В. В. Вьюгин

Математические основы машинного обучения и прогнозирования

УДК 519.2+519.83+004.8
ББК 22.17+22.18+32.813
В96

Вьюгин В. В.
Математические основы машинного обучения и прогнозирования
Электронное издание
М.: МЦНМО, 2014
304 с.
ISBN 978-5-4439-2014-6

Книга предназначена для первоначального знакомства с математическими основами современной теории машинного обучения (Machine Learning) и теории игр с предсказаниями. В первой части излагаются основы статистической теории машинного обучения, рассматриваются задачи классификации и регрессии с опорными векторами, теория обобщения и алгоритмы построения разделяющих гиперплоскостей. Во второй и третьей частях рассматриваются задачи адаптивного прогнозирования в нестохастических теоретико-игровой и сравнительной постановках: предсказания с использованием экспертных стратегий и игры с предсказаниями.

Для студентов и аспирантов, специализирующихся в области машинного обучения и искусственного интеллекта.

Подготовлено на основе книги: *В. В. Вьюгин*. Математические основы машинного обучения и прогнозирования. — М.: МЦНМО, 2013.

Издательство Московского центра
непрерывного математического образования
119002, Москва, Большой Власьевский пер., 11. Тел. (499) 241-74-83
http://www.mccme.ru

ISBN 978-5-4439-2014-6

ОГЛАВЛЕНИЕ

Часть I

Статистическая теория машинного обучения

ГЛАВА 1

ЭЛЕМЕНТЫ ТЕОРИИ КЛАССИФИКАЦИИ

Как было замечено во введении, теория машинного обучения решает задачи предсказания будущего поведения сложных систем в том случае, когда отсутствуют точные гипотезы о механизмах, управляющих поведением таких систем.

Имеется ряд категорий машинного обучения: контролируемое обучение, или «обучение с учителем» (supervised learning); неконтролируемое обучение (unsupervised learning; в частности, кластеризация); обучение с подкреплением (reinforcement learning). В этой и следующей главах нас будет интересовать первый тип машинного обучения — контролируемое обучение. Мы начинаем с обучающей выборки, представляющей собой примеры — пары вида «вход — выход». Целью обучения является восстановление зависимости между элементами этих пар для предсказания будущего выхода по заданному входу.

В основе статистической теории машинного обучения лежит гипотеза о существовании стохастического механизма, генерирующего такие пары. В этом случае мы можем оценивать вероятность ошибки классификации будущих примеров. При этом делаются минимальные предположения о виде вероятностного источника, генерирующего данные. Теория обобщения предоставляет оценки таких ошибок, равномерные относительно максимально широких классов вероятностных распределений, генерирующих данные.

Мы рассмотрим два основных класса задач: *задачи классификации* и *задачи регрессии*. В данной главе рассматривается задача классификации, в которой выход — это метка класса, к которому принадлежит вход.

1.1. Задача классификации

1.1.1. Байесовский классификатор

Предварительно рассмотрим один простейший метод классификации. Рассмотрим пару случайных переменных (X, Y), принимающую значения в множестве $\mathscr{X} \times \{-1, 1\}$. Предполагаем, что этой паре соответствует распределение вероятностей P и определены апостериорные вероятности принадлежности объекта $x \in \mathscr{X}$ к первому и второму классам: $P\{Y = 1 \mid X = x\}$ и $P\{Y = -1 \mid X = x\}$.

Легко построить оптимальный классификатор, если распределение вероятностей P, генерирующее пары «вход-выход», известно.

Обозначим условную вероятность того, что объект x принадлежит первому классу

$$\eta(x) = P\{Y = 1 \mid X = x\}.$$

Для произвольного классификатора $g : \mathscr{X} \to \{-1, 1\}$ вероятность ошибки классификации равна

$$\mathrm{err}_P(g) = P\{g(X) \neq Y\}.$$

Байесовский классификатор определяется как

$$h(x) = \begin{cases} 1, & \text{если} \quad \eta(x) > \dfrac{1}{2}, \\ -1 & \text{в противном случае.} \end{cases}$$

Следующая лемма показывает, что байесовский классификатор минимизирует вероятность ошибки $\mathrm{err}_P(h)$, которая в данном случае называется *байесовской ошибкой*.

Лемма 1.1. *Для любого классификатора $g : \mathscr{X} \to \{-1, 1\}$*

$$P\{h(X) \neq Y\} \leqslant P\{g(X) \neq Y\}. \tag{1.1}$$

Доказательство. Для произвольного классификатора g условная вероятность ошибки классификации при $X = x$ выражается в виде

$$\begin{gathered} P\{g(X) \neq Y \mid X = x\} = 1 - P\{g(X) = Y \mid X = x\} = \\ = 1 - \big(P\{Y = 1, g(X) = 1 \mid X = x\} + P\{Y = -1, g(X) = -1 \mid X = x\}\big) = \\ = 1 - \big(1_{g(x)=1} P\{Y = 1 \mid X = x\} + 1_{g(x)=-1} P\{Y = -1 \mid X = x\}\big) = \\ = 1 - \big(1_{g(x)=1}\eta(x) + 1_{g(x)=-1}(1 - \eta(x))\big), \end{gathered}$$

где для любого условия $R(x)$ будет $1_{R(x)} = 1$, если $R(x)$ выполнено, и $1_{R(x)} = 0$, в противном случае.

Аналогичное равенство выполнено для классификатора $h(x)$.

Заметим, что $1_{g(x)=-1} = 1 - 1_{g(x)=1}$ для любой функции классификации g. Таким образом, для каждого $x \in \mathscr{X}$

$$P\{g(X) \neq Y \mid X = x\} - P\{h(X) \neq Y \mid X = x\} =$$
$$= \eta(x)\big(1_{h(x)=1} - 1_{g(x)=1}\big) + (1 - \eta(x))\big(1_{h(x)=-1} - 1_{g(x)=-1}\big) =$$
$$= (2\eta(x) - 1)\big(1_{h(x)=1} - 1_{g(x)=1}\big) \geqslant 0$$

по определению байесовского классификатора h.

Интегрируем обе части этого неравенства по x. Получим неравенство леммы. □

Байесовский классификатор служит эталоном для оценки качества алгоритмов классификации.

Обозначим посредством $\mathscr{D}$ множество всех измеримых функций классификаторов типа $g\colon \mathscr{X} \to \{-1, 1\}$. Условие (1.1) можно записать в виде

$$\operatorname{err}_P(h) = P\{h(X) \neq Y\} = \inf_{g \in \mathscr{D}} P\{g(X) \neq Y\}.$$

Пусть некоторый классификатор $g_l \in \mathscr{D}$ построен некоторым алгоритмом $\mathscr{A}$ по случайной выборке $S = ((x_1, y_1), \ldots, (x_l, y_l))$, сгенерированной распределением вероятностей P. Алгоритм классификации $\mathscr{A}$ называется *состоятельным* для распределения P, если случайная величина $\operatorname{err}_P(g_l)$ сходится к $\operatorname{err}_P(h)$ по вероятности P, т. е. для любого $\varepsilon > 0$

$$P\big\{|\operatorname{err}_P(g_l) - \operatorname{err}_P(h)| > \varepsilon\big\} \to 0 \tag{1.2}$$

при $l \to \infty$.

Алгоритм классификации $\mathscr{A}$ называется *универсально состоятельным*, если условие (1.2) имеет место для любого распределения P.

Недостатком байесовского классификатора является то, что он использует для вычисления значений функции $h(x)$ вероятностное распределение P, генерирующее пары (x, y). Прежде чем использовать байесовский классификатор, надо решить задачу восстановления вероятностного распределения P по его реализациям. На практике такое вероятностное распределение часто неизвестно и его трудно восстановить. Обычно для получения достоверного результата требуется довольно много реализаций случайной величины (X, Y).

Основные проблемы статистической теории классификации связаны с тем, что при построении классификаторов $h(x)$ мы не можем использовать распределения вероятностей, генерирующие пары (x, y).

Таким образом, в дальнейшем будут рассматриваться классификаторы, не зависящие от вероятностного распределения, генерирующего данные.

Байесовский классификатор служит для сравнения предсказательной способности других алгоритмов классификации.

1.1.2. Постановка задачи классификации

Важная проблема, возникающая в статистической теории машинного обучения, заключается в том, чтобы понять, как много случайных примеров необходимо использовать при обучении для того, чтобы гарантировать достаточно малую ошибку классификации с заданной степенью достоверности.

Сначала напомним основные идеи PAC-теории машинного обучения (Probably Approximately Correct Learning), предложенной Валиантом [36].

В данном случае формальная постановка задачи основана на вероятностных предположениях. Мы предполагаем, что все примеры, представленные для обучения или проверки, независимо и одинаково распределены согласно некоторому фиксированному неизвестному распределению вероятностей P на множестве $\mathscr{X}$ этих примеров, имеющем структуру вероятностного пространства.

Предполагаем, что каждый пример x имеет метку — признак принадлежности к некоторому классу. Метки классов образуют множество D и задаются с помощью неизвестной нам функции $c \in C$ типа $c\colon \mathscr{X} \to D$, которая называется *концептом*: $c(x)$ — метка x.

Допустим, что по некоторой случайной выборке

$$S = \big((x_1, c(x_1)), \ldots, (x_l, c(x_l))\big),$$

порожденной распределением P, мы построили гипотезу $h = h_S$, которая выражает принадлежность объектов x к классам (подмножествам $\mathscr{X}$), порожденным неизвестным нам концептом c. Ошибка гипотезы h определяется как

$$\mathrm{err}_P(h) = P\{h(x) \neq c(x)\}.$$

Ошибка $\mathrm{err}_P(h)$ является случайной величиной, так как функция $h = h_S$ зависит от S.

Рассмотрим задачу: найти такую гипотезу h, для которой вероятность события, заключающегося в том, что ошибка $\mathrm{err}_P(h)$ велика, является малой. Другими словами, мы хотели бы утверждать, что гипотеза h *вероятно приблизительно верна* (probably approximately correct).

Степень «приблизительности» количественно будет выражаться с помощью параметра ε: мы будем требовать выполнения неравенства $\mathrm{err}_P(h) \leqslant \varepsilon$.

Степень «вероятности» будет измеряться с помощью параметра уровня доверия δ. Мы хотим получить хорошую аппроксимацию концепта $c \in C$ с высокой вероятностью. В частности, мы требуем, чтобы неравенство $\mathrm{err}_P(h) \leqslant \varepsilon$ выполнялось бы с вероятностью не меньшей чем $1 - \delta$.

Все эти соображения приводят к следующей точной формулировке PAC-теории машинного обучения.

Алгоритм A восстанавливает класс концептов C с помощью класса гипотез H, если для любого концепта $c \in C$, для любого распределения вероятностей P на примерах x, а также для любых $\varepsilon \in (0, 1/2)$ и $\delta \in \in (0, 1/2)$ выполнено следующее:

- алгоритм A получает на вход обучающую выборку, состоящую из случайных пар $(x, c(x))$, независимо и одинаково распределенных согласно P, число которых полиномиально зависит от $1/\varepsilon$ и $1/\delta$;
- алгоритм A выдает в качестве результата функцию h, для которой $\mathrm{err}_P(h) \leqslant \varepsilon$ с вероятностью не менее $1 - \delta$.

В этом случае говорим, что класс концептов C является PAC-изучаемым (Probably Approximately Correct Learnable).

В данной работе будет рассматриваться постановка задачи, несколько отличная от классической постановки задачи PAC-теории машинного обучения.

Мы не используем понятие концепта, вместо этого мы просто предполагаем, что пары (x, y) объектов x и их меток y одинаково и независимо распределены согласно некоторому неизвестному вероятностному распределению P на множестве $\mathscr{X} \times D$. Подобная постановка принята в современной *статистической теории машинного обучения*. В остальном все идеи PAC-теории сохраняются.

Мы предполагаем, что выборка $S = ((x_1, y_1), \ldots, (x_l, y_l))$ генерируется (порождается) некоторым источником. Основное предположение об источнике, порождающем выборку S, заключается в том, что на парах (x, y), т. е. на пространстве $\mathscr{X} \times D$, задано распределение вероятностей P, а пары (x_i, y_i), образующие выборку S, одинаково и независимо распределены.

Соответственно на множестве $(\mathscr{X} \times D)^l$ задано распределение вероятностей $P^l = P \times P \times \ldots \times P$.

Правило, или *функция* (*гипотеза*), *классификации* — это функция типа $h\colon \mathscr{X} \to D$, которая разбивает элементы $x_i \in \mathscr{X}$ на несколько классов. Мы будем также называть функцию h классификатором, или решающим правилом.

В дальнейшем у нас всегда будет рассматриваться случай бинарной классификации $D = \{-1, 1\}$, а функция $h\colon \mathscr{X} \to D$ будет называться индикаторной. В этом случае вся выборка S разбивается на две подвыборки: $S^+ = ((x_i, y_i)\colon y_i = 1)$ — положительные примеры (или первый класс) и $S^- = ((x_i, y_i)\colon y_i = -1)$ — отрицательные примеры (или второй класс).

В некоторых случаях индикаторная функция классификации h задается с помощью некоторой вещественной функции f и числа $r \in \mathbb{R}$:

$$h(x) = \begin{cases} 1, & \text{если} \quad f(x) \geqslant r, \\ -1 & \text{в противном случае.} \end{cases}$$

Предсказательная способность произвольной функции классификации h будет оцениваться по *ошибке классификации*, которая определяется как вероятность неправильной классификации

$$\mathrm{err}_P(h) = P\{h(x) \neq y\} = P\{(x, y)\colon h(x) \neq y\}.$$

Функция $\mathrm{err}_P(h)$ также называется *риск-функционалом*.

Основная цель при решении задачи классификации — для заданного класса функций классификации H построить оптимальный классификатор, т. е. такую функцию классификации $h \in H$, при которой ошибка классификации $\mathrm{err}_P(h)$ является наименьшей в классе H.

В этой главе в основном будет рассматриваться задача классификации n-мерных векторов — элементов множества $\mathbb{R}^n$, где $\mathbb{R}$ — множество всех действительных чисел. Далее D — множество классов этих векторов, которое является конечным множеством с небольшим числом элементов. Размерность n евклидова пространства $\mathbb{R}^n$ обычно велика по сравнению с числом классов.

Далее элементы $\mathbb{R}^n$ будем обозначать подчеркнутыми сверху буквами: $\overline{x}, \overline{y}, \ldots \in \mathbb{R}^n$; в координатах — $\overline{x} = (x_1, \ldots, x_n)$. Будут рассматриваться операции сложения векторов

$$\overline{x} + \overline{y} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \cdots \\ x_n + y_n \end{pmatrix},$$

умножения на вещественное число

$$\alpha\overline{x} = \begin{pmatrix} \alpha x_1 \\ \alpha x_2 \\ \cdots \\ \alpha x_n \end{pmatrix},$$

где $\overline{x} = (x_1, \ldots, x_n)'$ и $\overline{y} = (y_1, \ldots, y_n)'$. С помощью штриха мы уточняем форму представления вектора в виде матрицы — простую или транспонированную, но только в тех случаях, когда это имеет существенное значение.

На векторах из $\mathbb{R}^n$ также определено их скалярное произведение $(\overline{x} \cdot \overline{y}) = x_1 y_1 + \ldots + x_n y_n$. Норма (длина) вектора $\overline{x}$ определяется как $\|\overline{x}\| = \sqrt{(\overline{x} \cdot \overline{x})} = \sqrt{\sum_{i=1}^{n} x_i^2}$. При решении задачи классификации мы исходим из *обучающей выборки* $S = ((\overline{x}_1, y_1), \ldots, (\overline{x}_l, y_l))$, где $\overline{x}_i \in \mathscr{X}$ — вектор евклидова пространства $\mathbb{R}^n$ размерности n (например, это может быть цифровой образ какого-либо изображения), y_i — это элемент конечного множества D с небольшим числом элементов (метка класса), например, $y_i \in \{-1, 1\}$. Элементы $y_i \in D$ определяют классы объектов $\overline{x}_i$.

При решении задачи многомерной регрессии также рассматривается обучающая выборка $S = ((\overline{x}_1, y_1), \ldots, (\overline{x}_l, y_l))$, при этом элементы y_i обычно являются вещественными числами, т. е. $D = \mathbb{R}$. Задача регрессии будет рассмотрена в разделах 2.8, 2.9, а также в разделе 5.9.

1.1.3. Линейные классификаторы: персептрон

Рассмотрим один из наиболее старых алгоритмов классификации — персептрон.

Персептрон представляет собой некоторую техническую модель восприятия. Модель имеет два слоя. Первый, рецепторный слой подает сигнал на входы пороговых элементов — нейронов преобразующего слоя.

Математическая модель *персептрона* будет задаваться следующим образом. Задано пространство $\mathscr{X}$ исходных описаний объекта. Преобразование $\overline{y} = \overline{\varphi}(\overline{x})$, которое в координатном виде записывается как $y_i = \varphi_i(\overline{x})$, $i = 1, \ldots, n$, ставит исходному описанию $\overline{x} = (x_1, \ldots, x_m) \in \mathscr{X}$ объекта преобразованное описание объекта $\overline{y} = (y_1, \ldots, y_n) \in \mathscr{Y}$. Предполагаем, что $\mathscr{X} \subseteq \mathbb{R}^m$ и $\mathscr{Y} \subseteq \mathbb{R}^n$ для некоторых m, n.

Персептрон задается однородной линейной функцией

$$L(\overline{x}) = (\Lambda \cdot \overline{\varphi}(\overline{x})) = \sum_{i=1}^{n} \lambda_i \varphi_i(\overline{x}) = \sum_{i=1}^{n} \lambda_i y_i,$$

где действительные числа λ_i интерпретируются как веса, приписываемые преобразованным признакам y_i. Здесь $(\Lambda \cdot \overline{\varphi}(\overline{x}))$ обозначает скалярное произведение двух векторов $\Lambda = (\lambda_1, \ldots, \lambda_n)$ и $\overline{\varphi}(\overline{x}) = (\varphi_1(\overline{x}), \ldots, \varphi_n(\overline{x}))$ в евклидовом пространстве $\mathbb{R}^n$.

Будем связывать с персептроном *функцию активации*:

$$f(\overline{x}) = \sigma\left(\sum_{i=1}^{n} \lambda_i \varphi_i(\overline{x})\right).$$

Примеры функций активации:

$$\sigma(t) = \operatorname{sign}(t),$$
$$\sigma(t) = \frac{1}{1+e^{-t}},$$
$$\sigma(t) = \frac{e^t - e^{-t}}{e^t + e^{-t}},$$

где

$$\operatorname{sign}(t) = \begin{cases} 1, & \text{если} \quad t \geqslant 0, \\ -1, & \text{если} \quad t < 0. \end{cases}$$

В дальнейшем для простоты будем использовать бинарную функцию активации $\sigma(t) = \operatorname{sign}(t)$, которая определяет следующий классификатор. Считаем, что вектор $\overline{x}$ принадлежит первому классу, если

$$\sum_{i=1}^{n} \lambda_i \varphi_i(\overline{x}) \geqslant 0.$$

В противном случае вектор $\overline{x}$ принадлежит второму классу.

Геометрически это означает, что в пространстве признаков $\mathscr{X}$ задана гиперповерхность

$$\sum_{i=1}^{n} \lambda_i \varphi_i(\overline{x}) = 0, \tag{1.3}$$

которая делит пространство $\mathscr{X}$ на два полупространства. Объекты первого класса относятся к одному полупространству, объекты второго класса относятся ко второму полупространству. Подобная гиперповерхность называется *разделяющей*.

Каждой разделяющей гиперповерхности (1.3) соответствует разделяющая гиперплоскость

$$\sum_{i=1}^{n} \lambda_i y_i = 0$$

в пространстве преобразованных признаков $\mathscr{Y}$. Пространство $\mathscr{Y}$ также называется *спрямляющим*.

Пусть задана бесконечная обучающая выборка в спрямляющем пространстве

$$S = \big((\overline{y}_1, \varepsilon_1), (\overline{y}_2, \varepsilon_2), \ldots\big),$$

где ε_i обозначает принадлежность объекта $\overline{y}_i = \overline{\varphi}(\overline{x}_i)$ классу $\varepsilon_i \in \{-1, 1\}$, $i = 1, 2, \ldots$

Допустим, что существует гиперплоскость, строго разделяющая выборку S. Пусть $\Lambda = (\lambda_1, \ldots, \lambda_n)$ — вектор коэффициентов этой разделяющей гиперплоскости. По определению гиперплоскость строго разделяет выборку, если выполнено неравенство

$$\inf_i \varepsilon_i (\Lambda \cdot \overline{y}_i) > 0. \tag{1.4}$$

Для удобства преобразуем обучающую выборку следующим образом. Рассмотрим последовательность векторов $\tilde{y}_1, \tilde{y}_2, \ldots$, где

$$\tilde{y}_i = \begin{cases} \overline{y}_i, & \text{если} \quad \varepsilon_i = 1, \\ -\overline{y}_i, & \text{если} \quad \varepsilon_i = -1, \end{cases}$$

для всех i. Тогда условие строгого разделения (1.4) запишется в виде

$$\inf_i (\Lambda \cdot \tilde{y}_i) > 0.$$

Обозначим

$$\rho(\Lambda) = \inf_i \frac{(\Lambda \cdot \tilde{y}_i)}{\|\Lambda\|}, \quad \rho_0 = \sup_{\Lambda \neq \overline{0}} \rho(\Lambda), \tag{1.5}$$

где $\|\Lambda\| = \sqrt{\sum_{i=1}^{n} \lambda_i^2}$ — длина вектора Λ в пространстве $\mathbb{R}^n$.

Условие строгой разделимости выборки S может быть записано в виде $\rho_0 > 0$.

Алгоритм Розенблатта построения разделяющей гиперплоскости

Пусть задана произвольная бесконечная обучающая выборка

$$(\overline{y}_1, \varepsilon_1), \ (\overline{y}_2, \varepsilon_2), \ \ldots$$

и пусть существует гиперплоскость, проходящая через начало координат $(\Lambda^* \cdot \overline{y}) = 0$, строго разделяющая эту выборку, т. е. такая, что

$$\inf_i (\Lambda^* \cdot \tilde{y}_i) > 0.$$

Считаем, что $\|\Lambda^*\| = 1$.

Пусть задан порог разделения — число $\rho_0 > 0$ такое, что

$$(\Lambda^* \cdot \tilde{y}_i) > \rho_0 \tag{1.6}$$

для всех i. Также предполагаем, что векторы $\overline{y}_i$ равномерно ограничены по модулю

$$\sup_i |\overline{y}_i| = D < \infty.$$

Обучение персептрона заключается в изменении координат вектора весов Λ на каждом шаге алгоритма 1.1.

Пусть $\Lambda_t = (\lambda_{1,t}, \ldots, \lambda_{n,t})$ — текущий вектор коэффициентов гиперплоскости, вычисленный на шаге t алгоритма, $t = 1, 2, \ldots$ Алгоритм использует преобразованную последовательность векторов $\tilde{y}_1, \tilde{y}_2, \ldots$

1.1. Алгоритм Розенблатта построения разделяющей гиперплоскости

Полагаем $\Lambda_0 = (0, \ldots, 0)$.

FOR $t = 1, 2, \ldots$

Если $(\Lambda_{t-1} \cdot \tilde{y}_t) \geqslant 0$, то полагаем $\Lambda_t = \Lambda_{t-1}$ (т. е. если очередной вектор классифицируется правильно, то текущая гиперплоскость не изменяется).

Если $(\Lambda_{t-1} \cdot \tilde{y}_t) < 0$ (очередной вектор классифицируется неправильно), то производим корректировку направляющего вектора гиперплоскости $\Lambda_t = \Lambda_{t-1} + \tilde{y}_t$, назовем эту операцию также *исправлением ошибки*.

ENDFOR

Следующая теорема, принадлежащая А. А. Новикову, утверждает, что в том случае, когда существует гиперплоскость разделяющая выборку с положительным порогом, алгоритм Розенблатта после многократного предъявления обучающей последовательности, составленной из элементов выборки, построит за конечное число шагов гиперплоскость, строго разделяющую всю выборку.

Теорема 1.1. *Если существует гиперплоскость, разделяющая бесконечную выборку*

$$(\overline{y}_1, \varepsilon_1),\ (\overline{y}_2, \varepsilon_2),\ \ldots$$

с положительным порогом, то в алгоритме Розенблатта исправление ошибки происходит не более чем $\left\lfloor \dfrac{D^2}{\rho_0^2} \right\rfloor$ *раз*[1)]. *Это значит, что неравенство* $\Lambda_t \neq \Lambda_{t-1}$ *выполнено для не более чем* $\left\lfloor \dfrac{D^2}{\rho_0^2} \right\rfloor$ *различных* t.

После этого разделяющая гиперплоскость стабилизируется и будет безошибочно делить всю бесконечную оставшуюся часть последовательности.

Доказательство. Если на шаге t происходит изменение вектора Λ_t, то

$$\|\Lambda_t\|^2 = \|\Lambda_{t-1}\|^2 + 2(\Lambda_{t-1} \cdot \tilde{y}_t) + \|\tilde{y}_t\|^2.$$

[1)] Здесь и далее для любого вещественного числа r через $\lfloor r \rfloor$ обозначается максимальное целое число не больше r, а через $\lceil r \rceil$ — минимальное целое число не меньше r.

Так как $(\Lambda_{t-1} \cdot \tilde{y}_t) < 0$ (классификация t-го вектора неправильная) и $\|\tilde{y}_t\| \leqslant D$, получаем

$$\|\Lambda_t\|^2 \leqslant \|\Lambda_{t-1}\|^2 + D^2.$$

Если до шага T включительно произошло k таких исправлений, то получаем

$$\|\Lambda_t\|^2 \leqslant kD^2. \tag{1.7}$$

По условию разделимости (1.6) существует единичный вектор Λ^* такой, что

$$(\Lambda^* \cdot \tilde{y}_i) \geqslant \rho_0$$

для всех i.

Оценим величину $(\Lambda_t \cdot \Lambda^*)$. По определению $(\Lambda_0 \cdot \Lambda^*) = 0$. Если на шаге t алгоритм производит исправление, то

$$(\Lambda_t \cdot \Lambda^*) = (\Lambda_{t-1} \cdot \Lambda^*) + (\Lambda^* \cdot \tilde{y}_t) \geqslant (\Lambda_{t-1} \cdot \Lambda^*) + \rho_0.$$

Если на шаге t исправления не происходит, то

$$(\Lambda_t \cdot \Lambda^*) = (\Lambda_{t-1} \cdot \Lambda^*).$$

Таким образом, если к шагу t алгоритм произвел k исправлений, то

$$(\Lambda_t \cdot \Lambda^*) \geqslant k\rho_0.$$

По неравенству Коши–Буняковского

$$(\Lambda_t \cdot \Lambda^*) \leqslant \|\Lambda_t\| \cdot \|\Lambda^*\| = \|\Lambda_t\|.$$

Поэтому имеет место неравенство

$$\|\Lambda_t\| \geqslant k\rho_0. \tag{1.8}$$

Объединяем неравенства (1.7) и (1.8), получаем

$$k \leqslant \frac{D^2}{\rho_0^2}.$$

Таким образом, число исправлений не превосходит

$$k \leqslant \left\lfloor \frac{D^2}{\rho_0^2} \right\rfloor.$$

Теорема доказана. □

По теореме 1.1, какова бы ни была бесконечная разделимая с положительным порогом выборка, алгоритм Розенблатта, сделав конечное число исправлений, не превосходящее $\left\lfloor \frac{D^2}{\rho_0^2} \right\rfloor$, найдет какую-либо гиперплоскость, строго разделяющую всю выборку.

В некоторых случаях в персептроне рассматривается бинарное спрямляющее пространство, т. е. $\mathscr{Y} = \{-1, 1\}^n$.

В этом случае ясно, что $D^2 \leqslant n$. Тогда оценка теоремы 1.1 имеет вид

$$k \leqslant \left\lfloor \frac{n}{\rho_0^2} \right\rfloor,$$

т. е. число коррекций алгоритма обучения персептрона растет линейно с размерностью пространства.

В этом разделе была рассмотрена двухуровневая модель персептрона. На первом уровне определяется отображение $\overline{y} = \overline{\phi}(\overline{x})$ исходного пространства описаний объектов $\mathscr{X}$ в спрямляющее пространство $\mathscr{Y}$. На втором уровне реализуется алгоритм обучения — построение разделяющей гиперплоскости в пространстве $\mathscr{Y}$ на основе обучающей последовательности. Основное требование к отображению $\overline{\phi}$ диктует вторая часть модели, а именно, множества векторов — образов $\overline{y}$, принадлежащих к различным классам, должны быть разделимы гиперплоскостью.

Многослойная нейронная сеть. Персептроны можно комбинировать в виде *многослойных нейронных* сетей. В каждой вершине v такой сети располагается некоторая функция

$$f^v(\overline{x}) = \sigma\left((\overline{w}^v \cdot \overline{x}) + b^v\right),$$

σ — функция активации; на место аргумента в ней подставлено значение персептрона.

Рассмотрим *сеть* вершин, состоящую из l слоев. Заданы натуральные числа $n_1, \ldots, n_l$ — размеры слоев (число вершин в слое), причем самый верхний слой состоит из одной вершины: $n_l = 1$.

С каждой j-й вершиной i-го слоя сети ассоциируется функция

$$f_{i,j}(\overline{x}) = \sigma\left((\overline{w}^{i,j} \cdot \overline{x}) + b^{i,j}\right),$$

где $\overline{w}^{i,j}, \overline{x} \in \mathbb{R}^{n_{i-1}}$ и $b^{i,j} \in \mathbb{R}$, $n_0 > 0$.

Нейронная сеть может быть представлена в виде набора векторнозначных функций

$$f_i \colon \mathbb{R}^{n_{i-1}} \to \mathbb{R}^{n_i},$$

$i = 1, \ldots, l$, где $f_i = (f_{i,1}, \ldots, f_{i,n_{i-1}})$.

Выход нейронной сети задается одномерной функцией — композицией

$$f_l \circ f_{l-1} \circ \ldots \circ f_2 \circ f_1.$$

Векторы $\overline{w}^{i,j}$ называются весами, которые приписаны вершинам (i, j) нейронной сети.

1.2. Теория обобщения

1.2.1. Верхние оценки вероятности ошибки классификации

В теории обобщения вычисляются вероятности ошибки классификации на *тестовой выборке*, после того как функция классификации определена по *обучающей выборке*, т. е. проведено обучение алгоритма классификации.

В этом разделе мы приведем основные положения статистической теории обобщения.

Статистическая теория машинного обучения использует гипотезу о том, что пары (x_i, y_i) генерируются некоторым *неизвестным* нам распределением вероятностей, при этом, как правило, рассматривается очень широкий класс таких распределений. Используется только предположение о том, что данные независимо и одинаково распределены.

В статистической теории машинного обучения исходят из *обучающей выборки*, по которой определяется функция классификации или регрессии, наилучшим образом описывающая эту выборку. Класс функций классификации может быть очень широк – от разделяющих гиперплоскостей в n-мерном пространстве до произвольных многообразий, которые отображаются с помощью ядерных методов в гиперплоскости, расположенные в пространствах размерности $m > n$. Никакие распределения вероятностей не используются в алгоритмах, вычисляющих значения функции классификации.

Функция классификации проверяется на *тестовой выборке*. Задача теории обобщения состоит в том, чтобы оценить вероятность ошибки классификации на произвольной тестовой выборке.

Теория обобщения Вапника–Червоненкиса позволяет вычислить вероятность ошибки классификации или регрессии (относительно распределения вероятностей, генерирующего данные, возможно, неизвестного нам) для согласованной по обучающей выборке функции классификации или регрессии на любых будущих данных. Такая вероятность зависит от размера обучающей выборки и размерности или емкости класса функций, описывающих данные.

Емкость класса функций не зависит от числа параметров этих функций или от аналитического способа их задания. Она зависит от геометрических свойств класса — максимального размера проекций функций этого класса на выборки заданной длины.

В этом разделе будут даны равномерные верхние оценки вероятности ошибки в зависимости от длины обучающей выборки и размерности класса функций классификации.

Критерий выбора функции классификации основан на минимизации верхней оценки вероятности ошибки обобщения.

Пусть $S = \big((x_1, y_1), \ldots, (x_l, y_l)\big)$ — обучающая выборка. Здесь $x_i \in \mathscr{X}$ и $y_i \in \{-1, 1\}$ при $1 \leqslant i \leqslant l$. Элементы $\mathscr{X}$ называются *объектами*, а элементы D называются *метками*. В приложениях обычно $\mathscr{X} \subseteq \mathbb{R}^n$, где $\mathbb{R}^n$ — n-мерное евклидово векторное пространство.

Предполагаем, что на множестве $\mathscr{X} \times D$ задана структура вероятностного пространства с распределением P. Посредством P^l обозначаем вероятностное распределение на выборках S длины l, которое есть произведение l экземпляров распределения P.

Пусть задано правило (или функция) $h\colon \mathscr{X} \to \{-1, 1\}$. Риск-функционал (или ошибка классификации) определяется как

$$\operatorname{err}_P(h) = P\big\{(x, y)\colon h(x) \neq y\big\}.$$

Эта величина равна вероятности неправильной классификации.

Гипотеза классификации h согласована с выборкой

$$S = \big((x_1, y_1), \ldots, (x_l, y_l)\big),$$

если $h(x_i) = y_i$ для всех $1 \leqslant i \leqslant l$. Обозначим через

$$\operatorname{err}_S(h) = \frac{1}{l}\Big|\big\{i\colon h(x_i) \neq y_i, 1 \leqslant i \leqslant l\big\}\Big|$$

относительное число ошибок классификации h на выборке S. Здесь $|A|$ — число элементов множества A. Тогда гипотеза классификации h согласована с выборкой S, если $\operatorname{err}_S(h) = 0$.

Для произвольной гипотезы классификации h и $\varepsilon > 0$ имеем

$$P^l\big\{S\colon \operatorname{err}_S(h) = 0 \;\&\; \operatorname{err}_P(h) > \varepsilon\big\} = \prod_{i=1}^{l} P\{h(x_i) = y_i\} =$$
$$= \prod_{i=1}^{l} \big(1 - P\{h(x_i) \neq y_i\}\big) = (1 - \operatorname{err}_P(h))^l \leqslant e^{-l\varepsilon}. \quad (1.9)$$

Здесь мы использовали независимость ошибок на элементах выборки.

Пусть H — некоторый класс гипотез классификации. Если класс H конечный, то из (1.9) получаем оценку

$$P^l\big\{S\colon (\exists h \in H)(\operatorname{err}_S(h) = 0 \;\&\; \operatorname{err}_P(h) > \varepsilon)\big\} \leqslant |H| e^{-l\varepsilon}. \quad (1.10)$$

Интерпретация (1.10) заключается в следующем.

Пусть задан критический уровень $\delta > 0$ принятия ошибочной гипотезы классификации $h \in H$, согласованный с обучающей выборкой S. Тогда по (1.10) мы можем утверждать, что с вероятностью не меньше

$1-\delta$ гипотеза классификации $h_S \in H$, построенная по случайной обучающей выборке S и согласованная с ней, будет иметь ошибку классификации [1)]

$$\operatorname{err}_P(h) \leqslant \varepsilon = \frac{1}{l} \ln \frac{|H|}{\delta}.$$

Другими словами, всякая гипотеза классификации h, имеющая ошибку $\operatorname{err}_P(h) > \varepsilon$, с вероятностью не меньше $1 - |H|e^{-l\varepsilon}$ не будет согласована со случайной выборкой длины l.

В случае бесконечного семейства функций H аналогичные оценки на ошибку классификации дает теория обобщения Вапника–Червоненкиса. Сложность класса H оценивается с помощью функции роста

$$B_H(l) = \max_{(x_1, x_2, \ldots, x_l)} \left| \{(h(x_1), h(x_2), \ldots, h(x_l)) \colon h \in H\} \right|.$$

Свойства этой функции будут изучаться далее.

Имеет место теорема — аналог соотношения (1.10) для бесконечного H.

Теорема 1.2. *При $l > 2/\varepsilon$ имеет место оценка*

$$P^l \{S \colon (\exists\, h \in H)(\operatorname{err}_S(h) = 0 \;\&\; \operatorname{err}_P(h) > \varepsilon)\} \leqslant 2B_H(2l)e^{-\varepsilon l/4}.$$

Доказательство теоремы. Пусть $1_{h(\overline{x}) \neq y}$ есть величина, равная 1, если $h(x) \neq y$, и равная 0 в противном случае. Тогда

$$E\, 1_{h(x) \neq y} = \operatorname{err}_P(h),$$

где E — математическое ожидание по мере P. По определению

$$\operatorname{err}_S(h) = \frac{1}{l} \sum_{i=1}^{l} 1_{h(x_i) \neq y_i}$$

— частота ошибок классификации на выборке S.

Утверждение теоремы будет следовать из следующих двух лемм.

Лемма 1.2. *Пусть задан класс H функций классификации. Рассматриваются две случайные выборки S, S' длины l. Тогда для любого $\varepsilon > 0$ при $l > 2/\varepsilon$ имеет место неравенство*

$$P^l \{S \colon (\exists\, h \in H)(\operatorname{err}_S(h) = 0 \;\&\; \operatorname{err}_P(h) > \varepsilon)\} \leqslant$$
$$\leqslant 2P^{2l} \left\{ SS' \colon (\exists h \in H) \left(\operatorname{err}_S(h) = 0 \;\&\; \operatorname{err}_{S'}(h) > \frac{1}{2}\varepsilon \right) \right\}, \quad (1.11)$$

где SS' — двойная выборка, составленная из элементов выборки S и следующих за ними элементов выборки S'.

[1)] В дальнейшем $\ln r$ обозначает натуральный логарифм положительного числа r, а $\log r$ будет обозначать логарифм r по основанию 2.

Доказательство. Легко видеть, что неравенство (1.11) эквивалентно неравенству

$$P^l\Big\{S\colon \sup_{h\colon \operatorname{err}_S(h)=0}\operatorname{err}_P(h)>\varepsilon\Big\}\leqslant 2P^{2l}\Big\{SS'\colon \sup_{h\colon \operatorname{err}_S(h)=0}\operatorname{err}_{S'}(h)>\frac{1}{2}\varepsilon\Big\}. \quad (1.12)$$

Докажем (1.12). Для каждой выборки S из множества левой части неравенства (1.12) обозначим посредством h_S какую-нибудь функцию из класса H, для которой выполняются равенство $\operatorname{err}_S(h_S)=0$ и неравенство $\operatorname{err}_P(h_S)>\varepsilon$. Это случайная величина, зависящая от выборки.

Имеет место следующее неравенство между случайными величинами [1)]

$$1_{\operatorname{err}_S(h_S)=0\ \&\ \operatorname{err}_P(h_S)>\varepsilon}1_{\operatorname{err}_P(h_S)-\operatorname{err}_{S'}(h_S)\leqslant\varepsilon/2}\leqslant 1_{\operatorname{err}_S(h_S)=0\ \&\ \operatorname{err}_{S'}(h_S)>\varepsilon/2}. \quad (1.13)$$

Возьмем математическое ожидание по второй выборке S' от обеих частей неравенства (1.13). Получим неравенство для случайных величин, зависящих от первой выборки S:

$$1_{\operatorname{err}_S(h_S)=0\ \&\ \operatorname{err}_P(h_S)>\varepsilon}P^l\Big\{S'\colon \operatorname{err}_P(h_S)-\operatorname{err}_{S'}(h_S)\leqslant\frac{1}{2}\varepsilon\Big\}\leqslant$$
$$\leqslant P^l\Big\{S'\colon \operatorname{err}_S(h_S)=0\ \&\ \operatorname{err}_{S'}(h_S)>\frac{1}{2}\varepsilon\Big\}. \quad (1.14)$$

Используя свойства биномиального распределения, получаем

$$P^l\Big\{S'\colon \operatorname{err}_P(h_S)-\operatorname{err}_{S'}(h_S)\leqslant\frac{1}{2}\varepsilon\Big\}=P^l\Big\{S'\colon \operatorname{err}_{S'}(h_S)\geqslant\operatorname{err}_P(h_S)-\frac{1}{2}\varepsilon\Big\}=$$
$$=\sum_{\{k\colon k/l\geqslant p-\varepsilon/2\}}\binom{l}{k}p^k(1-p)^{l-k}>\frac{1}{2} \quad (1.15)$$

при $l>2/\varepsilon$. Здесь $p=\operatorname{err}_P(h_S)$.

Действительно, при $l>2/\varepsilon$ будет $p-\varepsilon/2<p-1/l$. Поэтому достаточно доказать, что

$$\sum_{\{k\colon k/l\geqslant p-1/l\}}\binom{l}{k}p^k(1-p)^{n-k}=\sum_{\{k\colon k\geqslant lp-1\}}\binom{l}{k}p^k(1-p)^{n-k}>\frac{1}{2}.$$

Это неравенство эквивалентно неравенству

$$\sum_{\{k\colon k<lp-1\}}\binom{l}{k}p^k(1-p)^{n-k}<\frac{1}{2}.$$

[1)] Здесь h_S — произвольная и $1_{\operatorname{err}_S(h_S)=0\ \&\ \operatorname{err}_P(h_S)>\varepsilon}=0$, если S не лежит в множестве из левой части неравенства (1.12)

Делаем замену переменных в этой сумме:

$$\sum_{\{k:\,k<lp-1\}} \binom{l}{k} p^k (1-p)^{n-k} = \sum_{\{k:\,l-k>l(1-p)+1\}} \binom{l}{k} p^k (1-p)^{n-k} = \\ = \sum_{\{k:\,k>lp+1\}} \binom{l}{k} p^k (1-p)^{n-k}. \quad (1.16)$$

Сумма первой и третьей сумм из (1.16) меньше 1. Поэтому каждая из них меньше $1/2$.

Подставляя неравенство (1.15) в (1.14), получим

$$1_{\mathrm{err}_S(h_S)=0\ \&\ \mathrm{err}_P(h_S)>\varepsilon} \leqslant 2P^l \Big\{ S' \colon \mathrm{err}_S(h_S) = 0\ \&\ \mathrm{err}_{S'}(h_S) > \frac{1}{2}\varepsilon \Big\}. \quad (1.17)$$

Возьмем среднее по S и получим

$$P^l \big\{ S \colon \mathrm{err}_S(h_S) = 0\ \&\ \mathrm{err}_P(h_S) > \varepsilon \big\} \leqslant \\ \leqslant 2P^{2l} \Big\{ SS' \colon \mathrm{err}_S(h_S) = 0\ \&\ \mathrm{err}_{S'}(h_S) > \frac{1}{2}\varepsilon \Big\} \leqslant \\ \leqslant 2P^{2l} \Big\{ SS' \colon \sup_{h \colon \mathrm{err}_S(h)=0} \mathrm{err}_{S'}(h) > \frac{1}{2}\varepsilon \Big\}. \quad (1.18)$$

Отсюда получаем (1.12). Лемма доказана. □

Лемма 1.3. *Вероятность того, что на двух случайных выборках S и S' длины l некоторая функция классификации $h \in H$ согласована с первой из них и совершает более εl ошибок на второй выборке ограничена величиной*

$$P^{2l} \big\{ SS' \colon (\exists h \in H)(\mathrm{err}_S(h) = 0\ \&\ \mathrm{err}_{S'}(h) > \varepsilon) \big\} \leqslant B_H(2l) e^{-\varepsilon l/2}.$$

Доказательство. Определим функцию η, которая по произвольной выборке $SS' = \big((x_1, y_1), \ldots, (x_{2l}, y_{2l})\big)$ длины $2l$ выдает ее состав, т. е. множество пар ее составляющих вместе с кратностями:

$$\eta(SS') = \big\{ ((x_1, y_1), k_1), \ldots, ((x_L, y_L), k_L) \big\},$$

где k_i — число вхождений пары (x_i, y_i) в выборку SS', $i = 1, \ldots, L$, L — число различных пар (x_i, y_i) в выборке SS'; по определению $k_1 + \ldots + k_L = 2l$.

В отличие от выборки ее состав — неупорядоченное множество. Мера P^{2l} на выборках длины $2l$ индуцирует меру $\widehat{P}$ на их составах:

$$\widehat{P}(\Xi) = P^{2l} \big\{ SS' \colon \eta(SS') \in \Xi \big\},$$

где Ξ — множество, состоящее из составов.

Далее временно:

- фиксируем некоторый состав Υ для выборок длины $2l$;

- фиксируем некоторую функцию классификации h; пусть функция h делает m ошибок на всех выборках с составом Υ.

Для каждой двойной выборки $SS' = \big((x_1, y_1), \ldots, (x_{2l}, y_{2l})\big)$ с составом Υ определим бинарную последовательность $\varepsilon_1, \ldots, \varepsilon_{2l}$ ошибок классификации, где

$$\varepsilon_i = \begin{cases} 1, & \text{если} \quad h(x_i) \neq y_i, \\ -1, & \text{если} \quad h(x_i) = y_i, \end{cases}$$

$i = 1, \ldots, 2l$.

Поскольку ошибки классификации описываются бернуллиевским распределением с вероятностью ошибки $p = P\{h(x) \neq y\}$, любые два набора $\varepsilon_1, \ldots, \varepsilon_{2l}$ и $\varepsilon'_1, \ldots, \varepsilon'_{2l}$, описывающих распределение m ошибок на двух выборках с одним и тем же составом, Υ равновероятны.

Поэтому вероятность того, что на некоторой двойной выборке SS', имеющей состав Υ, все ошибки сосредоточены на второй половине этой выборки, оценивается сверху:

$$\frac{\binom{l}{m}}{\binom{2l}{m}} = \frac{l!}{(l-m)!m!} \cdot \frac{(2l-m)!m!}{(2l)!} =$$

$$= \frac{(2l-m)\ldots(l-m+1)}{2l\ldots(l+1)} \leqslant \left(1 - \frac{m}{2l}\right)^l \leqslant \left(1 - \frac{\varepsilon}{2}\right)^l < e^{-\varepsilon l/2} \qquad (1.19)$$

при $m \geqslant \varepsilon l$.

Можно заменить в (1.10) функции классификации $h \in H$, которые делают $m \geqslant \varepsilon l$ ошибок, на функции, которые получаются ограничением области определения функций из H на множество всех объектов $\{x_1, \ldots, x_{2l}\}$ из выборок SS' данного состава $\eta(SS') = \Upsilon$. Их число не превосходит числа элементов множества

$$\big\{(h(x_1), h(x_2), \ldots, h(x_{2l})) \colon h \in H\big\},$$

состоящего из бинарных последовательностей длины $2l$.

Оценку числа таких наборов дает *функция роста* семейства индикаторных функций H:

$$B_H(l) = \max_{(x_1, x_2, \ldots, x_l)} \Big|\big\{(h(x_1), h(x_2), \ldots, h(x_l)) \colon h \in H\big\}\Big|.$$

Ясно, что $B_H(l) \leqslant 2^l$. Точные оценки функции роста различных семейств классификаторов будут даны в п. 1.2.2.

Из определения функции роста следует, что число всех ограничений функций классификации из H на выборках длины $2l$ не превосходит $B_H(2l)$.

Поэтому условная вероятность того, что некоторая функция классификации из класса H делает более εl ошибок на двойной выборке с данным составом Υ и все они сосредоточены на второй половине этой выборки, ограничена сверху:

$$P^{2l}\{SS' : (\exists\, h \in H)(\mathrm{err}_S(h) = 0 \;\&\; \mathrm{err}_{S'}(h) > \varepsilon) \mid \eta(SS') = \Upsilon\} \leqslant B_H(2l)e^{-\varepsilon l/2}.$$

Левая часть этого неравенства представляет собой случайную величину (функцию от состава Υ). Правая часть неравенства не зависит от состава Υ.

Интегрируя это неравенство по мере $\widehat{P}$ на составах Υ, получим безусловное неравенство:

$$P^{2l}\{SS' : (\exists\, h \in H)(\mathrm{err}_S(h) = 0 \;\&\; \mathrm{err}_{S'}(h) > \varepsilon)\} \leqslant B_H(2l)e^{-\varepsilon l/2}.$$

Лемма 1.3 доказана. □

Теорема 1.2 непосредственно следует из лемм 1.2 и 1.3. □

Из теоремы 1.2 следует, что всякая гипотеза классификации h, имеющая ошибку $\mathrm{err}_P(h) > \varepsilon$, с вероятностью не меньше $1 - 2B_H(2l)e^{-\varepsilon l/4}$ не будет согласована со случайной выборкой длины $l > 2/\varepsilon$, т. е. будет отвергнута как ошибочная.

Обозначим $\delta = 2B_H(2l)e^{-\varepsilon l/4}$. Тогда при $0 < \delta < 1$ будет выполнено $l\varepsilon > 2$, т. е. условие теоремы 1.2 выполнено. Отсюда получаем следствие.

Следствие 1.1. *Допустим, что класс H функций классификации имеет конечную VC-размерность*[1] *d.*

Пусть задан критический уровень $0 < \delta < 1$ принятия ошибочной гипотезы классификации $h \in H$, согласованной с обучающей выборкой S.

Тогда при $l \geqslant d$ с вероятностью не ниже $1 - \delta$ гипотеза классификации $h_S \in H$, построенная по случайной обучающей выборке S и согласованная с ней, будет иметь ошибку классификации

$$\mathrm{err}_P(h_S) \leqslant \frac{4}{l}\left(d \ln \frac{2el}{d} + \ln \frac{2}{\delta}\right).$$

Все эти результаты можно усилить на случай обучения с ошибками. Аналогичным образом доказываются следующие две леммы 1.4 и 1.5, а также их следствие — теорема 1.3.

Лемма 1.4. *Пусть задан класс H функций классификации. Рассматриваются две случайные выборки S, S' длины l. Тогда для любого $\varepsilon > 0$*

[1] Определение VC-размерности дано в п. 1.2.2. Там же получена оценка $B_H(l) \leqslant \left(\frac{el}{d}\right)^d$ при $l \geqslant d$.

при $l > 2/\varepsilon^2$ *имеет место неравенство*

$$P^l\{S\colon (\exists\, h\in H)(\mathrm{err}_P(h)-\mathrm{err}_S(h)>\varepsilon)\}\leqslant$$
$$\leqslant 2P^{2l}\Big\{SS'\colon (\exists\, h\in H)\Big(\mathrm{err}_{S'}(h)-\mathrm{err}_S(h)>\frac{1}{2}\varepsilon\Big)\Big\}.$$

Доказательство этой леммы аналогично доказательству леммы 1.3.

Лемма 1.5. *Вероятность того, что на двух случайных выборках* S *и* S' *длины* l *частоты ошибок некоторой функции классификации* $h\in H$ *различаются более чем на* $\varepsilon>0$*, ограничена величиной*

$$P^{2l}\{SS'\colon (\exists h\in H)(\mathrm{err}_{S'}(h)-\mathrm{err}_S(h)>\varepsilon)\}\leqslant 2B_H(2l)e^{-2\varepsilon^2 l}.$$

Доказательство этой леммы аналогично доказательству леммы 1.5.

Следующая теорема дает оценку вероятности отклонения риск-функционала от среднего числа ошибок на обучающей выборке.

Теорема 1.3. *Имеет место оценка*

$$P^l\{S\colon (\exists\, h\in H)(\mathrm{err}_P(h)-\mathrm{err}_S(h)>\varepsilon)\}\leqslant 4B_H(2l)e^{-\varepsilon^2 l/2}$$

при $l>2/\varepsilon^2$.

Отсюда получаем следствие, связывающее вероятность ошибки обобщения и среднее число ошибок на обучающей выборке.

Следствие 1.2. *Допустим, что класс* H *функций классификации имеет конечную VC-размерность* d*,* $0<\delta<1$ *и* $l\geqslant d$*. Тогда с вероятностью не менее* $1-\delta$ *для* $h\in H$ *выполнено*

$$\mathrm{err}_P(h)\leqslant \mathrm{err}_S(h)+\sqrt{\frac{2}{l}\Big(d\ln\frac{2el}{d}+\ln\frac{4}{\delta}\Big)}.$$

Следует отметить, что оценки теорем 1.2 и 1.3, а также следствий 1.1 и 1.2 имеют в основном теоретическое значение, так как на практике VC-размерность d может быть сравнимой с длиной выборки l. Ближе к практике находятся оценки, не зависящие от размерности пространства (см. теорему 1.9).

1.2.2. VC-размерность

В этом пункте мы рассмотрим определение и свойства размерности Вапника–Червоненкиса — VC-размерности, которая характеризует «сложность» бесконечного класса функций классификации.

Пусть $\mathscr{X}$ — множество объектов и H — произвольный класс функций классификации на $\mathscr{X}$. Рассмотрим функцию $h\in H$ и произвольный набор — *выборку* элементов $(x_1,\ldots,x_l)$ из $\mathscr{X}$.

Бинарный набор $(h(x_1),\ldots,h(x_l))$, состоящий из элементов множества $\{-1,1\}$, определяет разделение множества $\{x_1,\ldots,x_l\}$ на два подмножества: $\{x_i\colon h(x_i)=1\}$ — положительные примеры и $\{x_i\colon h(x_i)=-1\}$ — отрицательные примеры.