \mapsto_S y).$$

Говоря неформально, $M_S(x)$ выражает меру правдоподобия утверждения о том, что значением $a_S(x)$ является именно y_x .

Каждому объекту $x \in X_S$ можно сопоставить один из четырех перечисляемых ниже типов, в соответствии со значением $M_S(x)$:

- x – **эталон**, если значение $M_S(x)$ – большое положительное,
- x – **периферийный** (или **неинформативный**) объект, если $M_S(x)$ – положительное, но не такое большое, как у эталонов,
- x – **пограничный** объект, если $M_S(x)$ близко к 0,
- x – **выброс** (т.е. зашумленный, или ошибочно размеченный объект), если $M_S(x) < 0$.

Для нахождения оптимальной АФ a_S рекомендуется строить её не по всей выборке S , а по ее подвыборке \hat{S} , содержащей только эталоны.

\hat{S} строится при помощи излагаемого ниже алгоритма, в котором используются следующие параметры:

- δ – порог фильтрации выбросов,
- l_0 – допустимая доля ошибок.

Алгоритм состоит из перечисляемых ниже действий.

- Из S удаляются все пары (x, y_x) , такие, что $M_S(x) < \delta$.
Ниже под S понимается не исходная выборка, а результат этого удаления.
- $\forall y \in Y$ в \hat{S} зачисляется такая пара $(x^*, y) \in S$, что значение $M_S(x^*)$ максимально среди $\{M_S(x) \mid x \in X_S, (x, y) \in S\}$.
- Далее выполняются итерации, состоящие из перечисляемых ниже действий. Итерации заканчиваются, если \hat{S} станет равно S .
 - $E := \{(x, y) \in S \setminus \hat{S} : M_{\hat{S}}(x) < 0\}$,
 - если $|E| < l_0$ то выход,
 - иначе $\hat{S} := \hat{S} \cup \{(x^*, y)\}$, где $(x^*, y) \in E$, и значение $M_{\hat{S}}(x^*)$ минимально среди $\{M_{\hat{S}}(x) \mid (x, y) \in E\}$.

2.9 Вероятностные модели обучения

2.9.1 Дискретная вероятностная модель обучения

В **дискретной вероятностной модели обучения (ДВМО)** предполагается, что множество X является конечным или счетным, на множестве $X \times Y$ задано **вероятностное распределение** (обычно называемое просто **распределением**), т.е. функция p вида

$$p : X \times Y \rightarrow [0, 1], \quad (2.102)$$

удовлетворяющая условию $\sum_{(x,y) \in X \times Y} p(x, y) = 1$.

Будем предполагать, что $\forall y \in Y \exists x \in X : p(x, y) > 0$.

$\forall (x, y) \in X \times Y$ значение $p(x, y)$ называется **вероятностью** появления пары (x, y) в обучающей выборке S .

Мы предполагаем, что все пары в S появляются независимо друг от друга.

Если выборка S большая, то число $p(x, y)$ можно понимать как приближительную

- долю тех пар в S , которые равны (x, y) , или

- частоту появления пары (x, y) в S .

$\forall X' \subseteq X, \forall y \in Y$ будем обозначать записью $p(X', y)$ значение

$$p(X', y) \stackrel{\text{def}}{=} \sum_{x \in X'} p(x, y). \quad (2.103)$$

Это значение можно понимать как приблизительную частоту появления в обучающей выборке объекта из X' с ответом y .

Если $X' = X$, то значение (2.103) обозначается записью $p(y)$, и понимается как приблизительная частота появления в обучающей выборке объекта с ответом y .

Отметим, что ДВМО концептуально отличается от рассмотренных выше моделей обучения:

- в рассмотренных выше моделях обучения $\forall x \in X$ обучающая выборка не может содержать пар вида (x, y) и (x, y') , где $y \neq y'$, т.к. вторая компонента пары (x, y) – это истинный ответ на объект x , который определяется однозначно по x ,
- а в ДВМО нет понятия истинного ответа на объект, в ней любой ответ на объект может появиться с некоторой вероятностью.

Одной из компонентов ДВМО является **функция потерь**

$$\lambda : Y \times Y \rightarrow \mathbf{R}_{\geq 0}, \quad (2.104)$$

которая сопоставляет каждой паре $(y, y') \in Y \times Y$ **потерю** $\lambda_{yy'} \geq 0$, возникающую в том случае, когда на какой-либо объект дается ответ y' , в то время когда правильным ответом на этот объект был бы y .

2.9.2 Оптимальные аппроксимирующие функции

Пусть a – функция вида $a : X \rightarrow Y$.

$\forall y \in Y$ запись $a^{-1}(y)$ обозначает прообраз y относительно a :

$$a^{-1}(y) = \{x \in X \mid a(x) = y\}.$$

Риск, соответствующий функции $a : X \rightarrow Y$ определяется как среднее значение потери при использовании a в качестве АФ:

$$R(a) = \sum_{y, y' \in Y} \lambda_{yy'} p(a^{-1}(y'), y). \quad (2.105)$$

Если $\lambda_{yy'} = \llbracket y \neq y' \rrbracket$, то $R(a)$ можно интерпретировать как вероятность ошибки при использовании a в качестве АФ.

Теорема 6.

Пусть заданы распределение (2.102) и функция потерь (2.104).

Минимальное значение риска (2.105) достигается на функции a , сопоставляющей каждому $x \in X$ такой ответ $y_x \in Y$, который минимизирует значение выражения

$$\sum_{y \in Y} \lambda_{yy_x} p(x, y).$$

Доказательство.

Согласно определению значения $p(a^{-1}(y'), y)$,

$$R(a) = \sum_{y, y' \in Y} \lambda_{yy'} \sum_{x \in a^{-1}(y')} p(x, y) = \sum_{y' \in Y} \sum_{x \in a^{-1}(y')} \sum_{y \in Y} \lambda_{yy'} p(x, y). \quad (2.106)$$

Заметим, что сумма вида $\sum_{y' \in Y} \sum_{x \in a^{-1}(y')} \dots$ — это сумма вида $\sum_{x \in X} \dots$, где выражение, изображаемое многоточием во второй сумме, получается из выражения, изображаемого многоточием в первой сумме заменой всех вхождений y' на $a(x)$. Поэтому (2.106) можно переписать в виде

$$R(a) = \sum_{x \in X} \sum_{y \in Y} \lambda_{y, a(x)} p(x, y),$$

откуда видно, что $R(a)$ достигает минимального значения в том и только в том случае, когда $\forall x \in X$ сумма $\sum_{y \in Y} \lambda_{y, a(x)} p(x, y)$ достигает минимального значения, что по определению y_x возможно при $a(x) = y_x$. ■

Теорема 7.

Пусть заданы распределение (2.102) и функция потерь (2.104), причем

- $\forall y \in Y \quad \lambda_{yy} = 0$, и
- $\forall y, y' \in Y$, если $y \neq y'$, то значение $\lambda_{yy'}$ не зависит от y' (обозначим это значение λ_y).

Минимальное значение риска (2.105) достигается на функции a , сопоставляющей каждому $x \in X$ такой ответ $y_x \in Y$, который максимизирует значение выражения $\lambda_{y_x} p(x, y_x)$, т.е.

$$a(x) = \arg \max_{y \in Y} \lambda_y p(x, y). \quad (2.107)$$

Доказательство.

Согласно определению риска (2.105) и условиям теоремы,

$$\begin{aligned}
 R(a) &= \sum_{y, y' \in Y} \lambda_{yy'} p(a^{-1}(y'), y) = \sum_{y, y' \in Y, y \neq y'} \lambda_y p(a^{-1}(y'), y) = \\
 &= \sum_{y, y' \in Y} \lambda_y p(a^{-1}(y'), y) - \sum_{y \in Y} \lambda_y p(a^{-1}(y), y) = \\
 &= \sum_{y \in Y} \sum_{y' \in Y} \sum_{x \in a^{-1}(y')} \lambda_y p(x, y) - \sum_{y \in Y} \sum_{x \in a^{-1}(y)} \lambda_y p(x, y).
 \end{aligned} \tag{2.108}$$

Нетрудно видеть, что

- сумму $\sum_{y' \in Y} \sum_{x \in a^{-1}(y')} \lambda_y p(x, y)$ можно заменить на $\sum_{x \in X} \lambda_y p(x, y)$, и
- сумму $\sum_{y \in Y} \sum_{x \in a^{-1}(y)} \lambda_y p(x, y)$ можно заменить на $\sum_{x \in X} \lambda_{a(x)} p(x, a(x))$,

поэтому (2.108) можно переписать в виде

$$\begin{aligned}
 R(a) &= \sum_{y \in Y} \sum_{x \in X} \lambda_y p(x, y) - \sum_{x \in X} \lambda_{a(x)} p(x, a(x)) = \\
 &= \sum_{y \in Y} \lambda_y p(y) - \sum_{x \in X} \lambda_{a(x)} p(x, a(x)),
 \end{aligned}$$

откуда видно, что $R(a)$ достигает минимального значения в том и только в том случае, когда $\forall x \in X$ выражение $\lambda_{a(x)} p(x, a(x))$ достигает максимального значения, что по определению y_x возможно при $a(x) = y_x$. ■

Отметим, что функция (2.107), совпадает с функцией

$$a(x) = \arg \max_{y \in Y} \lambda_y p(y) p(x|y), \tag{2.109}$$

где $p(x|y) = \frac{p(x,y)}{p(y)}$ – вероятность появления объекта x в обучающей выборке, при условии, что ответом на этот объект является y . Значение $p(x|y)$ можно понимать как приблизительную долю (или частоту появления) пары (x, y) в подвыборке S_y , где

$$S_y = \{(x, y_x) \in S \mid y_x = y\}. \tag{2.110}$$

В ряде случаев вместо (2.109) более удобно использовать формулу

$$a(x) = \arg \max_{y \in Y} \ln \left(\lambda_y p(y) p(x|y) \right), \tag{2.111}$$

которая, как нетрудно видеть, определяет ту же функцию, что и (2.109).

2.9.3 Построение АФ по обучающей выборке

Пусть задана некоторая ДВМО, причем

- распределение (2.102) неизвестно, и
- функция потерь (2.104) известна и удовлетворяет условиям теоремы 7.

Также задана некоторая обучающая выборка $S \subseteq X \times Y$, построенная в соответствии с этой ДВМО.

Требуется построить a_S вида (2.109), где вместо неизвестных истинных значений вероятностей $p(y)$ и $p(x|y)$ должны быть использованы оценки $\hat{p}(y)$ и $\hat{p}(x|y)$ этих вероятностей, вычисленные по S .

В качестве оценки $\hat{p}(y)$ можно взять, например, $\frac{|S_y|}{|S|}$.

Для вычисления оценки $\hat{p}(x|y)$ обозначим записью x_1, \dots, x_m последовательность первых компонентов пар, входящих в S_y , и полагаем

$$\hat{p}(x|y) \stackrel{\text{def}}{=} \frac{1}{m} \sum_{i=1}^m \mathbb{I}[x_i = x].$$

2.9.4 Непрерывная вероятностная модель обучения

Во многих задачах машинного обучения множеством объектов X является \mathbf{R} или \mathbf{R}^n . Для данного случая можно определить аналог рассмотренного выше понятия ДВМО, который называется **непрерывной вероятностной моделью обучения (НВМО)**. В НВМО вместо понятия вероятностного распределения используется понятие **плотности вероятности** (обычно называемой просто **плотностью**), которая представляет собой функцию p вида

$$p : X \times Y \rightarrow [0, 1],$$

имеющую излагаемый ниже смысл. Для каждой пары $(x, y) \in X \times Y$ значение p на паре (x, y) будем обозначать записью $p(x|y)$.

Предполагается, что

- на множестве X задана мера Лебега μ ,
- $\forall y \in Y$ функция вида $X \rightarrow \mathbf{R}_{\geq 0}$, отображающая каждый $x \in X$ в $p(x|y)$, предполагается интегрируемой по Лебегу, и

$$\int_X p(x|y) d\mu = 1. \quad (2.112)$$

Для каждого измеримого подмножества $X' \subseteq X$ и каждого $y \in Y$ интеграл $\int_{X'} p(x|y)d\mu$ интерпретируется как вероятность появления в обучающей выборке S пары с первой компонентой из множества X' , при условии, что вторая компонента этой пары равна y . Так же, как и в ДВМО, мы предполагаем, что все пары в S появляются независимо друг от друга. Если выборка S большая, то число $\int_{X'} p(x|y)d\mu$ можно понимать как приближительную

- долю тех пар (x, y) в S_y (см. (2.110)), у которых $x \in X'$, или
- частоту появления в S_y пар с первой компонентой из X' .

Множество Y ответов предполагается конечным или счетным.

Одной из компонентов НВМО является распределение (обозначаемое тем же символом p) на Y . $\forall y \in Y$ число $p(y)$ имеет тот же смысл, что и в ДВМО.

Как и выше, в НВМО можно

- определить понятие риска, соответствующего функции $a : X \rightarrow Y$,
- и доказать, что если функция потерь (2.104) известна и удовлетворяет условиям теоремы 7, то оптимальная АФ имеет вид (2.109), где $p(x|y)$ понимается как плотность в точке x .

Таким образом, если

- задана некоторая НВМО, причем
 - плотность $p(x|y)$ либо неизвестна, либо известна лишь частично (например, известно, что она имеет вид $\varphi(x, \theta)$, где функция φ известна, а θ – неизвестный параметр), и
 - функция потерь (2.104) известна и удовлетворяет условиям теоремы 7,
- и задана некоторая обучающая выборка $S \subseteq X \times Y$, построенная в соответствии с этой НВМО,

то оптимальную АФ a_S можно искать в виде (2.109), где вместо неизвестных истинных значений вероятностей $p(y)$ и $p(x|y)$ используются оценки $\hat{p}(y)$ и $\hat{p}(x|y)$ этих вероятностей, вычисленные по S ,

Как и выше, в качестве оценки $\hat{p}(y)$ можно взять $\frac{|S_y|}{|S|}$.

Ниже излагаются различные варианты вычисления оценки $\hat{p}(x|y)$, в зависимости от вида множества объектов X и предположений о классе функций, в котором содержится функция $p(x|y) : X \rightarrow [0, 1]$.

Обозначим записью X_{S_y} множество первых компонентов пар, входящих в S_y . Элементы X_{S_y} будем обозначать записями x_1, \dots, x_m .

Рассмотрим различные виды, которые может иметь множество X .

1. $X = \mathbf{R}$.

- Согласно определению плотности, при небольшом числе $h > 0$ произведение $p(x) \cdot 2h$ приблизительно равно количеству точек из X_{S_y} , попавших в отрезок $[x - h, x + h]$. Поэтому $\hat{p}(x|y)$ можно определить как долю точек из X_{S_y} , лежащих внутри отрезка $[x - h, x + h]$:

$$\hat{p}(x|y) = \frac{1}{2mh} \sum_{i=1}^m \mathbb{I}[|x - x_i| < h]. \quad (2.113)$$

- Другой вид оценки $p(x|y)$ – оценка Парзена-Розенблатта:

$$\hat{p}(x|y) = \frac{1}{mh} \sum_{i=1}^m K\left(\frac{x - x_i}{h}\right) \quad (2.114)$$

где $K(z)$ – четная функция, называемая **ядром**, такая, что

$$\int_{-\infty}^{+\infty} K(z) dz = 1,$$

и h – параметр, называемый **шириной окна**, небольшое положительное число.

АФ, построенную в соответствии с приведенным выше определением, будем обозначать записью a_S^h , явно указывая ширину окна h . Оптимальным является такое h , которое минимизирует риск

$$\sum_{(x, y_x) \in S} \mathbb{I}[a_{S \setminus \{(x, y_x)\}}^h(x) \neq y_x]. \quad (2.115)$$

2. X – метрическое пространство, т.е. на X задана метрика ρ .

В этом случае м.б. использована следующая оценка $\hat{p}(x|y)$:

$$\hat{p}(x|y) = \frac{1}{mV(h)} \sum_{i=1}^m K\left(\frac{\rho(x, x_i)}{h}\right) \quad (2.116)$$

где K, h – параметры, имеющие тот же смысл, что и аналогичные параметры выше (ядро и ширина окна, соответственно), и $V(h)$ –

нормирующий множитель, предназначенный для того, чтобы (2.116) было плотностью, т.е. удовлетворяло условию (2.112).

Если распределение объектов в пространстве X сильно неравномерно, то лучше использовать переменную ширину окна, определяемую в каждой точке $x \in X$ как расстояние от x до $(k + 1)$ -го соседа, где оптимальное значение k м.б. найдено из условия, аналогичного условию (2.115).

3. $X = \mathbf{R}^n$.

- Если нет никаких предположений о том, какой вид может иметь плотность $p(x|y)$, то можно использовать оценку

$$\hat{p}(x|y) = \frac{1}{m} \sum_{i=1}^m \prod_{j=1}^n \frac{1}{h_j} K\left(\frac{x_{0j} - x_{ij}}{h_j}\right),$$

где

- $\forall i = 1, \dots, m \quad x_i = (x_{i1}, \dots, x_{in}), x = (x_{01}, \dots, x_{0n})$, и
- K, h_1, \dots, h_n – параметры, имеющие тот же смысл, что и аналогичные параметры в предыдущем пункте (т.е. K – ядро, и h_1, \dots, h_n – ширины окон, соответствующих каждому из признаков).

- Если известно, что $p(x|y)$ имеет гауссов вид

$$\mathcal{N}(x; \mu, \Sigma) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)} \quad (2.117)$$

(\top обозначает транспонирование),

где μ и Σ – неизвестные параметры:

- $\mu \in \mathbf{R}^n$,
- Σ – симметричная, невырожденная, положительно определенная матрица порядка n , называемая **ковариационной матрицей**,

то нахождение оценки $\hat{p}(x|y)$ сводится к нахождению оценок $\hat{\mu}$ и $\hat{\Sigma}$, которые м.б. вычислены по правилам

$$\hat{\mu} = \frac{1}{m} \sum_{i=1}^m x_i; \quad \hat{\Sigma} = \frac{1}{m-1} \sum_{i=1}^m (x_i - \hat{\mu})(x_i - \hat{\mu})^\top.$$

Некоторое обоснование данных правил заключается в том, что

- μ совпадает с мат. ожиданием Ex случайной величины x с плотностью (2.117), т.е. с интегралом $\int_{\mathbf{R}^n} x \mathcal{N}(x; \mu, \Sigma) dx$,
- Σ совпадает с мат. ожиданием $E(x - \mu)(x - \mu)^\top$.
- Если $\forall y \in Y$ $p(x|y)$ имеет вид (2.117), и известно, что ковариационные матрицы в (2.117) одинаковы для всех $y \in Y$, то оценка $\hat{\Sigma}$ этих матриц м.б. вычислена по формуле

$$\hat{\Sigma} = \frac{1}{|S| - |Y|} \sum_{(x, y_x) \in S} (x - \hat{\mu}_{y_x})(x - \hat{\mu}_{y_x})^\top.$$

В данном случае АФ, вычисленную по формуле (2.111), можно записать (опуская сложение с константой в $\arg \max_{y \in Y}(\dots)$) в виде

$$\begin{aligned} a(x) &= \arg \max_{y \in Y} \left(\ln(\lambda_y p(y)) - \frac{1}{2} (x - \hat{\mu}_y)^\top \hat{\Sigma}^{-1} (x - \hat{\mu}_y) \right) = \\ &= \arg \max_{y \in Y} \left(\ln(\lambda_y p(y)) - \frac{1}{2} \hat{\mu}_y^\top \hat{\Sigma}^{-1} \hat{\mu}_y + x^\top \hat{\Sigma}^{-1} \hat{\mu}_y \right) = \\ &= \arg \max_{y \in Y} (x^\top \alpha_y + \beta_y), \end{aligned} \quad (2.118)$$

где $\alpha_y = \hat{\Sigma}^{-1} \hat{\mu}_y$, $\beta_y = \ln(\lambda_y p(y)) - \frac{1}{2} \hat{\mu}_y^\top \hat{\Sigma}^{-1} \hat{\mu}_y$.

АФ (2.118) называется **линейным дискриминантом Фишера**.

После вычисления оценки $\hat{p}(x|y)$

- те объекты из S_y , для которых значение $\hat{p}(x|y)$ мало, рассматриваются как выбросы,
- соответствующие пары (x, y_x) удаляются из S ,
- после чего оценка $\hat{p}(x|y)$ перевычисляется.

2.9.5 ЕМ-алгоритм

ЕМ-алгоритм предназначен для вычисления оценки $\hat{p}(x|y)$, в предположении, что плотность $p(x|y)$ является **смесью** плотностей вида $\varphi(x, \theta_j)$, где $j = 1, \dots, k$, и функция φ предполагается известной, т.е.

$$p(x|y) = \sum_{j=1}^k w_j \varphi(x, \theta_j), \quad (2.119)$$

где $w_1, \dots, w_k, \theta_1, \dots, \theta_k$ – неизвестные параметры, причем

$$\forall j = 1, \dots, k \quad w_j \in \mathbf{R}_{\geq 0}, \quad \sum_{j=1}^k w_j = 1. \quad (2.120)$$

Обозначим символом Θ вектор параметров, входящих в (2.119), т.е.

$$\Theta = (w_1, \dots, w_k, \theta_1, \dots, \theta_k).$$

Нахождение оценки $\hat{p}(x|y)$ сводится к нахождению вектора $\hat{\Theta}$ оценок всех параметров, входящих в Θ .

Для решения данной задачи отдельно рассматриваются случаи, когда число k компонентов смеси известно, и когда это число неизвестно.

Число k компонентов смеси известно

В данном случае строится последовательность $\hat{\Theta}^{(1)}, \hat{\Theta}^{(2)}, \dots$ приближений к искомому вектору оценок параметров $\hat{\Theta}$. Алгоритм

- использует матрицу $G = (g_{ij})_{m \times k}$ скрытых (hidden) переменных, и
- заключается в итерационном повторении двух шагов:
 - **Е-шаг**: вычисление G по текущему приближению $\hat{\Theta}^{(s)}$,
 - **М-шаг**: вычисление следующего приближения $\hat{\Theta}^{(s+1)}$ по текущим матрице G и вектору $\hat{\Theta}^{(s)}$.

Одним из входных данных алгоритма является небольшое положительное число δ , используемое в критерии остановки.

Первое приближение $\hat{\Theta}^{(1)}$ выбирается произвольно (или исходя из каких-либо соображений), с условием (2.120).

s -я итерация (где $s \geq 1$) имеет вид:

Е-шаг (expectation) :

$$\forall i = 1, \dots, m, \quad \forall j = 1, \dots, k \quad g_{ij} = \frac{w_j \varphi(x_i, \theta_j)}{p(x_i)}, \quad \text{где}$$

- w_j и θ_j – компоненты текущего приближения $\hat{\Theta}^{(s)}$,
- $p(x_i) = \sum_{j=1}^k w_j \varphi(x_i, \theta_j)$.

Если $s \geq 2$, и $\max_{i,j} |g_{ij} - g'_{ij}| < \delta$, где g'_{ij} – соответствующая компонента матрицы G , вычисленной на предыдущей итерации, то алгоритм заканчивает свою работу, его результатом является $\hat{\Theta}^{(s)}$.

M-шаг (maximization) :

Целью данного шага является решение оптимизационной задачи

$$\prod_{i=1}^m p(x_i) \rightarrow \max_{\Theta},$$

(где $p(x_i) = \sum_{j=1}^k w_j \varphi(x_i, \theta_j)$) при ограничении $\sum_{j=1}^k w_j = 1$ (данная задача называется **задачей максимизации правдоподобия**), которая эквивалентна оптимизационной задаче

$$\ln \prod_{i=1}^m p(x_i) = \sum_{i=1}^m \ln p(x_i) \rightarrow \max_{\Theta}$$

при указанном выше ограничении.

Функция Лагранжа этой оптимизационной задачи имеет вид

$$L = \sum_{i=1}^m \ln p(x_i) - \lambda \left(\sum_{j=1}^k w_j - 1 \right).$$

Одним из необходимых условий оптимальности является соотношение

$$\frac{\partial L}{\partial w_j} = \sum_{i=1}^m \frac{\varphi(x_i, \theta_j)}{p(x_i)} - \lambda = 0, \quad \forall j = 1, \dots, k. \quad (2.121)$$

Из (2.121) следуют равенства

$$w_j \sum_{i=1}^m \frac{\varphi(x_i, \theta_j)}{p(x_i)} = w_j \lambda \quad (\forall j = 1, \dots, k), \quad (2.122)$$

просуммировав которые по j , получаем равенство

$$\sum_{j=1}^k w_j \sum_{i=1}^m \frac{\varphi(x_i, \theta_j)}{p(x_i)} = \sum_{j=1}^k w_j \lambda,$$

из которого следует равенство

$$\sum_{i=1}^m \sum_{j=1}^k w_j \frac{\varphi(x_i, \theta_j)}{p(x_i)} = \lambda \sum_{j=1}^k w_j = \lambda. \quad (2.123)$$

Левая часть (2.123) равна сумме m единиц, т.е. m , поэтому $\lambda = m$. Подставляя в (2.122) $\lambda = m$, и вспоминая определение g_{ij} , получаем равенства $\sum_{i=1}^m g_{ij} = w_j m \quad (\forall j = 1, \dots, k)$, т.е.

$$w_j = \frac{1}{m} \sum_{i=1}^m g_{ij}, \quad \forall j = 1, \dots, k. \quad (2.124)$$

Неравенства $w_j \geq 0$ будут выполнены на каждой итерации, т.к. они выполнены для $\hat{\Theta}^{(1)}$.

Другие необходимые условия оптимальности имеют вид

$$\begin{aligned} \frac{\partial L}{\partial \theta_j} &= \sum_{i=1}^m \frac{w_j}{p(x_i)} \frac{\partial}{\partial \theta_j} \varphi(x_i, \theta_j) = \\ &= \sum_{i=1}^m \frac{w_j \varphi(x_i, \theta_j)}{p(x_i)} \frac{\partial}{\partial \theta_j} \ln \varphi(x_i, \theta_j) = \\ &= \sum_{i=1}^m g_{ij} \frac{\partial}{\partial \theta_j} \ln \varphi(x_i, \theta_j) = \\ &= \frac{\partial}{\partial \theta_j} \sum_{i=1}^m g_{ij} \ln \varphi(x_i, \theta_j) = 0 \quad (\forall j = 1, \dots, k). \end{aligned} \quad (2.125)$$

Последнее равенство в (2.125) эквивалентно соотношению

$$\theta_j = \arg \max_{\theta} \sum_{i=1}^m g_{ij} \ln \varphi(x_i, \theta) \quad (\forall j = 1, \dots, k). \quad (2.126)$$

Соотношения (2.124) и (2.126) являются теми правилами, в соответствии с которыми вычисляются соответствующие компоненты вектора $\hat{\Theta}^{(s+1)}$.

Если $\varphi(x, \theta_j)$ имеет гауссов вид, т.е. $\varphi(x, \theta_j) = \mathcal{N}(x, \mu_j, \Sigma_j)$, то значение параметра $\theta_j = (\mu_j, \Sigma_j)$, удовлетворяющее условию (2.126), можно выразить в явном виде:

$$\mu_j = \frac{1}{mw_j} \sum_{i=1}^m g_{ij} x_i, \quad \Sigma_j = \frac{1}{mw_j} \sum_{i=1}^m g_{ij} (x_i - \mu_j)(x_i - \mu_j)^\top.$$

Число компонентов смеси неизвестно

В данном случае компоненты смеси строятся в процессе обучения.

Параметры алгоритма:

- R – максимально допустимый разброс значений $p(x_i)$,

- m_0 – минимальная длина выборки, по которой можно восстановить плотность,
- δ – параметр критерия останковки.

Сначала строится первая компонента: $k := 1$, $w_1 := 1$, и

$$\theta_1 := \arg \max_{\theta} \sum_{i=1}^m \ln \varphi(x_i, \theta).$$

Затем выполняется цикл по k (начиная с $k = 2$), тело этого цикла состоит из следующих действий:

1. Строится множество $U := \{x_i \in S_y : p(x_i) < \frac{1}{R} \max_j p(x_j)\}$,
(U – объекты, не подходящие ни к одной из компонентов).
2. **Если** $|U| < m_0$, **то** работа алгоритма завершается.
(В данном случае объекты из U рассматриваются как выбросы.)
3. **Иначе** добавляется новая (k -я) компонента с параметрами

$$w_k := \frac{1}{m}|U|, \quad \theta_k := \arg \max_{\theta} \sum_{x_i \in U} \ln \varphi(x_i, \theta),$$

$$\forall j = 1, \dots, k-1 \quad w_j := w_j(1 - w_k).$$

4. Для уточнения построенного в предыдущем действии вектора параметров Θ запускается EM-алгоритм на S_y с k компонентами смеси и параметром останковки δ .

Литература

- [1] Вьюгин В.В. Математические основы машинного обучения и прогнозирования. Москва, издательство МЦНМО, 2018. 384 с.
- [2] Ветров Д.П., Кропотов Д.А. Алгоритмы выбора моделей и построения коллективных решений в задачах классификации, основанные на принципе устойчивости. Москва, URSS, 2006. 112 с.
- [3] Информационно-аналитический ресурс по машинному обучению
<http://www.machinelearning.ru/>
- [4] Воронцов К. В., Математические методы обучения по прецедентам (теория обучения машин).
<http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf>
- [5] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386. (1958)
- [6] A. Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, volume 12, pages 615–622. (1962)
- [7] M. Mohri and A. Rostamizadeh. Perceptron Mistake Bounds. (2013)
<https://arxiv.org/pdf/1305.0208.pdf>
- [8] В. Н. Вапник, А. Я. Червоненкис. Теория распознавания образов. Статистические проблемы обучения. М., Наука. (1974)
- [9] Айзерман М. А., Браверман Э. М., Розоноэр Л. И. Метод потенциальных функций в теории обучения машин. - М.: Наука, 1970. - 320 с.
- [10] Головкин В. А. Нейронные сети: обучение, организация и применение. - М.: ИПР- ЖР, 2001.

- [11] Загоруйко Н. Г. Прикладные методы анализа данных и знаний. - Новосибирск: ИМ СО РАН, 1999.
- [12] Лапко А. В., Ченцов С. В., Крохов С. И., Фельдман Л. А. Обучающиеся системы обработки информации и принятия решений. Непараметрический подход. - Новосибирск: Наука, 1996.
- [13] Нейроинформатика / А. Н. Горбань, В. Л. Дунин-Барковский, А. Н. Кирдин, Е. М. Миркес, А. Ю. Новоходько, Д. А. Россиев, С. А. Терехов и др. - Новосибирск: Наука, 1998. - 296 с.
- [14] Хардле В. Прикладная непараметрическая регрессия. - М.: Мир, 1993.
- [15] Bishop C. M. Pattern Recognition and Machine Learning. - Springer, Series: Information Science and Statistics, 2006. - 740 pp.
- [16] Burges C. J. C. A tutorial on support vector machines for pattern recognition // Data Mining and Knowledge Discovery. - 1998. - Vol. 2, no. 2. - Pp. 121–167.
<http://citeseer.ist.psu.edu/burges98tutorial.html>
- [17] Murphy Kevin P. Machine Learning: A Probabilistic Perspective. The MIT Press, 2012, 1104 с.
- [18] Хайкин С. Нейронные сети. Полный курс. Вильямс, 2018, 1104 с.
- [19] Николенко С., Кадурын А., Архангельская Е. Глубокое обучение. Погружение в мир нейронных сетей, Питер, 2018. 480 с.
- [20] Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение. ДМК Пресс, 2017, 652 с.
- [21] Лесковец Ю., Раджараман А., Ульман Д. Анализ больших наборов данных. ДМК Пресс, 2016, 498 с.
- [22] Kung S.Y. Kernel Methods and Machine Learning, Cambridge University Press, 2014, 578 с.
- [23] Skansi S. Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence. Springer, 2018. 191 с.
- [24] Smith J. Machine Learning Systems. Manning Publications Co., 2018. 253 с.

- [25] Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python, Москва, 2017. 393 с.
- [26] Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. ДМК Пресс, 2015. 402 с.
- [27] Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Springer, 2014. 739 p.
- [28] Мерков А. Б. Распознавание образов. Введение в методы статистического обучения. 2011. 256 с.
- [29] Мерков А. Б. Распознавание образов. Построение и обучение вероятностных моделей. 2014. 238 с.
- [30] Коэлю Л.П., Ричарт В. Построение систем машинного обучения на языке Python. 2016. 302 с.
- [31] Barber D. Bayesian Reasoning and Machine Learning. Cambridge University Press, 2012.
- [32] Mackay D.J.C. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, 2003.
- [33] Wainwright M.J., Jordan M.I. Graphical Models, Exponential Families, and Variational Inference. Foundations and Trends in Machine Learning, NOWPress, 2008.
- [34] Koller D., Friedman N. Probabilistic Graphical Models: Principles and Techniques. The MIT Press, 2009.