Методы машинного обучения и поиск достоверных закономерностей в данных

(Распознавание и регрессионный анализ)

Сенько О.В.

Содержание

1. Введение.

1.1 Области применения.

. Задачи диагностики и прогнозирования некоторой величины Y по доступным значениям переменных $X_1,\ldots,X_n$ часто возникают в различных областях человеческой деятельности. В частности могут быть упомянуты:

- задачи диагностики хода технологического процесса по показаниям различных датчиков;
- задачи диагностики состояния технического оборудования;
- задачи медицинской диагностики по совокупности клинических и лабораторных показателей;
- задачи прогнозирования свойств ещё не синтезированного химического соединения по его молекулярной формуле;
- прогноз значений финансовых индикаторов.

Для решения подобных задач могут быть использованы методы, основанные на использовании точных знаний. Например, могут использоваться методы математического моделирования, основанные на использовании физических законов. Однако сложность точных математических моделей нередко оказывается слишком высокой. Кроме того при использовании физических моделей часто требуется знание различных параметров, характеризующих рассматриваемое явление или процесс. Значения некоторых из таких параметров часто известны только приблизительно или неизвестны вообще. Все эти обстоятельства ограничивают возможности эффективного использования физических моделей.

В прикладных исследованиях нередко возникают ситуации, когда математическое моделирование, основанное на использовании точных законов оказывается затруднительны, но в распоряжении исследователей оказывается выборка прецедентов - результатов наблюдений исследуемого процесса или явления, включающих значения прогнозируемой величины Y и переменных $X_1,\ldots,X_n$. В этих случаях для решения задач диагностики и прогнозирования могут быть использованы методы, основанные на обучении по прецедентам.

1.1 Основные понятия.

Предположим, что задача прогнозирования решается для некоторого процесса или явления F. Множество объектов, которые потенциально могут возникать в рамках F, называется генеральной совокупностью, далее обозначаемой Ω.

Поиск алгоритма, вычисляющего осуществляется по выборке прецедентов, которая обычно является случайной выборкой объектов из Ω с известными значениями $Y, X_1,\ldots, X_n$, Выборку прецедентов также принято называть **обучающей выборкой.**

Обучающая выборка имеет вид $\tilde{S}_t = \{s_1 = (y_1, \mathbf{x}_1),\ldots, s_m = (y_m, \mathbf{x}_m)\}$,

где y_j - значение переменной для объекта s_j ;

$\mathbf{x}_j$ - вектора переменных $X_1,\ldots, X_n$ для объекта s_j ;

m - число объектов в $\tilde{S}_t$.

В процессе обучения производится поиск эмпирических закономерностей, связывающих прогнозируемую переменную Y с переменными $X_1,\ldots, X_n$.

Данные закономерности далее используются при прогнозировании.

Методы, основанные на обучении по прецедентам, также принято называть

Методами машинного обучения (Machine learning)

Прогнозируемая величина Y может иметь различную природу:

- принимать значения из отрезка непрерывной оси;
- принимать значения из конечного множества;
- являться кривой, описывающей вероятность возникновения некоторого критического события до различных моментов времени.

***Задачи распознавания*.** Задачи, в которых прогнозируемая величина принимает значения из множества, содержащего несколько элементов принято называть задачей распознавания, Например, к задачам распознавания относятся задачи прогнозирования категориальных переменных. Подмножества объектов с одинаковым значением Y обычно принято называть классами. Поэтому задача

распознавания часто формулируется в следующем виде. Предположим, что множество объектов Ω является объединением непересекающихся классов $K_1,\ldots,K_l$. Тогда задача распознавания состоит в поиске по обучающей выборке $\tilde{S}_t$ алгоритма, относящего произвольный объект s из множества Ω к одному из классов $K_1,\ldots,K_l$. При этом искомый номер класса вычисляется по описанию объекта s , представляющему собой вектор значений на s переменных $X_1,\ldots,X_n$

Задачи регрессии. Задачи, в которых прогнозируемая величина принимает значения из некоторого подмножества оси вещественных чисел $\mathbf{R}$, обычно принято называть задачами регрессии.

Приведём конкретный пример закономерности, которая может быть использована при решении задачи регрессии.

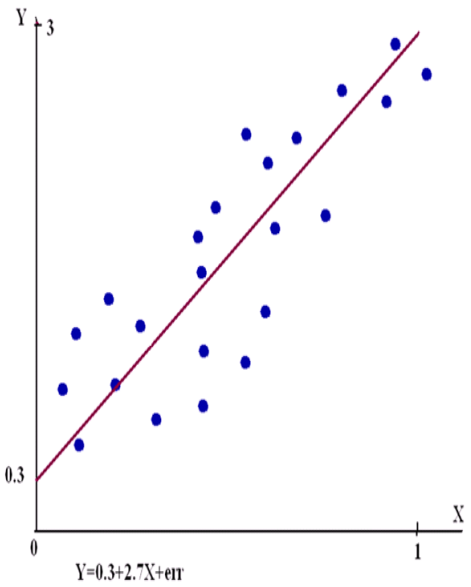


Рис. 1.1

На рисунке 1 изображена закономерность, описывающая линейную связь между переменными Y и X . Видно, что график линейной функции $0.3+2X$ проходит достаточно близко к значениям переменной Y . Поэтому данная функция может быть использована для предсказания значений Y по соответствующим значениям X .

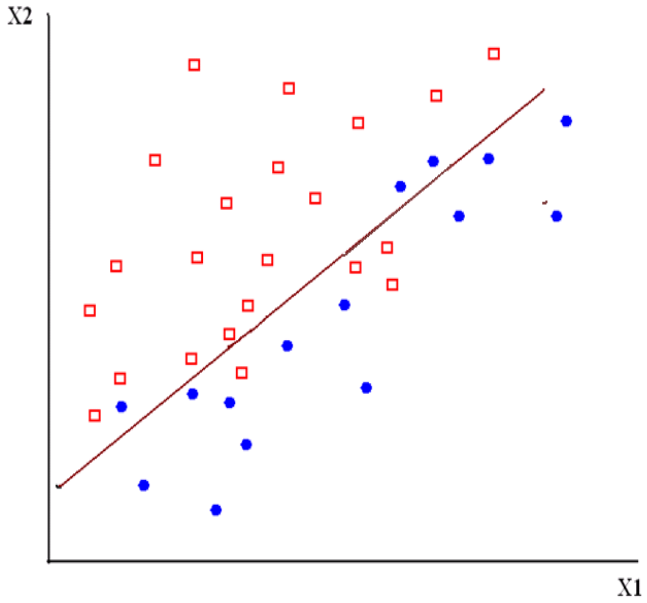


Рис. 1.2

На рисунке 2 показана закономерность, которая может быть использована для решения задачи распознавания: объекты класса K1 (обозначены ▫) и класса K2 (обозначены ●) из обучающей выборке St находятся по разные стороны прямой, проходящей через точки .

1.3. Методы распознавания и регрессии с максимальной обобщающей способностью

Для каждой задачи регрессии или распознавания существует объективно оптимальный метод, для которого обобщающая способность объективно является наилучшей. Для задач регрессионного анализа оптимальным является алгоритм A, вычисляющий прогноз Y для произвольного вектора $\mathbf{x}$ равный условному математическому ожиданию Y в точке $\mathbf{x}$: $A(\mathbf{x}) = E(Y \mid \mathbf{x})$.

Для задач распознавания наиболее высокой точностью обладает байесовский классификатор. Пусть в точке $\mathbf{x} \in \mathbf{R}^n$ объекты из классов $K_1,\ldots,K_L$ встречаются с вероятностями $\mathbf{P}(K_1 \mid \mathbf{x}),\ldots,\mathbf{P}(K_L \mid \mathbf{x})$.

Тогда распознаваемый объект со значением вектора прогностических переменных $\mathbf{x}$ может быть отнесён в класс $K_{i'}$ только в случае выполнения набора неравенства $\mathbf{P}(K_{i'} \mid \mathbf{x}) \geq \mathbf{P}(K_i \mid \mathbf{x})$ при $i \in \{1,\ldots,l\}$. Иными словами распознаваемый объект может быть отнесён к одному из классов, вероятность принадлежности которому в точке

$\mathbf{x}$ максимальна . Таким образом, максимально точное решение задач регрессии может быть легко получено, если в каждой точке известны условных математических ожиданий $E(Y \mid \mathbf{x})$. Аналогично максимально точное решение задач распознавания может быть получено при знании условных вероятностей $\mathbf{P}(K_1 \mid \mathbf{x}), \ldots, \mathbf{P}(K_l \mid \mathbf{x})$.

1.4 Способы обучения.

1.4.1 Метод максимального правдоподобия.

Условные математические ожидания $E(Y \mid \mathbf{x})$ могут быть вычислены, когда известна плотность совместного распределения переменных $Y, X_1, \ldots, X_n$ - $p(Y, X_1, \ldots, X_n)$. Условные вероятности $\mathbf{P}(K_1 \mid \mathbf{x}), \ldots, \mathbf{P}(K_l \mid \mathbf{x})$ могут быть вычислены, когда известны плотности совместного распределения переменных $X_1, \ldots, X_n$ для каждого из классов $K_1, \ldots, K_l$, а также вероятности каждого из классов во всей генеральной совокупности. Плотности совместного распределения принципе могут быть получены с использованием известного метода максимального правдоподобия.

Метод максимального правдоподобия (ММП) используется в математической статистике для аппроксимации вероятностных распределений по выборкам данных. В общем случае ММП требует априорных предположений о типе распределений. Чаще всего используется гипотеза о нормальности распределения. Значения параметров $\theta_1, \ldots, \theta_r$, задающих конкретный вид распределений, ищутся путём максимизации функционала правдоподобия, представляющего собой произведение плотностей вероятностей на объектах обучающей выборки. Рассмотрим в качестве примера задачу распознавания двух классов K_1 и K_2 по единственному признаку X . При этом предполагается, что классы K_1 и K_2 распределены нормально, то есть точки, описывающие объекты из данных классов, имеют плотности распределения $p_1(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu_1)^2}{2\sigma^2}}$ и $p_2(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu_2)^2}{2\sigma^2}}$ соответственно при значении признака X равном x. До пустим, что нам неизвестны параметры μ_1 и μ_2 , являющиеся математическими

ожиданиями признака X в классах K_1 и K_2. Математические ожидания μ_1 и μ_2 могут быть найдены с помощью ММП по обучающей выборке $\tilde{S}_t = \{s_1 = (y_1, x_1), \ldots, s_m = (y_m, x_m)\}$, где $y_j = 1$ при $s_j \in K_1$ и $y_j = 2$ при $s_j \in K_2$. При этом подбираются такие значения μ_1 и μ_2, при которых достигают максимума функционалы правдоподобия

$$L_1(\tilde{S}_t, \mu_1) = \prod_{s_j \in K_1} p_1(x_j) = \prod_{s_j \in K_1} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu_1)^2}{2\sigma^2}} \qquad (1)$$

И

$$L_2(\tilde{S}_t, \mu_1) = \prod_{s_j \in K_2} p_2(x_j) = \prod_{s_j \in K_2} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu_2)^2}{2\sigma^2}} \qquad (2)$$

соответственно. То есть функционал $L_1(\tilde{S}_t, \mu_1)$ является произведением плотностей вероятности в точках, соответствующим объектам обучающей выборки из класса K_1, функционал $L_2(\tilde{S}_t, \mu_1)$ является произведением плотностей вероятности в точках, соответствующим объектам обучающей выборки из класса K_2.

Поскольку натуральный логарифм $\ln(z)$ является монотонной функцией аргумента z, то задача максимизации $L_1(\tilde{S}_t, \mu_1)$ эквивалентна задаче максимизации $\ln[L_1(\tilde{S}_t, \mu_1)]$. Отметим, что

$$\ln[L_1(\tilde{S}_t, \mu_1)] = \prod_{s_j \in K_1}^{m} \ln[\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_j-\mu_1)^2}{2\sigma^2}}] = \sum_{s_j \in K_1} \{\ln(\frac{1}{\sqrt{2\pi}\sigma}) + \ln[e^{\frac{-(x_j-\mu_1)^2}{2\sigma^2}}]\} =$$

$$= -m_1 \ln(\sigma) - \tfrac{1}{2} m_1 \ln(2\pi) - \sum_{s_j \in K_1} \frac{(x_j - \mu_1)^2}{2\sigma^2}, \qquad (3)$$

где m_1 - число объектов из класса K_1 в выборке $\tilde{S}_t$. Из формулы (3) следует, что задача максимизации $\ln[L_1(\tilde{S}_t, \mu_1)]$ сводится к задаче минимизации

$Q(\tilde{S}_t, \mu_1) = \sum_{s_j \in K_1} \frac{(x_j - \mu_1)^2}{2\sigma^2}$. Необходимым условием минимума $Q(\tilde{S}_t, \mu_1)$ является выполнение равенства $\frac{\partial Q(\tilde{S}_t, \mu_1)}{\partial \mu_1} = 0$, что эквивалентно выполнение равенства

$$\sum_{s_j \in K_1} \frac{2(x_j - \mu_1)}{2\sigma^2} = 0. \qquad (4)$$

Из равенства (4) следует, что $\mu_1 = \frac{1}{m_1} \sum_{s_j \in K_1} x_j$. То есть применение ММП приводит к тривиальному выводу о равенстве параметра μ_1 среднему значению признака X по всем объектам обучающей выборки из класса K_1. Очевидно, что поиск оптимального значения параметра μ_2 с помощью ММП совершенно аналогичен поиску оптимального значения параметра μ_1 и приводит к одинаковому результату.

В общем случае нам требуется найти параметры $\theta_1, \ldots, \theta_r$ совместного распределения переменных $Y, X_1, \ldots, X_n$. Данная задача может быть решена с помощью максимизации функционал правдоподобия

$$L(\tilde{S}_t, \theta_1, \ldots, \theta_r) = \prod_{j=1}^{m} p(y_j, \mathbf{x}_j, \theta_1, \ldots, \theta_r),$$

(5)

который является произведением плотностей вероятностей в точках, соответствующих объектам обучающей выборки $\tilde{S}_t = \{s_1 = (y_1, \mathbf{x}_1), \ldots, s_m = (y_m, \mathbf{x}_m)\}$. Метод ММП является одним из важнейших инструментов настройки алгоритмов распознавания или регрессионных моделей в математической статистике. Однако использованием ММП требует знания вида вероятностного распределения. На практике чаще используется метод минимизации эмпирического риска, который требует знания только общего вида алгоритма прогнозирования.

Метод минимизации эмпмрмческого риска (ММЭР). Основным способом поиска закономерностей является поиск некотором априори заданном семействе алгоритмов прогнозирования $\tilde{M} = \{A : \tilde{X} \to \tilde{Y}\}$ алгоритма, наилучшим образом

аппроксимирующего связь переменных из набора $X_1,\ldots,X_n$ с переменной Y на обучающей выборке, где $\tilde{X}$ - область возможных значений векторов переменных $X_1,\ldots,X_n$, $\tilde{Y}$ - область возможных значений переменной Y. Отметим, что чаще всего алгоритм A задаётся с помощью прогнозирующей функции.

Пусть $\lambda[y_j, A(\mathbf{x}_j)]$ - величина “потерь”, произошедших в результате использования в качестве прогноза величины $A(\mathbf{x}_j)$. Одним из способов обучения является минимизация на обучающей выборке функционала эмпирического риска $Q(\tilde{S}_t, A) = \frac{1}{m}\sum_{j=1}^{m} \lambda[y_j, A(\mathbf{x}_j)]$

Приведём примеры конкретного вида функции потерь $\lambda[y_j, A(\mathbf{x}_j)]$ **.** В задачах регрессии чаще всего используется квадрат ошибки прогноза $\lambda[y_j, A(\mathbf{x}_j)] = [y_j - A(\mathbf{x}_j)]^2$. Также может быть использован модуль ошибки $\lambda[y_j, A(\mathbf{x}_j)] = |[y_j - A(\mathbf{x}_j)]|$.

В случае задачи распознавания функция потерь может быть равной 0 при правильной классификации и 1 при ошибочной. При этом функционал эмпирического риска равен числу ошибочных классификаций.

Следует отметить тесную связь между ММП и ММЭР. Данная связь буде проанализирована далее при рассмотрении методов регрессионного анализа.

Точность алгоритма прогнозирования на всевозможных новых не использованных для обучения объектах, которые возникают в результате процесса, соответствующего рассматриваемой задаче прогнозирования принято называть обобщающей способностью. Иными словами обобщающую способность алгоритма прогнозирования можно определить как точность на всей генеральной совокупности. Мерой обобщающей способности служит математическое ожидание потерь по генеральной совокупности - $E_\Omega\{\lambda[Y, A(\mathbf{x})]\}$. При решении задач прогнозирования основной целью является достижение наилучшей обобщающей способности, при которой математическое ожидание потерь $E_\Omega\{\lambda[Y, A(\mathbf{x})]\}$ минимально.

1.5 Эффект переобучения.

Расширение модели $\tilde{M} = \{A : \tilde{X} \to \tilde{Y}\}$, увеличение её сложности всегда приводит к повышению точности аппроксимации на обучающей выборке. Однако повышение точности на обучающей выборке, связанное с увеличением сложности модели, часто не ведёт к увеличению обобщающей способности. Более того, обобщающая способность может даже снижаться. Различие между точностью на обучающей выборке и обобщающей способностью при этом возрастает. Данный эффект называется эффектом переобучения.

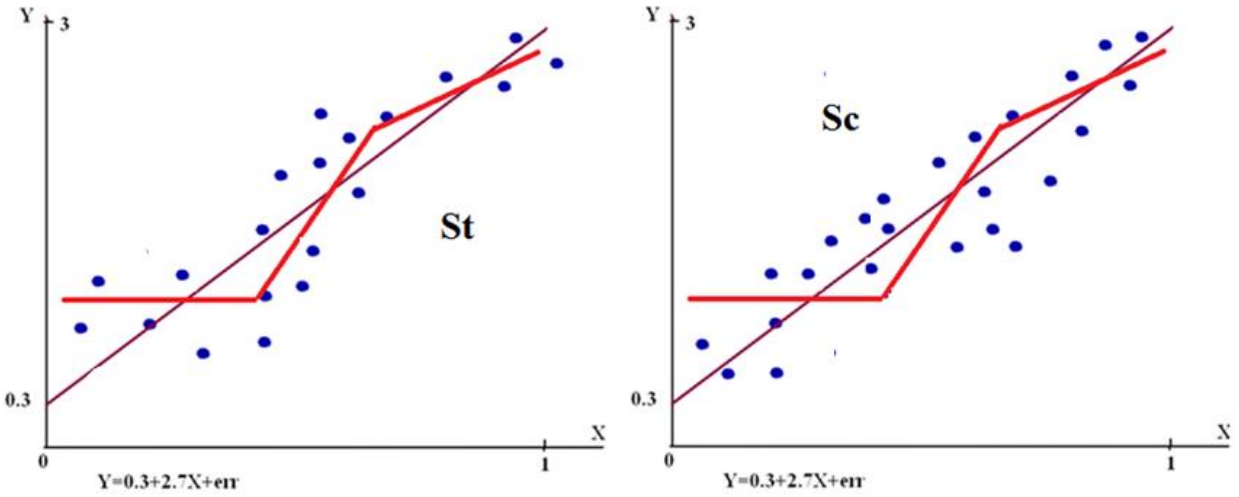


Рис. 1.4

На левой части показано, что использование кусочно-линейной модели (красная линия)позволяет значительно лучше аппроксимировать зависимость на обучающей выборке S_t, чем простая линейная регрессия (тёмно-синяя прямая). Однако оказывается (правый слайд), что точность аппроксимации новой контрольной выборки S_c, взятой из той же самой генеральной совокупности, для простой линейной регрессии значительно лучше, чем для кусочно-линейной.

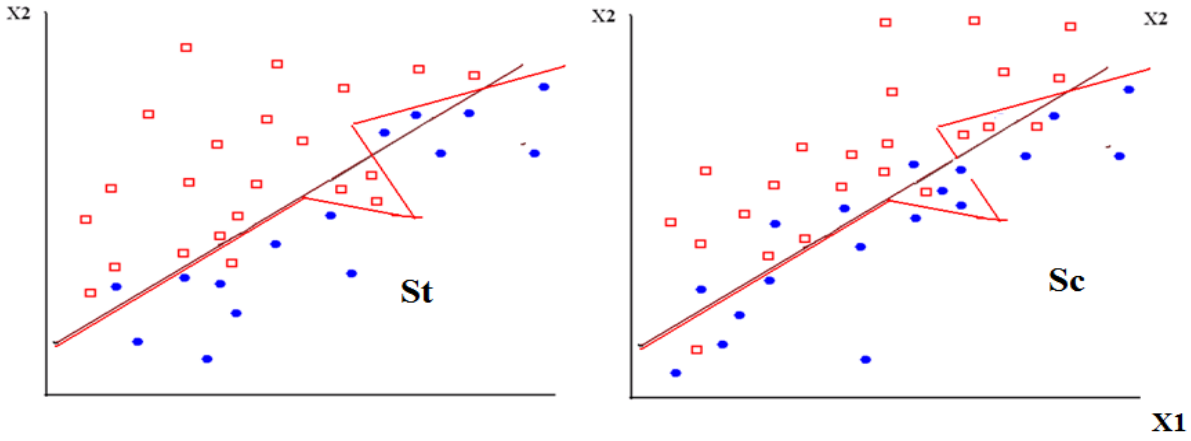


Рис. 1.5

На левой части показано, что использование кусочно-линейной границы (красная линия)позволяет значительно лучше разделить объекты класса K1

и класса K2 обучающей выборке S_t, чем простая линейная граница (тёмно-синяя прямая). Однако оказывается (правый слайд), что точность на новой контрольной выборки S_c, взятой из той же самой генеральной совокупности, для простой линейной границы значительно лучше, чем для кусочно-линейной

1.5 Методы оценивания обобщающей способности.

Обобщающая способность алгоритма прогнозирования на генеральной совокупности Ω, может оцениваться по случайной выборке объектов из Ω, которую обычно называют контрольной выборкой. При этом контрольная выборка не должна содержать объектов из обучающей выборки. В противном случае величина потерь может оказаться завышенной.

Контрольная выборка имеет вид $\tilde{S}_c = \{(y_1, \mathbf{x}_1), \ldots, (y_{m_c}, \mathbf{x}_{m_c})\}$, где

y_j - значение переменной Y для j-го объекта;

$\mathbf{x}_j$ - значение вектора переменных $X_1, \ldots, X_n$ для j-го объекта;

m_c - число объектов в $\tilde{S}_c$.

Обобщающая способность алгоритма прогнозирования A может оцениваться с помощью функционала риска

$$Q(\tilde{S}_c, A) = \frac{1}{m}\sum_{j=1}^{m_c} \lambda[y_j, A(\mathbf{x}_j)]$$

При $m_c \to \infty$ согласно закону больших чисел

$$Q(\tilde{S}_c, A) \quad \to \quad E_\Omega\{\lambda[Y, A(\mathbf{x})]\}$$

Обычно при решении задачи прогнозирования по прецедентам в распоряжении исследователей сразу оказывается весь массив существующих эмпирических данных $\tilde{S}_{in}$, по которому необходимо построить алгоритм прогнозирования и оценить его точность. Для оценки точности прогнозирования могут быть использованы следующие стратегии.

1) Выборка $\tilde{S}_{in}$ случайным образом расщепляется на выборку $\tilde{S}_t$ для обучения алгоритма прогнозирования и выборку $\tilde{S}_c$ для оценки точности

2) Процедура кросс-проверки. Выборка $\tilde{S}_{in}$ случайным образом расщепляется на выборки $\tilde{S}_1$ и $\tilde{S}_2$. На первом шаге $\tilde{S}_1$ используется для обучения и $\tilde{S}_1$ для контроля. На втором шаге, наоборот, для обучения используется $\tilde{S}_2$, а $\tilde{S}_1$ используется для контроля.

3) Процедура скользящего контроля выполняется по полной выборке $\tilde{S}_{in}$ за шагов $m = |\tilde{S}_{in}|$.

на j -ом шаге формируется обучающая выборка $\tilde{S}_t = \tilde{S}_t^j \setminus s_j$, где s_j - j- ый объект в $\tilde{S}_{in}$, и контрольная выборка , состоящая из единственного объекта s_j . Величина потерь методе скользящий контроль оценивается с помощью функционала

$$Q_{sc}(\tilde{S}_{in}, A) = \frac{1}{m}\sum_{j=1}^{m} \lambda[y_j, A(\mathbf{x}_j, \tilde{S}_t^j)]$$

В книге [1] было показано, что функционал $Q_{sc}(\tilde{S}_{in}, A)$ является несмещённой оценкой математического ожидания потерь

1.6 Существующие методы и модели для решения задач прогнозирования и распознавания

Для подавляющего числа приложений вид распределений или значения конкретных их параметров неизвестны. Не известен обычно также вид регрессионной зависимости, или разделяющей поверхности в задачах распознавания. В связи с эти возникло большое число разнообразных подходов, в которых поиск оптимальных алгоритма прогнозирования производится внутри достаточно обширных семейств (моделей). Обычно такие семейства задаются с помощью набора параметров. Для поиска оптимальных значений параметров используются ММП или ММЭР. При этом для повышения устойчивости обучения нередко используются модифицированные варианты ММП или ММЭР, позволяющие добиваться более высокой устойчивости обучения. Использование которы данных подходов позволяет добиваться определённых успехов при решении конкретных задач. Для решения задач распознавания часто используются

- статистические методы, включая байесовские метод;
- методы, основанные на линейной разделимости;
- методы, основанные на ядерных оценках;
- нейросетевые методы;
- комбинаторно-логические методы и алгоритмы вычисления оценок;
- алгебраические методы;
- решающие деревья и леса;
- методы, основанные на принятии коллективных решений по системам закономерностей
- методы, основанные на опорных векторах.

Для решения задач регрессии используются

многомерная линейная регрессия;

ядерные оценки;

нейросетевые методы;

метод опорных векторов.

2. Линейная регрессия

2.1 Методы настройки моделей

Распространённым средством решения задач прогнозирования величины Y по переменным $X_1,\ldots,X_n$ является использование метода множественной линейной регрессии. В данном методе связь переменной Y с переменными $X_1,\ldots,X_n$ задаётся с помощью линейной модели

$$Y=\beta_0+\beta_1X_1+\ldots\beta_nX_n+\varepsilon,$$

где $\beta_0,\beta_1,\ldots,\beta_n$ вещественные регрессионные коэффициенты, ε - случайная величина, являющаяся ошибкой прогнозирования.

Регрессионные коэффициенты ищутся по обучающей выборке $\tilde{S}_t=\{s_1=(y_1,\mathbf{x}_1),\ldots,s_m=(y_m,\mathbf{x}_m)\}$, где y_j - значение прогнозируемой переменной Y, $\mathbf{x}_j=(x_{1j},\ldots,x_{nj})$ - вектор значений переменных $X_1,\ldots,X_n$, $j=1,\ldots,n$.

Предположим, что ошибка ε распределена нормально с нулевым ожиданием $\mu=0$ и стандартным отклонением σ. Откуда следует, что разность $Y-\beta_0+\beta_1X_1+\ldots\beta_nX_n$ также распределена нормально с нулевым ожиданием $\mu=0$ и стандартным отклонением σ. Откуда следует, что функционал правдоподобия (1.9) может быть записан в виде $L(\tilde{S}_t,\beta_0,\beta_1,\ldots,\beta_n)=\prod_{j=1}^{m}\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(y_j-\beta_0-\sum_{i=1}^{n}\beta_ix_{ij})^2}{2\sigma^2}}$.

Прологарифмировав функцию правдоподобия

$$\ln[L_1(\tilde{S}_t,\mu_1)]=\prod_{s_j\in K_1}^{m}\ln[\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(y_j-\beta_0-\sum_{i=1}^{n}\beta_ix_{ij})^2}{2\sigma^2}}]=$$

$$= \sum_{s_j \in K_1} \{\ln(\frac{1}{\sqrt{2\pi}\sigma}) + \ln[e^{\frac{-(y_j - \beta_0 - \sum_{i=1}^{n} \beta_i x_{ij})^2}{2\sigma^2}}]\} =$$

$$= -m_1 \ln(\sigma) - \tfrac{1}{2} m_1 \ln(2\pi) - \sum_{s_j \in K_1} \frac{-(y_j - \beta_0 - \sum_{i=1}^{n} \beta_i x_{ij})^2}{2\sigma^2}$$

Традиционным способом поиска регрессионных коэффициентов является метод наименьших квадратов (МНК). МНК заключается в минимизации функционала эмпирического риска с квадратичными потерями

$Q(\tilde{S}_t, \beta_0, \ldots, \beta_n) = \frac{1}{m} \sum_{j=1}^{m} [y_j - \beta_0 - \sum_{i=1}^{n} x_{ji} \beta_i]^2$. То есть оценки $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_n$ регрессионных коэффициентов $\beta_0, \beta_1, \ldots, \beta_n$ по методу МНК удовлетворяют условию $(\hat{\beta}_0, \ldots, \hat{\beta}_n) = \arg\min[Q(\tilde{S}_t, \beta_0, \ldots, \beta_n)]$. Очевидно, МНК является вариантом метода минимизации эмпирического риска с квадратичной функцией потерь. Покажем, что для задач, в которых величина случайной ошибки ε не зависит от переменных

2.2 Одномерная регрессия.

Рассмотрим простейший вариант линейной регрессии, описывающей связь между переменной Y и единственной переменной X: $Y = \beta_0 + \beta X + \varepsilon$. . Функционал эмпирического риска на выборке $\tilde{S}_t = \{(y_1, x_1), \ldots, (y_m, x_m)\}$ принимает вид $Q(\tilde{S}_t, \beta_0, \beta_1) = \frac{1}{m} \sum_{j=1}^{m} [y_j - \beta_0 - \beta_1 x_j]^2$.

Необходимым условием минимума функционала $Q(\tilde{S}_t, \beta_0, \beta_1)$ является выполнение системы из двух уравнений

$$\frac{\partial Q(\tilde{S}_t, \beta_0, \beta_1)}{\partial \beta_0} = -\frac{2}{m} \sum_{j=1}^{m} y_j + 2\beta_0 + \frac{2\beta_1}{m} \sum_{j=1}^{m} x_j = 0$$

(2)

$$\frac{\partial Q(\tilde{S}_t, \beta_0, \beta_1)}{\partial \beta_1} = -\frac{2}{m}\sum_{j=1}^{m} x_j y_j + 2\beta_0 \sum_{j=1}^{m} x_j + \frac{2\beta_1}{m}\sum_{j=1}^{m} x_j^2 = 0$$

Оценки $(\hat{\beta}_0, \hat{\beta}_1)$ являются решением системы (2) относительно параметров (β_0, β_1) соответственно .

Таким образом оценки могут быть записаны в виде

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{m} x_j y_j - \frac{1}{m}\sum_{j=1}^{m} x_j \sum_{j=1}^{m} y_j}{\sum_{j=1}^{m} x_j^2 - \frac{1}{m}(\sum_{j=1}^{m} x_j)^2}, \quad \hat{\beta}_0 = \overline{y} - \beta_1 \overline{x}, \text{ где } \overline{y} = \frac{1}{m}\sum_{j=1}^{m} y_j, \quad \overline{x} = \frac{1}{m}\sum_{j=1}^{m} x_j$$

Выражение для $\hat{\beta}_1$ может быть переписано в виде $\hat{\beta}_1 = \frac{Cov(Y, X \mid \tilde{S}_t)}{D(X)}$, где

$Cov(Y, X \mid \tilde{S}_t) = \frac{1}{m}\sum_{j=1}^{m}(y_j - \overline{y})(x_j - \overline{x})$ является выборочной ковариацией переменных Y и X, $D(X \mid \tilde{S}_t) = \frac{1}{m}\sum_{j=1}^{m}(x_j - \overline{x})^2$ - выборочная дисперсия переменной X.

2.3 Многомерная регрессия.

При вычислении оценки вектора параметров $\beta_0, \ldots, \beta_n$ в случае многомерной линейной регрессии удобно использовать матрицу плана $\mathbf{X}$ размера $m \times (n+1)$,

которая строится по обучающей выборке $\tilde{S}_t = \{(y_1, \mathbf{x}_1), \ldots, (y_m, \mathbf{x}_m)\}$, где $\mathbf{x}_j = (x_{j1}, \ldots, x_{jm})$ - вектор значений переменных $X_1, \ldots, X_n$. Матрица плана имеет вид $\mathbf{X} = \begin{pmatrix} 1\, x_{11} \ldots\ldots\ldots x_{1n} \\ \ldots\ldots\ldots\ldots\ldots \\ 1\, x_{j1} \ldots\ldots\ldots x_{jn} \\ \ldots\ldots\ldots\ldots\ldots \\ 1\, x_{m1} \ldots\ldots\ldots x_{mn} \end{pmatrix}$.

Пусть $\mathbf{y} = (y_1, \ldots, y_m)$ - вектор значений переменной Y . Связь значений Y с переменными $X_1, \ldots, X_n$ на объектах обучающей выборки может быть описана с помощью матричного уравнения $\mathbf{y} = \boldsymbol{\beta}\mathbf{X}^t + \boldsymbol{\varepsilon}$, где $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_m)$ - вектор ошибок прогнозирования для объектов $\tilde{S}_t$.

Функционал $Q(\tilde{S}_t, \beta_0, \ldots, \beta_n)$ может быть записан в виде

$Q(\tilde{S}_t, \beta_0, \ldots, \beta_n) = \frac{1}{m} \sum_{j=1}^{m} [y_j - \sum_{i=1}^{n+1} \beta_i \hat{x}_{ji}]^2$, где $\hat{x}_{ji}$ - элементы матрицы плана $\mathbf{X}$, определяемые равенствами $\hat{x}_{j1} = 1$, $\hat{x}_{j1} = x_{j(i-1)}$ при $i > 1$.

Необходимым условием минимума функционала $Q(\tilde{S}_t, \beta_0, \ldots, \beta_n)$ является выполнение системы из $n+1$ уравнений

$$\begin{aligned} &\frac{\partial Q(\tilde{S}_t, \beta_0, \ldots, \beta_n)}{\partial \beta_0} = 2[\sum_{j=1}^{m} y_j \hat{x}_{j1} - \sum_{j=1}^{m} \sum_{i=1}^{n+1} \beta_i \hat{x}_{ji} \hat{x}_{j1}] = 0 \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &\frac{\partial Q(\tilde{S}_t, \beta_0, \ldots, \beta_n)}{\partial \beta_n} = 2[\sum_{j=1}^{m} y_j \hat{x}_{jn} - \sum_{j=1}^{m} \sum_{i=1}^{n+1} \beta_i \hat{x}_{ji} \hat{x}_{jn}] = 0 \end{aligned} \qquad (3)$$

Вектор оценок значений регрессионных коэффициентов $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_n)$ является решением системы уравнений (3) . В матричной форме система (3) может быть записана в виде

$$-2\mathbf{X}^t\mathbf{y}^t + 2\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}^t = 0 \qquad (4)$$

Решение системы (4) существует, если $\det(\mathbf{X}^t\mathbf{X}) \neq 0$. В этом случае для $\mathbf{X}^t\mathbf{X}$ существует обратная матрица и решение (4) относительно вектора может быть записано в виде: $\hat{\boldsymbol{\beta}}^t = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t y^t$. Из теории матриц следует, что $\det(\mathbf{X}^t\mathbf{X}) = 0$ если ранг матрицы $\mathbf{X}$ по строкам менее $n+1$, что происходит, если m-мерный вектор значений одной из переменных $X_{i'} \in \{X_1, \ldots, X_n\}$ на выборке $\tilde{S}_t$ является линейной комбинаций m-мерных векторов значений на $\tilde{S}_t$ других переменных из $\{X_1, \ldots, X_n\}$. При сильной коррелированности m-мерного вектора значений одной из переменных $X_{i'} \in \{X_1, \ldots, X_n\}$ на выборке $\tilde{S}_t$ с какой-либо линейной комбинацией других переменных значение $\det(\mathbf{X}^t\mathbf{X})$ оказывается близким к 0. При этом вычисленный вектор оценок $\hat{\boldsymbol{\beta}}^t$ может сильно изменяться при относительно небольших чисто случайных изменениях вектора $\mathbf{y} = (y_1, \ldots, y_m)$. Таким образом оценивание с использованием МНК при наличии мультиколлинеарности оказывается неустойчивым. Отметим также, что $\det(\mathbf{X}^t\mathbf{X}) = 0$ при $n+1 \geq m$. Поэтому МНК не может использоваться для оценивания регрессионных коэффициентов, когда число переменных превышает число объектов в обучающей выборке. На практике высокая устойчивость достигается только, когда число объектов в выборках по крайней мере в 3-5 раз превышает число переменных. Для подробного изучения методов многомерной линейно регрессии может быть рекомендована, например, книга [27]

2.4. Методы, основанные на регуляризации по Тихонову

Одним из возможных способов борьбы с неустойчивостью является использование методов, основанных на включение в исходный оптимизируемый функционал $Q(\tilde{S}_t, \beta_0, \ldots, \beta_n)$ дополнительной штрафной компоненты. Введение такой компоненты позволяет получить решение, на котором $Q(\tilde{S}_t, \beta_0, \ldots, \beta_n)$ достаточно близок к своему глобальному минимуму. Однако данное решение оказывается значительно более устойчивым и благодаря устойчивости позволяет достигать существенно более высокой обобщающей способности. Подход к получению более эффективных решений с помощью

включения штрафного слагаемого в оптимизируемый функционал принято называть регуляризацией по Тихонову.

На первом этапе переходим от исходных переменных $X_1,\ldots,X_n$ к стандартизированным $X_n^s,\ldots,X_n^s$, где $X_i^s=\frac{X_i-\hat{X}_i}{\hat{\sigma}_i}$, $\hat{X}_i=\frac{1}{m}\sum_{j=1}^{m}x_{ji}$ $,\hat{\sigma}_i=\sqrt{\frac{1}{m}\sum_{j=1}^{m}(x_{ji}-\hat{X}_i)^2}$ а также от исходной прогнозируемой переменной Y к стандартизованной прогнозируемой переменной $Y_s=Y-\frac{1}{m}\sum_{j=1}^{m}y_j$. Пусть $\hat{x}_{j1}^s=1,\ \hat{x}_{ji}^s=x_{j(i-1)}^s$ при $i>1$, где $x_{j(i-1)}^s$ - значение признака X_i^s для j-го объекта. Пусть также $\mathbf{X}_s=\begin{pmatrix}\hat{x}_{11}^s.\ldots\ldots\hat{x}_{1n}^s\\ \ldots\ldots\ldots\ldots\\ \hat{x}_{j1}\ldots\ldots\hat{x}_{jn}\\ \ldots\ldots\ldots\ldots\\ \hat{x}_{m1}\ldots\ldots\hat{x}_{mn}\end{pmatrix}$ - матрица плана для стандартизированных переменных, $\mathbf{y}_s=(y_1^s,\ldots,y_m^s)$- вектор значений стандартизованной переменной Y_s.

Одним из первых методов регрессии, использующих принцип регуляризации, является метод гребневой регрессии (ridge regression). В гребневой регрессии в оптимизируемый функционал дополнительно включается сумма квадратов регрессионных коэффициентов при переменных $X_1^s,\ldots,X_n^s$. В результате функционал имеет вид

$$Q_{ridge}(\tilde{S}_t,\beta_0,\ldots,\beta_n)=\frac{1}{m}\sum_{j=1}^{m}[y_j^s-\sum_{i=1}^{n+1}\beta_i\hat{x}_{ji}^s]^2+\gamma\sum_{i=0}^{n}\beta_i^2,$$

где γ - положительный вещественный параметр, $X_1^s,\ldots,X_n^s$ для j-го объекта, Пусть $\hat{\boldsymbol{\beta}}_r$ является вектором оценок регрессионных коэффициентов, полученным в результате минимизации $Q_{ridge}(\tilde{S}_t,\beta_0,\ldots,\beta_n)$. Отметим, что увеличение регрессионных коэффициентов приводит к увеличению $Q_{ridge}(\tilde{S}_t,\beta_0,\ldots,\beta_n)$. Таким образом

использование гребневой регрессии приводит к снижению длины вектора регрессионных коэффициентов при переменных $X_n^s,\ldots,X_n^s$.

Рассмотрим конкретный вид вектора регрессионных коэффициентов $\hat{\boldsymbol{\beta}}^r$. Необходимым условием минимума функционала $Q_{ridge}(\tilde{S}_t,\beta_0,\ldots,\beta_n)$ является выполнение системы из $n+1$ уравнений

$$\frac{\partial Q(\tilde{S}_t,\beta_0,\ldots,\beta_n)}{\partial \beta_0}=2[\sum_{j=1}^{m}y_j\hat{x}_{j1}^s-\sum_{j=1}^{m}\sum_{i=1}^{n+1}\beta_i\hat{x}_{ji}^s\hat{x}_{j1}^s+\gamma\beta_0]=0 \quad (5)$$

……

$$\frac{\partial Q(\tilde{S}_t,\beta_0,\ldots,\beta_n)}{\partial \beta_n}=2[\sum_{j=1}^{m}y_j\hat{x}_{jn}^s-\sum_{j=1}^{m}\sum_{i=1}^{n+1}\beta_i\hat{x}_{ji}^s\hat{x}_{jn}^s+\gamma\beta_n]=0$$

Поэтому вектор оценок регрессионных коэффициентов в методе гребневая регрессия является решением системы (5).

В матричной форме система (5) может быть записана в виде $-\mathbf{X}_s^t\mathbf{y}_s^t+(\mathbf{X}_s^t\mathbf{X}_s+\gamma I]\hat{\boldsymbol{\beta}}^t=0$ или в виде $\hat{\boldsymbol{\beta}}^t=\mathbf{X}_s^t\mathbf{y}_s^t[\mathbf{X}_s^t\mathbf{X}_s+\gamma\mathbf{I}]^{-1}$, где $\mathbf{I}$ – единичная матрица.

Отметим, что произведение $\mathbf{X}_s^t\mathbf{X}_s$ представляет собой симметрическую неотрицательно определённую матрицу. Матрица $\mathbf{X}_s^t\mathbf{X}_s+\gamma\mathbf{I}$ также является симметрической матрицей. Каждому собственному значению λ_k матрицы $\mathbf{X}_s^t\mathbf{X}_s$ соответствует собственное значение $\lambda_i+\gamma$ матрицы $\mathbf{X}_s^t\mathbf{X}_s+\gamma\mathbf{I}$. Таким образом минимальное собственное значение матрицы $\mathbf{X}_s^t\mathbf{X}_s+\gamma\mathbf{I}$ удовлетворяет неравенству $\lambda_{\min}\geq\gamma$. Откуда следует, что всегда $\det(\mathbf{X}_s^t\mathbf{X}_s+\gamma I)>0$, а обратная матрица $[(\mathbf{X}^s)^t\mathbf{X}+\gamma I]^{-1}$ всегда существует. Большая величина $\det(\mathbf{X}_s^t\mathbf{X}_s+\gamma I)>0$ приводит к относительно небольшим изменениям оценок регрессионных коэффициентов при небольших изменениях в обучающих выборках.

Наряду с гребневой регрессией в последние годы получил распространение метод Лассо, основанный на минимизации функционала

$Q_{Lasso}(\tilde{S}_t, \beta_0, \ldots, \beta_n) = \frac{1}{m}\sum_{j=1}^{m} [y_j^s - \sum_{i=1}^{n+1} \beta_i \hat{x}_{ji}^s]^2 + \gamma \sum_{i=0}^{n} |\beta_i|$. Интересной особенностью метода Лассо является равенство 0 части из регрессионных коэффициентов $(\beta_1, \ldots, \beta_n)$. Однако равенство 0 коэффициента на самом деле означает исключение из модели соответствующей ему переменной. Поэтому метод Лассо не только строит оптимальную регрессионную модель, но и производит отбор переменных. Метод может быть использован для отбора переменных в условиях, когда размерность данных превышает размер выборки. Отметим, что общее число отобранных переменных не может превышать размера обучающей выборки m . Эксперименты показали, что эффективность отбора переменных методом Лассо снижается, при высокой взаимной корреляции некоторых из них.

Данными недостатками не обладает другой метод построения регрессионной модели, основанный на регуляризации по Тихонову, который называется эластичная сеть. Метод эластичная сеть основан на минимизации функционала

$Q_r(\tilde{S}_t, \beta_0, \ldots, \beta_n) = \frac{1}{m}\sum_{j=1}^{m} [y_j^s - \sum_{i=1}^{n+1} \beta_i \hat{x}_{ji}^s]^2 + \gamma \sum_{i=0}^{n} [\alpha |\beta_i| + (1-\alpha)\frac{1}{2}\beta_i^2]$, где $\alpha \in [0,1]$.

Метод эластичная сеть включает в себя метод гребневая регрессия и Лассо как частные случаи.

Методы регрессионного анализа подробно рассматриваются в большом числе публикации. Например можно привести учебное пособие [4]. Методы регрессионного анализа, основанные на регуляризации по Тихонову рассматриваются в курсе лекций [3] и книге [16]

3. Методы распознавания

3.1 Методы оценки эффективности алгоритмов распознавания

Каждый алгоритм распознавания классов $K_1, \ldots, K_L$ независимо от задачи или используемой модели может быть представлен как последовательное выполнение распознающего оператора R и решающего правила C : $A = R \otimes C$. Оператор оценок вычисляет для распознаваемого объекта s вещественные оценки $\gamma_1(s), \ldots, \gamma_L(s)$ за классы $K_1, \ldots, K_L$ соответственно. Решающее правило C производит отнесение объекта s по вектору оценок $[\gamma_1(s), \ldots, \gamma_L(s)]$ к одному из классов. Распространённым решающим правилом является простая процедура, относящая объект в тот класс, оценка за который максимальна. В случае распознавания двух классов K_1 и K_2 распознаваемый объект s будет отнесён к классу K_1 , если $\gamma_1(s) - \gamma_2(s) \geq 0$ и классу K_2 в противном случае.

Назовём приведённое выше правило правилом $C(0)$. Однако точность распознавания правила $C(0)$ может оказаться слишком низкой для того, чтобы обеспечить требуемую величину потерь, связанных с неправильной классификацией объектов, на самом деле принадлежащих классу K_1. Для достижения необходимой величины потерь может быть использовано пороговое решающее правило $C(\delta)$: распознаваемый объект s будет отнесён к классу K_1, если $\gamma_1(s) - \gamma_2(s) \geq \delta$ и классу K_2 в противном случае.

Обозначим через $p_{ci}(\delta, s)$ вероятность правильной классификации правилом объекта s , на самом деле принадлежащего K_i, $i \in \{1,2\}$. При $\delta < 0$

$p_{c1}(\delta, s) \geq p_{c1}(0, s)$, но $p_{c2}(\delta, s) \leq p_{c2}(0, s)$. Уменьшая δ , мы увеличиваем $p_{c1}(\delta, s)$ и уменьшаем $p_{c2}(\delta, s)$. Напротив, увеличивая δ , мы уменьшаем $p_{c1}(\delta, s)$ и увеличиваем $p_{c2}(\delta, s)$. Зависимость между $p_{c1}(\delta, s)$ и $p_{c2}(\delta, s)$ может быть приближённо восстановлена по обучающей выборке $\tilde{S}_t$, включающей описания объектов $\{s_1, \ldots, s_m\}$

Пусть $\begin{pmatrix} \gamma_1(s_1) \ldots \gamma_1(s_m) \\ \gamma_2(s_1) \ldots \gamma_2(s_m) \end{pmatrix}$ - матрица оценок за классы объектов $\{s_1, \ldots, s_m\}$. По данной матрице оценок легко получить множество величин $\{\gamma(s_i) = \gamma_1(s_i) - \gamma_2(s_i) \mid i = 1, \ldots, m\}$, , где $i = 1, \ldots, m$.

Предположим, что величины $\gamma(s_i)$ принимают ***r*** различных значений $\Gamma_1, \ldots, \Gamma_r$, Данным величинам можно сопоставить решающие правила $C(\Gamma_1), \ldots, C(\Gamma_r)$. Для каждого из правил $C(\Gamma_i)$ вычислим две величины:

a) долю K_1 среди объектов обучающей выборки, удовлетворяющих условию $\gamma(s) \geq \Gamma_i$, которую обозначим $\nu_{c1}(\Gamma_i)$;
b) долю K_2 среди объектов обучающей выборки, удовлетворяющих условию $\gamma(s_*) < \Gamma_i$, которую обозначим $\nu_{c2}(\Gamma_i)$.

В результате мы получим ***r*** пар чисел

$$\{[\nu_{c1}(\Gamma_1), \nu_{c2}(\Gamma_1)], \ldots, [\nu_{c1}(\Gamma_r), \nu_{c2}(\Gamma_r)]\}.$$

Каждая пара чисел может рассматриваться как точка на плоскости в декартовой системе координат. Таким образом, набору пороговых элементов $\Gamma_1, \ldots, \Gamma_r$ соответствует набор точек на плоскости.

Соединив соседние по номеру точки отрезками прямых, получим ломаную линию, соединяющую точки (1,0) и (0,1), которая изображена на рисунке 3.1. Данная линия графически отображает аппроксимацию по обучающей выборке взаимозависимости между $p_{c1}(\delta, s)$ и $p_{c2}(\delta, s)$ при всевозможных значениях δ . Соответствующий пример представлен на рисунке 2. Взаимозависимость между ν_{c1} и

ν_{c2} наиболее полно оценивает эффективность распознающего оператора $\boldsymbol{R}$. Отметим, что ν_{c1} постепенно убывает по мере роста ν_{c2}.

.

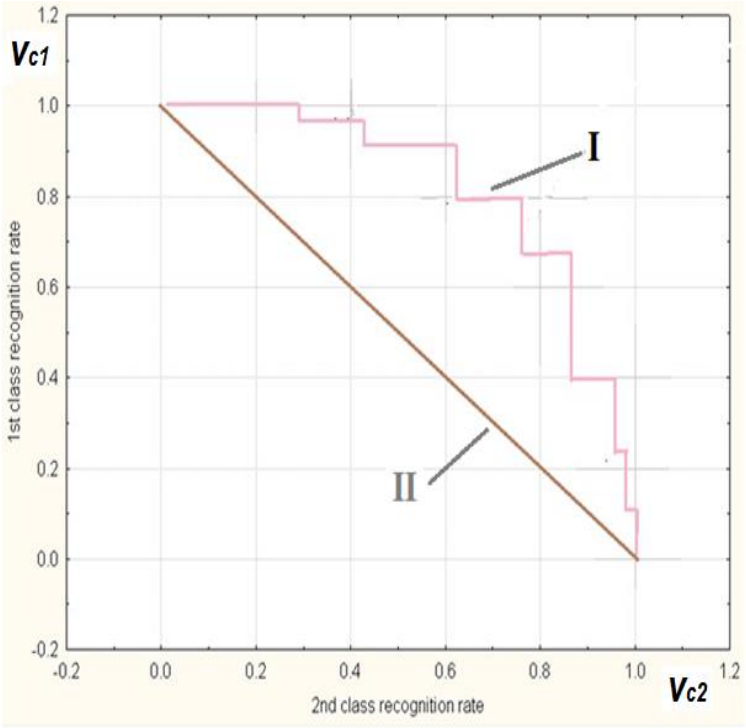


Рис 3.1. Ломаная (I) соединяет точки на двумерной плоскости в декартовой системе координат, которые являются соседними в ряду (1.1) .

Однако сохранение высокого значения ν_{c1} при высоких значениях ν_{c2} соответствует существованию решающего правила, при котором точность распознавания обоих классов высока. Наиболее эффективному распознающему оператору, обеспечивающему полное распознавание классов соответствует совпадение линии **I** с прямой, связывающей точки (0,1) и (1,1). Отсутствию распознающей способности соответствует совпадение с прямой **II**, связывающей точки (0, 1) и (1,0). В целом эффективность распознающего оператора может характеризоваться формой линии I. Чем ближе линия I к прямой, связывающей точки (0,1) и (1,1), тем лучше распознающий оператор и соответствующий ему метод распознавания. Наоборот, приближенность линии I к прямой, связывающей точки (0,1) и (1,1), соответствует низкой эффективности соответствующего метода распознавания.

На рисунке 3 сравниваются линии, характеризующие эффективность распознающих операторов, принадлежащих к трём методам распознавания, при решении задач

диагностики двух видов аутизма по психометрическим показателям. Изучалась эффективность

-линейного дискриминанта Фишера (ЛДФ) с соответствующей линией обозначенной O;

- метода опорных векторов (МОВ) с линией, обозначенной O ;

-метода статистически взвешенные синдромов (СВС) с линией, обозначенной O.

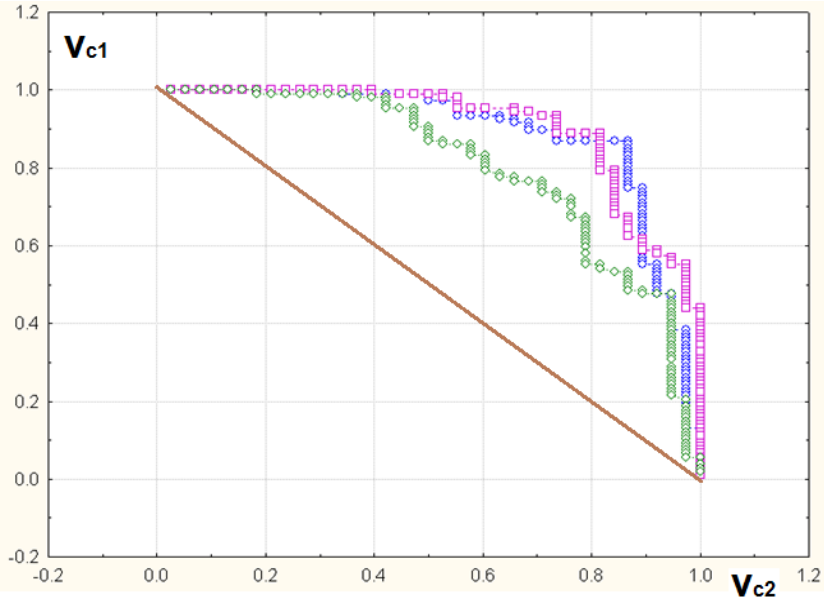


Рис. 3.2 Сравнение трёх метод распознавания с помощью

Методы распознавания используются при решении многих задач идентификации объектов, представляющих важность для пользователя. Эффективность идентификации для таких задач удобно описывать в терминах:

«Чувствительность» - доля правильно распознанных объектов целевого класса

«Ложная тревога» - доля объектов ошибочно отнесённых в целевой класс.

Пример кривой, связывающей параметры «Чувствительность» и «Ложная тревога» представлен на рисунке 4.

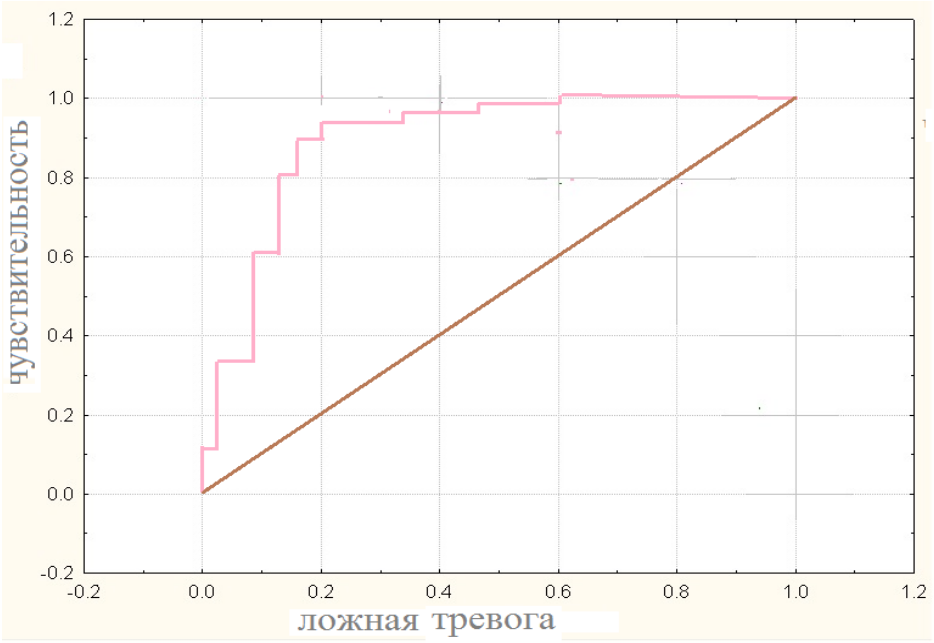


Рис. 3.3 Вид ROC кривой в координатах чувствительность (ось Y) и ложная тревога (ось X)

Анализ, основанный на построении и анализе линий, связывающих параметры «Чувствительность» и «Ложная тревога» принято называть анализом Receiver Operating Characteristic или ROC-анализом.

Отметим, что по мере увеличения числа пороговых точек , что обычно происходит при возрастании объёма выборки, ломаная линия I постепенно приближается к некоторой кривой. Поэтому линию Линии, связывающих параметры «Чувствительность» и «Ложная тревога» принято называть ROC-кривыми. В качестве меры близости к прямой, связывающей точки (0,0) и (1,1), соответствующей абсолютно точному распознаванию, используется площадь под ROC – кривой.

Задача к разделу «Методы оценки эффективности алгоритмов распознавания»

Банк использует 2 метода распознавания для повышения прибыли при кредитовании. Используемая технология основана на распознавании в заёмщиков, для которых риск отказа от выплат по кредиту является высоким. Предполагается, что доход банка с одного добросовестного заёмщика составляет $d = 10000$ условных единиц (у.е.). Потери банка при отказе от выплат по кредиту составляет $L = 45000$ у.е. Доля заёмщиков, отказывающихся от выплат по кредиту составляет $p_{rej} = 0.05$. В таблице приведены значения чувствительности и ложной тревоги при некотором наборе пороговых значений для методов распознавания A и B.

Таблица 1

Метод A			Метод B	
Чувствительность	Ложная тревога		Чувствительность	Ложная тревога
0.03	0.001		0.03	0.001
0.08	0.002		0.16	0.002
0.13	0.01		0.28	0.02
0.19	0.03		0.44	0.06
0.27	0.07		0.57	0.08
0.34	0.09		0.61	0.09
0.47	0.11		0.67	0.11
0.61	0.14		0.69	0.14
0.74	0.17		0.72	0.17
0.91	0.21		0.78	0.2
0.97	0.24		0.83	0.23
1	0.28		0.88	0.27
			0.92	0.32
			0.98	0.35
			1	0.37

Вопросы. Позволяют ли приведённые в таблице 1 данные сделать вывод о потенциальной возможности увеличении дохода банка при использовании метода A или метода B? Какой из двух методов позволяет получить более высокий доход?

Решение. Средний доход банка на одну поданную заявку на кредит в D случае, когда методы распознавания не используются очевидно может быть найден по формуле

$$D = d * (1 - p_{rej}) - p_{rej} * L = 10000 * 0.95 - 45000 * 0.05 = 7250,$$

При использовании метода распознавания с чувствительностью Sen и уровнем ложной тревоги Fa. Величина потерь, произошедших непосредственно из-за отказов от выплат по кредиту, которая без применения методов распознавания была равна $p_{rej} * L$, становится равной $p_{rej} * L * (1 - Sen)$. Величина дохода, полученная на добросовестных заёмщиков, которая без применения методов распознавания была равна $d * (1 - p_{rej})$, в случае применения метода распознавания оказывается равной $d * (1 - p_{rej}) * (1 - Fa)$. Таким образом величина дохода в случае использование метода распознавания рассчитывается по формуле

$$D = d * (1 - p_{rej}) * (1 - Fa) - p_{rej} * L * (1 - Sen)$$

3.2 Байесовские методы

Ранее было показано, что максимальную точность распознавания классов $K_1,\ldots,K_L$ обеспечивает байесовское решающее правило, относящее распознаваемый объект, описываемый вектором переменных (признаков) $X_1,\ldots,X_n$ к классу K_{i^b}, для которого условная вероятность $\mathbf{P}(K_{i^b}\,|\,\mathbf{x})$ максимальна.

Байесовские методы обучения основаны на аппроксимации условных вероятностей классов в точках признакового пространства с использованием формулы Байеса. Формула Байеса позволяет рассчитать условные вероятности классов в точке признакового пространства:

$$\mathbf{P}(K_i\,|\,\mathbf{x})=\frac{p_i(\mathbf{x})\mathbf{P}(K_i)}{\sum_{j=1}^{L}p_i(\mathbf{x})\mathbf{P}(K_i)},$$

где $p_i(\mathbf{x})$ - плотность распределения вероятности для класса K_i; $\mathbf{P}(K_i)$ - вероятность класса K_i безотносительно к признаковым описаниям (априорная вероятность).

При этом в качестве оценок априорных вероятностей $\mathbf{P}(K_i)$ могут быть взята доля объектов класса K_i в обучающей выборке, которая далее будет обозначаться ν_i. Плотности вероятностей $p_1(\mathbf{x}),\ldots,p_L(\mathbf{x})$ восстанавливаются исходя из предположения об их принадлежности фиксированному типу распределения. Чаще всего используется многомерное нормальное распределения. Плотность данного распределения в общем виде представляется выражением

$$p(\mathbf{x})=\frac{1}{(2\pi)^{n/2}\,|\,\Sigma\,|^{1/2}}\exp[-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})^t]$$

где $\boldsymbol{\mu}$ - математическое ожидание вектора признаков $X_1,\ldots,X_n$

$\boldsymbol{\Sigma}$ - матрица ковариаций признаков $X_1,\ldots,X_n$; $|\boldsymbol{\Sigma}|$ - детерминант матрицы $\boldsymbol{\Sigma}$.

Для построения распознающего алгоритма достаточно оценить вектора атематических ожиданий $\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_L$ и матрицы ковариаций $\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_L$ для классов $K_1,\ldots,K_L$ соответственно. Оценка $\boldsymbol{\mu}_i$ вычисляется как среднее значение векторов признаков по объектам обучающей выборки из класса K_i :

$$\hat{\boldsymbol{\mu}}_i = \tfrac{1}{m_i} \sum_{s_j \in \tilde{S}_t \cap K_i} \mathbf{x}_j \ ,$$

где m_i - число объектов класса K_i в обучающей выборке.

Оценка элемента матрицы ковариаций для класса K_i вычисляется по формуле

$$\hat{\sigma}^i_{kk'} = \tfrac{1}{m_i} \sum_{s_j \in \tilde{S}_t \cap K_i} (x_{jk} - \mu^i_k)(x_{jk'} - \mu^i_{k'}), \quad k,k' \in \{1,\ldots,n\} \ ,$$

где μ^i_k - k-я компонента вектора $\boldsymbol{\mu}^i$. Матрицу ковариации, состоящую из элементов $\sigma^i_{kk'}$ обозначим $\hat{\Sigma}_i$. Очевидно, что согласно формуле Байеса максимум $\mathbf{P}(K_i \mid \mathbf{x})$ достигается для тех же самых классов для которых максимально произведение $p_i(\mathbf{x})\mathbf{P}(K_i)$. На практике для классификации удобнее использовать натуральный логарифм $\ln[p_i(\mathbf{x})\mathbf{P}(K_i)]$, который согласно вышеизложенному может быть оценён выражением $g_i(\mathbf{x}) = -\tfrac{1}{2}(\mathbf{x}\hat{\boldsymbol{\Sigma}}_i\mathbf{x}^t) + \mathbf{w}_i\mathbf{x}^t + g$, где $\mathbf{w}_i = \hat{\boldsymbol{\mu}}_i\hat{\boldsymbol{\Sigma}}_i^{-1}$,

$g_i^0 = -\tfrac{1}{2}(\hat{\boldsymbol{\mu}}_i\hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{\mu}}_i^t) - \tfrac{1}{2}\ln(|\hat{\boldsymbol{\Sigma}}_i|) + \ln(\nu_i) - \tfrac{n}{2}\ln(2\pi)$ - не зависящее от $\mathbf{x}$ слагаемое;

Таким образом объект с признаковым описанием будет отнесён построенной выше аппроксимацией байесовского классификатора к классу, для которого оценка является максимальной. Следует отметить, что построенный классификатор в общем случае является квадратичным по признакам. Однако классификатор превращается в линейный, если оценки ковариационных матриц разных классов оказываются равными.

Задача к разделу Байесовские методы

Пусть априорные вероятности классов K_1 и K_2 равны 0.3 и 0.7 соответственно. Предположим, что значения некоторого признака X для обоих классов распределены нормально. Для класса K_1 $\mu_1 = 2, \sigma = 1.$ Для класса K_2 $\mu_2 = -2, \sigma = 1.5.$ Выделить на числовой оси области значений признака X, при которых байесовский классификатор относит классифицируемые объекты классу K_1.

Решение. Как было показано байесовский классификатор относит объект, для которого $X = x_*$, классу K_1. при выполнении неравенства

$$\ln[P(K_1)\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x_*-\mu_1)^2}{2\sigma_1}}] > \ln[P(K_2)\frac{1}{\sqrt{2\pi}\sigma_2}e^{-\frac{(x_*-\mu_2)^2}{2\sigma_2}}].$$

Откуда следует, что

$$\ln[\frac{P(K_1)}{P(K_2)}\frac{\sigma_2}{\sigma_1}] - \frac{(x_*-\mu_1)^2}{2\sigma_1} + \frac{(x_*-\mu_2)^2}{2\sigma_2} > 0. \qquad (1)$$

.

Введём дополнительные обозначения

$$\alpha = \frac{\sigma_1 - \sigma_2}{2\sigma_1\sigma_2},\ \beta = \frac{\mu_1\sigma_2 - \mu_2\sigma_1}{\sigma_1 - \sigma_2},\ \gamma = \frac{\sigma_1\mu_2^2 - \sigma_2\mu_1^2}{\sigma_1 - \sigma_2}.$$

Нетрудно показать, что неравенство (1) эквивалентно неравенству

$$\alpha(x^* + \beta)^2 > \ln[\frac{P(K_2)}{P(K_1)}\frac{\sigma_1}{\sigma_2}] + \alpha(\beta^2 - \gamma), \qquad (2)$$

Введём обозначение $\theta = \frac{1}{\alpha}\ln[\frac{P(K_2)}{P(K_1)}\frac{\sigma_1}{\sigma_2}] + (\beta^2 - \gamma)$. Неравенство (2) эквивалентно неравенству $(x^* + \beta)^2 > \theta,$ при $\sigma_1 > \sigma_2$ или неравенству $(x^* + \beta)^2 < \theta$ при $\sigma_2 > \sigma_1$.

Неравенство $(x^* + \beta)^2 > \theta$ выполняется всегда при $\theta < 0$. При $\theta > 0$ неравенство $(x^* + \beta)^2 > \theta$ эквивалентно одновременному выполнению неравенств $x^* > \theta - \beta, x^* < -\theta - \beta.$

Неравенство $(x^* + \beta)^2 < \theta$ не выполняется при $\theta < 0$. При $\theta > 0$ неравенство $(x^* + \beta)^2 < \theta$ эквивалентно одновременному выполнению неравенств $x^* < \theta - \beta, x^* > -\theta - \beta.$

3.2.2 Линейный дискриминант Фишера

Рассмотрим вариант метода Линейный дискриминант Фишера (ЛДФ) для распознавания двух классов K_1 и K_2. В основе метода лежит поиск в многомерном признаковом пространстве такого направления $\mathbf{w}$, чтобы средние значения проекции на него объектов обучающей выборки из классов K_1 и K_2 максимально различались. Проекцией произвольного вектора $\mathbf{x}$ на направление $\mathbf{w}$ является отношение $\frac{(\mathbf{w},\mathbf{x}^t)}{|\mathbf{w}|}$. В качестве меры различий проекций классов на $\mathbf{w}$ используется функционал

$$\Phi(\mathbf{w}) = \frac{[\hat{\mathbf{X}}_{w1}(\mathbf{w}) - \hat{\mathbf{X}}_{w2}(\mathbf{w})]^2}{\hat{d}_1(\mathbf{w}) + \hat{d}_2(\mathbf{w})},$$

где $\hat{X}_{wi}(\mathbf{w}) = \frac{1}{m_i} \sum_{s_j \in \tilde{S}_t \cap K_i} \frac{(\mathbf{w}\mathbf{x}_j^t)}{|\mathbf{w}|}$ - среднее значение проекции векторов, описывающих объекты из класса K_i;

$$\hat{d}_{wi}(\mathbf{w}) = \frac{1}{m_i} \sum_{s_j \in \tilde{S}_t \cap K_i} [\frac{(\mathbf{w}\mathbf{x}_j^t)}{|\mathbf{w}|} - \hat{X}_{wi}(\mathbf{w})]^2$$

- выборочная дисперсия проекций векторов, описывающих объекты из класса $K_i, \quad i \in \{1,2\}$.

Смысл функционала $\Phi(\mathbf{w})$ ясен из его структуры. Он является по сути квадратом отличия между средними значениями проекций классов на направление $\mathbf{w}$, нормированным на сумму внутриклассовых выборочных дисперсий

Можно показать, что $\Phi(\mathbf{w})$ достигает максимума при

$$\mathbf{w} = \hat{\Sigma}_{12}^{-1}(\boldsymbol{\mu}_1^t - \boldsymbol{\mu}_2^t), \qquad (1)$$

где $\hat{\Sigma}_{12} = \hat{\Sigma}_1 + \hat{\Sigma}_2$. Таким образом оценка направления, оптимального для распознавания K_1 и K_2 может быть записана в виде (1).

Распознавание нового объекта s_* по признаковому описанию $\mathbf{x}_*$ производится по величине проекции $\gamma(\mathbf{x}_*) = \frac{(\mathbf{w}, \mathbf{x}_*^t)}{|\mathbf{w}|}$ с помощью простого порогового правила: при $\gamma(\mathbf{x}_*) > \delta$ объект s_* относится к классу K_1 и s_* относится к классу K_2 в противном случае.

Граничный параметр δ подбирается по обучающей выборке таким образом, чтобы проекции объектов разных классов на оптимальное направление $\mathbf{w}$ оказались бы максимально разделёнными. Простой, но эффективной, стратегией является выбор в качестве порогового параметра δ средней проекции объектов обучающей выборки на направление $\mathbf{w}$. Метод ЛДФ легко обобщается на случай с несколькими классами.

При этом исходная задача распознавания классов $K_1, \ldots, K_L$ сводится к последовательности задач с двумя классами и :

Зад. 1. Класс $K_1' = K_1$, класс $K_2' = \Omega \setminus K_1$

……

Зад. L. Класс $K_1' = K_L$, класс $K_2' = \Omega \setminus K_L$

Для каждой из L задач ищется оптимальное направление и пороговое правило.В результате получается набор из L направлений $\mathbf{w}_1, \ldots, \mathbf{w}_L$. При распознавании нового объекта по признаковому описанию вычисляются проекции на $\mathbf{w}_1, \ldots, \mathbf{w}_L$

$$\gamma_1(\mathbf{x}_*) = \frac{(\mathbf{w}_1 \mathbf{x}_*^t)}{|\mathbf{w}_1|}, \ldots, \gamma_L(\mathbf{x}_*) = \frac{(\mathbf{w}_L \mathbf{x}_*^t)}{|\mathbf{w}_L|}$$

Распознаваемый объект относится к тому классу, соответствующему максимальной величине проекции. Распознавание может производится также по величинам $[\gamma_1(\mathbf{x}_*) - b_1], \ldots, [\gamma_L(\mathbf{x}_*) - b_L]$.

3.2 3 Логистическая регрессия

Целью логистической регрессии является аппроксимация плотности условных вероятностей классов в точках признакового пространства. При этом аппроксимация производится с использованием логистической функции:

$$g(z) = \frac{e^z}{e^z + 1} = \frac{1}{e^{-z} + 1}.$$

График логистической функции приведён на рисунке

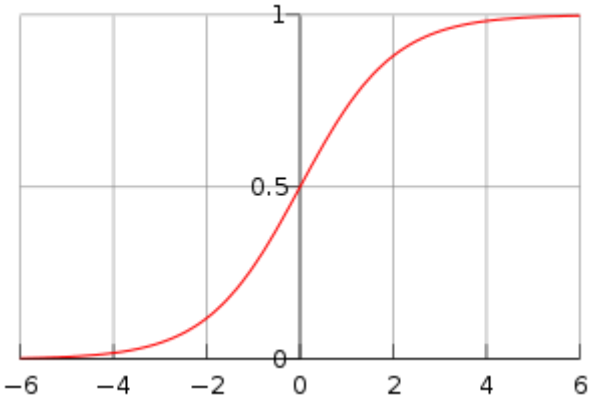


Рис.

В методе логистическая регрессия связь условной вероятности класса с прогностическими признаками осуществляются через переменную , которая задаётся как линейная комбинация признаков: $z = \beta_0 + \beta_1 X_1 + \ldots \beta_n X_n$

Таким образом условная вероятность K в точке векторного пространства $\mathbf{x}_* = (x_{*1}, \ldots, x_{*n})$ задаётся в виде

$$\mathbf{P}(K \mid \mathbf{x}) = \frac{1}{e^{-\beta_0 - \beta_1 x_{*1} - \ldots - \beta_1 x_{*n}} + 1} = \frac{e^{\beta_0 + \beta_1 x_{*1} + \ldots + \beta_1 x_{*n}}}{e^{\beta_0 + \beta_1 x_{*1} + \ldots + \beta_1 x_{*n}} + 1}$$

Оценки регрессионных параметров $\beta_0, \beta_1, \ldots, \beta_n$ могут быть вычислены по обучающей выборке с помощью различных вариантов метода максимального правдоподобия.

Метод k-ближайших соседей

Простым, но достаточно эффективным подходом к решению задач распознавания является метод k-ближайших соседей. Оценка условных вероятностей $\mathbf{P}(K_i \mid \mathbf{x})$ ведётся по ближайшей окрестности V_k точки $\mathbf{x}$, содержащей k признаковых описаний объектов обучающей выборки. В качестве оценки за класс K_i выступает отношение $\frac{k_i}{k}$, где k_i - число признаковых описаний объектов обучающей

выборки из K_i внутри V_k. Окрестность V_k задаётся с помощью функции расстояния $\rho(\mathbf{x}', \mathbf{x}'')$, заданной на декартовом произведении $\tilde{X} \times \tilde{X}$, где $\tilde{X}$ - область допустимых значений признаковых описаний. В качестве функции расстояния может быть использована стандартная эвклидова метрика $\rho(\mathbf{x}', \mathbf{x}'') = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x'_i - x''_i)^2}$.

Для задач с бинарными признаками в качестве функции расстояния может быть использована метрика Хэмминга, равная числу совпадающих позиций в двух сравниваемых признаковых описаниях.

Окрестность V_k ищется путём поиска в обучающей выборке $\tilde{S}_t$ векторных описаний, ближайших в смысле выбранной функции расстояний, к описанию распознаваемого объекта s_*. Единственным параметром, который может быть использован для настройки (обучения) алгоритмов в методе k–ближайших соседей является собственно само число ближайших соседей.

Для оптимизации параметра k обычно используется метод, основанный на скользящем контроле. Оценка точности распознавания производится по обучающей выборке при различных k и выбирается значение данного параметра, при котором полученная точность максимальна.

Разнообразные статистические методы распознавания рассмотрены в курсе лекций [3]. Следует отметить также книги [16],[17].

4 Модели распознавания, основанные на различных способах обучения

Статистические методы распознавания нередко обеспечивали достаточно высокую точность в прикладных исследованиях**.** Однако в различных областях науки и практической деятельности возникали задачи диагностики и прогнозирования, которые могли быть сведены к задачам распознавания. При этом исследователям удавалось собрать обучающую выборку весьма ограниченного объёма. а число показателей, которые потенциально могли быть использованы оказывалось достаточно большим. Для решения таких задач стали предлагаться новые подходы, не содержащие предположений о лежащих в основе изучаемого процесса вероятностных распределений. Оказалось, что такие подходы часто имеют более высокую эффективность, чем статистические методы.

4.1 Метод Линейная машина

. Метод «Линейная машина» предназначен для решения задачи распознавания с классами $K_1,\ldots,K_L$ . .

В процессе обучения классам $K_1,\ldots,K_L$ ставятся в соответствие линейные функции $f_1,\ldots,f_L$ от переменных $X_1,\ldots,X_n$, являющиеся оценками за классы $K_1,\ldots,K_L$. То есть для произвольного вектора значений переменных $\mathbf{x}=(x_1,\ldots,x_n)$

$$f_1(\mathbf{x}) = w_0^1 + w_1^1 x_1 + \ldots + w_n^1 x_n$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$f_L(\mathbf{x}) = w_0^L + w_1^L x_1 + \ldots + w_n^L x_n$$

Для того, чтобы распознать объект s, описание которого задаётся вектором $\mathbf{x}$. вычисляются значения функций $f_1,\ldots,f_L$ в точке $\mathbf{x}$. Объект s будет отнесён классу K_l, если выполняется набор неравенств: $f_l(\mathbf{x}) > f_j(\mathbf{x}), j \in \{1,\ldots,L\} \setminus \{l\}$

Таким образом алгоритм распознавания задаётся матрицей вещественных параметров

$$\mathbf{W} = \begin{pmatrix} w_0^1 w_1^1 \dots w_n^1 \\ \dots\dots\dots\dots \\ w_0^L w_1^L \dots w_n^L \end{pmatrix}.$$

Обучения ведётся по выборке $\tilde{S}_t = \{s_1 = (y_1, \mathbf{x}_1), \dots, s_m = (y_m, \mathbf{x}_m)\}$, где $\{y_1, \dots, y_m\}$ являются значениями дискретной прогнозируемой переменной, указывающей на номер класса, которому принадлежит соответствующий объект. Обучение состоит в поиске таких значений параметров из матрицы $\mathbf{W}$, при которых максимальное число объектов $\tilde{S}_t$ оказывается правильно распознанным. Обозначим через $r(i)$ номер класса, которому принадлежит объект s_i из обучающей выборки. Максимальная точность на $\tilde{S}_t$ соответствует выполнению максимального числа блоков неравенств:

$$f_{r(1)}(\mathbf{x}_1) > f_j(\mathbf{x}_1),\ j \in \{1, \dots, L\} \setminus \{r(1)\}$$

$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$

$$f_{r(m)}(\mathbf{x}_m) > f_j(\mathbf{x}_m),\ j \in \{1, \dots, L\} \setminus \{r(m)\}$$

Каждый из блоков соответствует одному из объектов выборки включает $l-1$ неравенство. Таким образом суммарное число неравенств составляет $(l-1)m$.

Поиск оптимальной матрицы коэффициентов $\mathbf{W}$ производится с помощью релаксационного алгоритма, подробно описанного в книге [10].

Приведём графический пример алгоритма распознавания, построенного с помощью метода линейная машина. Имеется задача распознавания с классами 1, 2, 3 по признакам X_1 и X_2

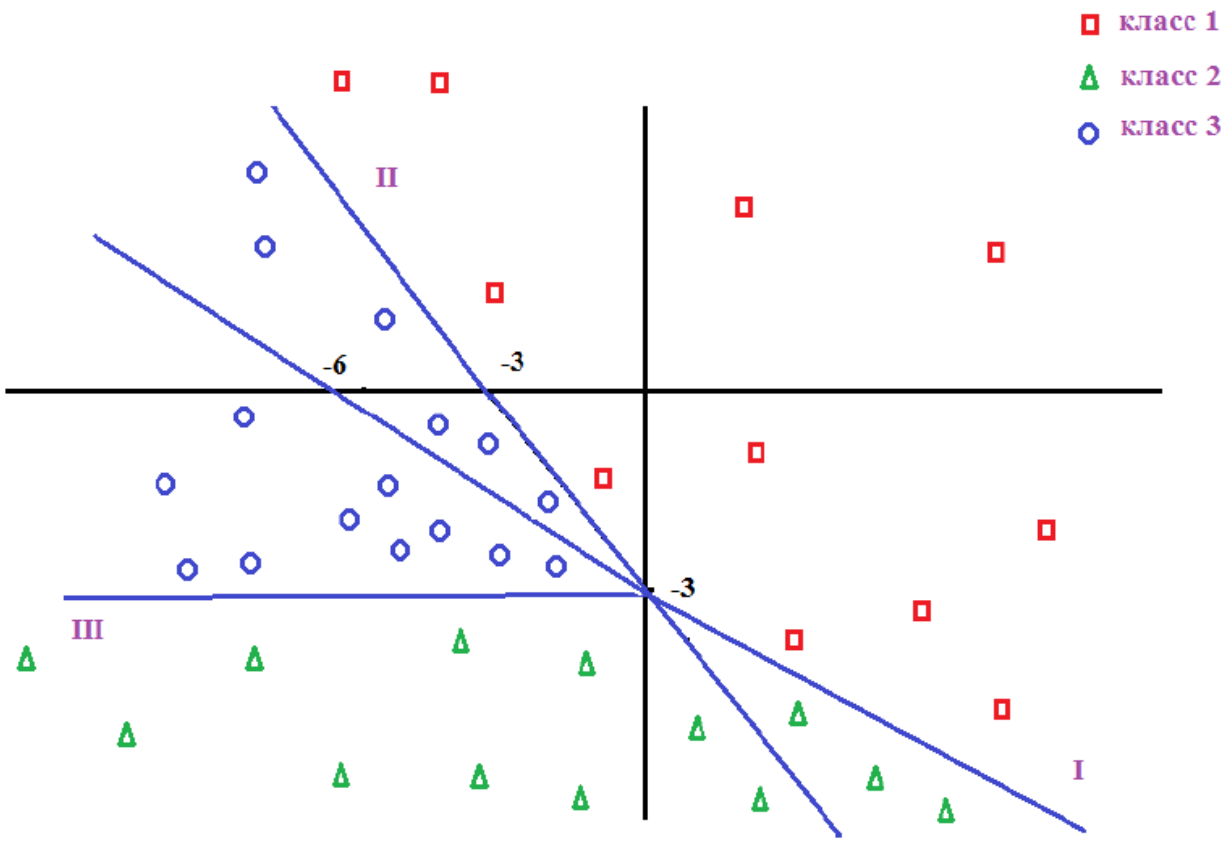


Рис. 1 Области, соответствующие отнесению распознаваемых объектов классам $K_1,\ldots,K_3$ методом линейная машина, вычисляющим оценки за классы по формулам (1).

Предполагается, что с использованием метода ЛМ для каждого класса найдены линейные функции оценок:

$$f_1(x_1,x_2)=4.0+2x_1-x_2 \quad \text{для класса 1;}$$

$$f_2(x_1,x_2)=-2.0+x_1-3x_2 \quad \text{для класса 2;} \qquad (1)$$

$$f_3(x_1,x_2)=1.0+x_1-2x_2 \quad \text{для класса 3.}$$

Изобразим на двумерной диаграмме области, соответствующие отнесению классов.

Очевидно, что классу 1 соответствует область, для которой выполняются неравенства $f_1(x_1,x_2)>f_2(x_1,x_2)$ и $f_1(x_1,x_2)>f_3(x_1,x_2)$. Неравенство $f_1(x_1,x_2)>f_2(x_1,x_2)$ эквивалентны неравенству $6+x_1+2x_2>0$, задающему границу I. Неравенство

$f_1(x_1, x_2) > f_3(x_1, x_2)$ эквивалентны неравенству $3 + x_1 + x_2 > 0$, задающему границу II. Область, соответствующая классу 1, помечена красными квадратами. Область, не относящаяся классу 1 при выполнении неравенства $f_2(x_1, x_2) > f_3(x_1, x_2)$.

Метод линейная машина подробно описан в книге [10].

4.2 Нейросетевые методы

4.2.1 Модель искусственного нейрона.

В основе нейросетевых методов лежит попытка компьютерного моделирования процессов обучения, используемых в живых организмах. Когнитивные способности живых существ связаны с функционированием сетей связанных между собой биологических нейронов – клеток нервной системы. Для моделирования биологических нейросетей используются сети, узлами которых являются искусственные нейроны (т.е. математические модели нейронов), Можно выделить три типа искусственных нейронов: нейроны-рецепторы, внутренние нейроны и реагирующие нейроны. Каждый внутренний или реагирующий нейрон имеет множество входных связей, по которым поступают сигналы от рецепторов или других внутренних нейронов. Пример модели внутреннего или реагирующего нейрона представлен на рисунке 1.

Представленный на рисунке 1 нейрон имеет r внешних связей, по которым на него поступают входные сигналы $u_1, \ldots, u_r$. Поступившие сигналы суммируются с весами $w_1, \ldots, w_r$. На выходе нейрона вырабатывается сигнал $\Phi(z)$, где $z = w_0 + \sum_{i=1}^{r} w_i u_i$, w_0 - параметр сдвига. Может быть использована также форма записи $z = \sum_{i=0}^{r} w_i u_i$, где фиктивный «сигнал» u_0 тождественно равен 1.

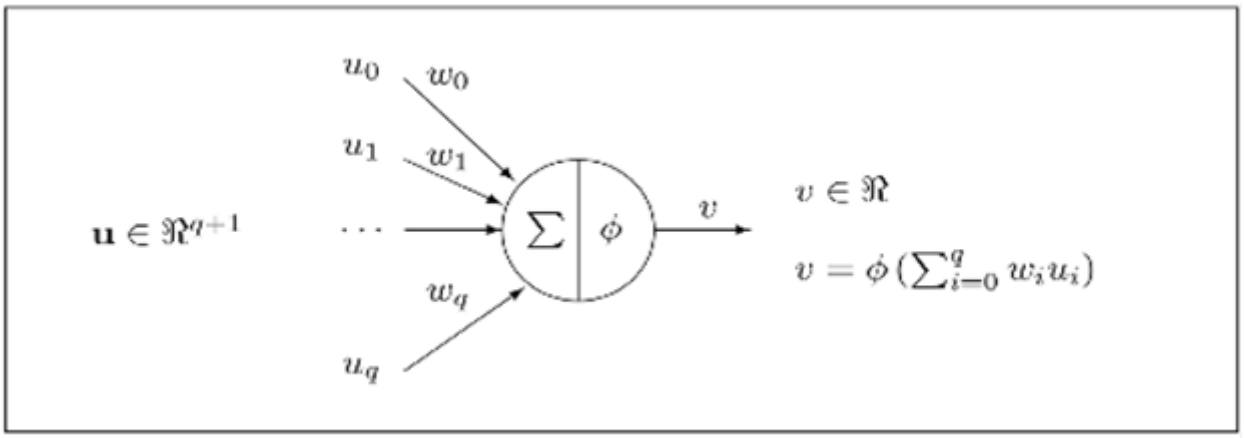


Рис.1. Модель внутреннего или реагирующего нейрона.

Функцию $\Phi(z)$ обычно называют активационной функцией. Могут использоваться различные виды активационных функций, включая

a) пороговую функцию, задаваемую с помощью пороговой величины b:

$$\Phi(z) = \begin{cases} 1 \text{ при } z > b \\ -1 \text{ при } z \le b \end{cases},$$

b) сигмоидная функция $\Phi(z) = \dfrac{1}{1+e^{-az}}$, где a - вещественная константа;

c) гиперболический тангенс;

d) тождественное преобразование $\Phi(z) = z$.

Первой нейросетевой моделью стал перцептрон Розенблатта, предложенный в 1957 году. В данной модели используется единственный реагирующий нейрон. Модель, реализующая линейную разделяющую функцию в пространстве входных сигналов, может быть использована для решении задач распознавания с двумя классами, помеченными метками 1 или -1. В качестве активационной функции используется пороговая функция:

$$\Phi(z) = \begin{cases} 1 \text{ при } z > 0 \\ -1 \text{ при } z \le 0 \end{cases}.$$

Особенностью модели Розенблатта является очень простая, но вместе с тем эффективная, процедура обучения, вычисляющая значения весовых коэффициентов

$w_0,\ldots,w_n$. Настройка параметров производится по обучающим выборкам, совершенно аналогичных тем, которые используются для обучения статистических алгоритмов.

На первом этапе производится преобразование векторов сигналов (признаковых описаний) для объектов обучающей выборки. В набор исходных признаков добавляется тождественно равная 1 нулевая компонента. Затем вектора описаний из класса K_2 умножаются на -1. Вектора описаний из класса K_1 не изменяются.

Нулевое приближение вектора весовых коэффициентов $w_0^0,\ldots,w_n^0$ выбирается случайным образом. Преобразованные описания объектов обучающей выборки $\tilde{S}_t$ последовательно подаются на вход перцептрона. В случае если описание $\mathbf{x}^{(k)}$, поданное на шаге k классифицируется неправильно, то происходит коррекция по правилу $\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \mathbf{x}^{(k)}$. В случае правильной классификации $\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)}$.

Отметим, что правильной классификации всегда соответствует выполнение равенства $(\mathbf{w}^{(k)}, \mathbf{x}^{(k)}) \geq 0$ а неправильной классификации соответствует выполнение равенства $(\mathbf{w}^{(k)}, \mathbf{x}^{(k)}) < 0$. Процедура повторяется до тех пор, пока не будет выполнено одно из следующих условий:

- достигается полное разделение объектов из классов K_1 и K_2;

- повторение подряд заранее заданного числа итераций не приводит к улучшению разделения;

- оказывается исчерпанным заранее заданный лимит итераций. Для описанной процедуры справедлива следующая теорема.

Теорема. В случае, если описания объектов обучающей выборки линейно разделимы в пространстве признаковых описаний, то процедура обучения перцептрона построит линейную гиперплоскость разделяющую объекты двух классов за конечное число шагов.

Отсутствие линейной разделимости двух классов приводит к бесконечному зацикливанию процедуры обучения перцептрона.

Существенно более высокой аппроксимирующей способностью обладают нейросетевые методы распознавания, задаваемые комбинациями является связанных между собой нейронов. Таким методом является многослойный перцептрон.

4.2.2 Многослойный перцептрон.

В методе многослойный перцептрон сеть формируется из нескольких слоёв нейронов.

В их число входит слой входных рецепторов, подающих сигналы на нейроны из внутренних слоёв. Слои внутренних нейронов осуществляют преобразование сигналов. Слой реагирующих нейронов производит окончательную классификацию объектов на основании сигналов, поступающих от нейронов, принадлежащих внутренним слоям. Обычно соблюдаются следующие правила формирования структуры сети.

Допускаются связи между только между нейронами, находящимися в соседних слоях.

Связи между нейронами внутри одного слоя отсутствуют.

Активационные функции для всех внутренних нейронов идентичны и задаются сигмоидными функциями.

Для решения задач распознавания с L классами $K_1,\ldots,K_L$ используется конфигурация с L реагирующими нейронами. Схема многослойного перцептрона с двумя внутренними слоями представлена на рисунке 3.

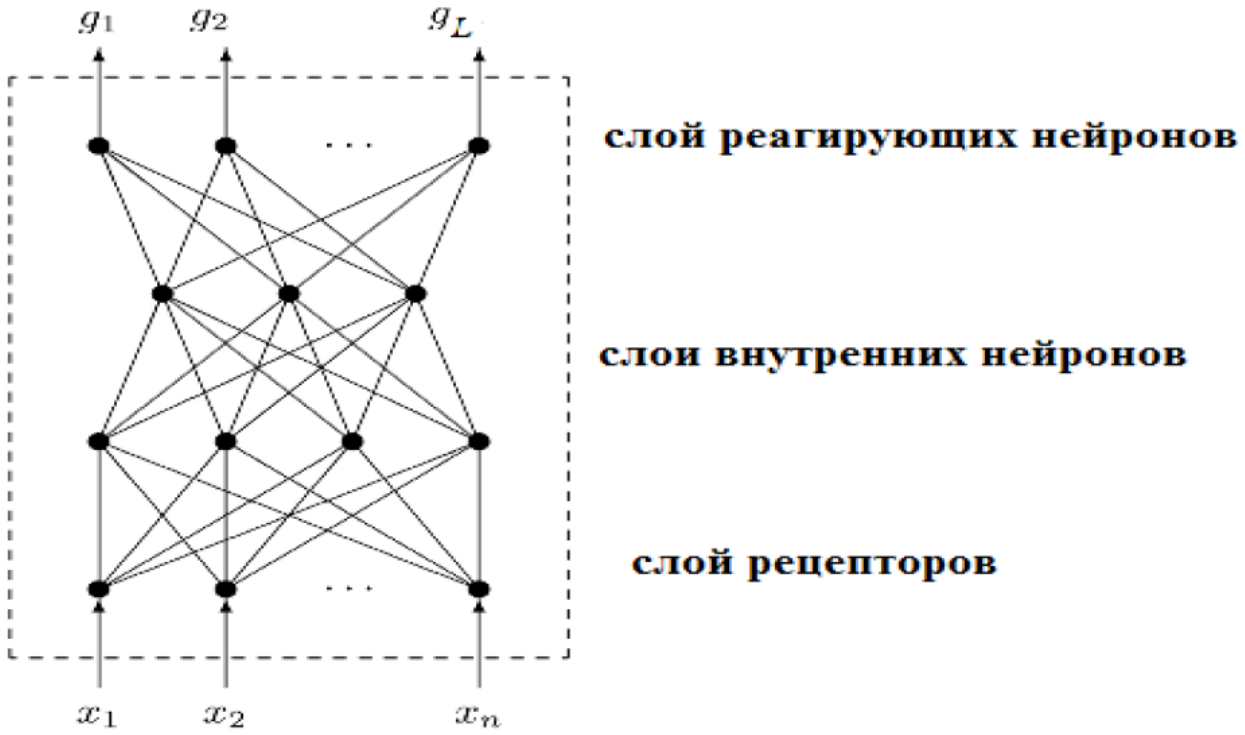


Рис. 3 Схема многослойного перцептрона с двумя внутренними слоями.

Отметим, что сигналы $g_1,\ldots,g_L$, вычисляемые на выходе реагирующих нейронов, интерпретируются как оценки за классы $K_1,\ldots,K_L$. Весовые коэффициенты w сопоставлены каждой из связей между нейронами из различных слоёв. Рассмотрим процедуру распознавания объектов с использованием многослойного перцептрона. Предположим, что конфигурация нейронной сети включает наряду со слоем рецепторов и слоем реагирующих нейронов также H внутренних слоёв искусственных нейронов. Заданы также количества нейронов в каждом слое. Пусть n – число входных нейронов-рецепторов, $r(h)$- число нейронов в внутреннем слое h.

На первом этапе вектор рецепторы формируют по информации, поступающей из внешней среды, вектор входных переменных (сигналов) $u_1^0,\ldots,u_n^0$. Отметим, что входные сигналы $u_1^0,\ldots,u_n^0$ могут интерпретироваться как признаки $X_1,\ldots,X_n$ в общей постановке задачи распознавания.

Предположим, что для нейрона с номером i из первого внутреннего слоя связь с рецепторами осуществляется с помощью весовых коэффициентов $w_1^{i0},\ldots,w_n^{i0}$. Сумматор нейрона i первого внутреннего слоя вычисляет взвешенную сумму $\xi^{i0}=\sum_{t=0}^{n} w_t^{i0}u_t^0$.

Сигнал на выходе нейрона i первого внутреннего слоя вычисляется по формуле $u_i^1=\Phi(\xi^{i0})$. Аналогичным образом вычисляются сигналы на выходе нейронов второго внутреннего слоя. Сигналы $g_1,\ldots,g_L$ рассчитываются с помощью той же самой процедуры, которая используется при вычислении сигналов на выходе нейронов из внутренних слоёв. То есть при вычислении g_i на первом шаге соответствующий сумматор вычисляет взвешенную сумму

$$\xi^{iH}=\sum_{t=0}^{r(H)} w_t^{iH}u_t^H,$$

где $w_1^{iH},\ldots,w_{r(h)}^{iH}$ - весовые коэффициенты, характеризующие связь реагирующего нейрона i с нейронами последнего внутреннего слоя H, $u_1^H,\ldots,u_{r(H)}^H$ - сигналы на выходе внутреннего слоя H. Сигнал на выходе реагирующего нейрона i

вычисляется по формуле $g_i = \Phi(\xi^{iH})$. Очевидно, что вектор выходных сигналов является функцией вектора входных сигналов (вектора признаков) и матрицы весовых коэффициентов связей между нейронами.

Аппроксимирующие способности многослойных перцептронов. Один реагирующий нейрон позволяет аппроксимировать области, являющиеся полупространствами, ограниченными гиперплоскостями. Нейронная сеть с одним внутренним слоем позволяет аппроксимировать произвольную выпуклую область в многомерном признаковом пространстве (открытую или закрытую).

Было доказано также, что МП с двумя внутренними слоями позволяет аппроксимировать произвольные области многомерного признакового пространства. Аппроксимирующая способность способность многослойного перцептрона с различным числом внутренних слоёв проиллюстрирована на рисунке 3.

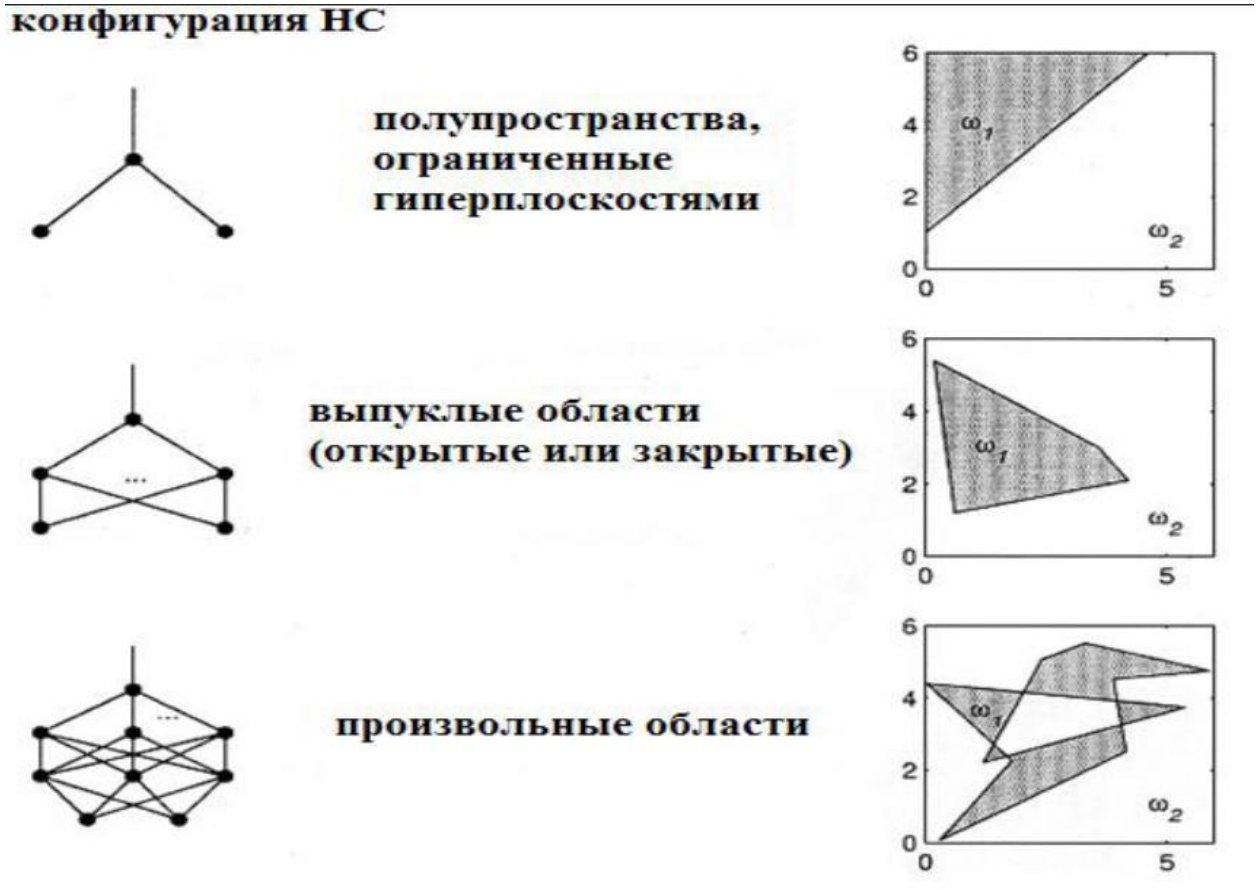


Рис. 3 На рисунке проиллюстрирована аппроксимирующая способность нейронных сетей. с различным числом внутренних слоёв.

Области, соответствующие классам ω_1 и ω_2 разделяются с помощью простого нейрона, а также с помощью многослойных перцептронов с одним и двумя внутренними слоями**.**

Верхняя конфигурация иллюстрирует разделяющую способность отдельного искусственного нейрон, функционирующего в соответствии с моделью Розенблатта.

Ниже представлена конфигурация с одним внутренним слоем нейронов. Данная конфигурация позволяет выделять в многомерном пространстве признаков выпуклые области произвольного типа. Наконец, в нижней части рисунка иллюстрируется разделяющая способность многослойного перцептрона с двумя внутренним слоями. Данная конфигурация позволяет выделять в многомерном пространстве признаков области, которые могут быть получены из набора выпуклых областей с помощью операций объединеия и пересечения. Очевидно, что многослойный перцептрон обладает очень высокой аппроксимирующей способностью.

Обучение многослойных перцептронов. Для обучения метода многослойный перцептрон обычно используется метод обратного распространения ошибки. Данный метод сходен с обучением перцептрона Розенблатта тем, что коррекция изначально произвольных значений весовых коэффициентов ω производится для каждого предъявленного в процессе обучения объекта. Коррекция производится с использованием метода градиентного спуска. То есть коррекция производится в направлении в пространстве коэффициентов ω, в котором максимально снижается целевой функционал. В качестве целевого функционала используется функционал эмпирического риска с квадратичными потерями. Принимается эффективный метод расчёта градиента, основанный на использовании аналитических формул.

4.3 Решающие деревья и леса

4.3.1 Решающие деревья

Структура решающих деревьев. Решающие деревья воспроизводят логические схемы, позволяющие получить окончательное решение о классификации объекта с помощью ответов на иерархически организованную систему вопросов. Причём вопрос, задаваемый на последующем иерархическом уровне, зависит от ответа, полученного на предыдущем уровне. Подобные логические модели издавна используются в ботанике, зоологии, минералогии, медицине и других областях. Пример, решающего дерева, позволяющая грубо оценить стоимость квадратного метра жилья в предполагаемом городе приведена на рисунке 4.

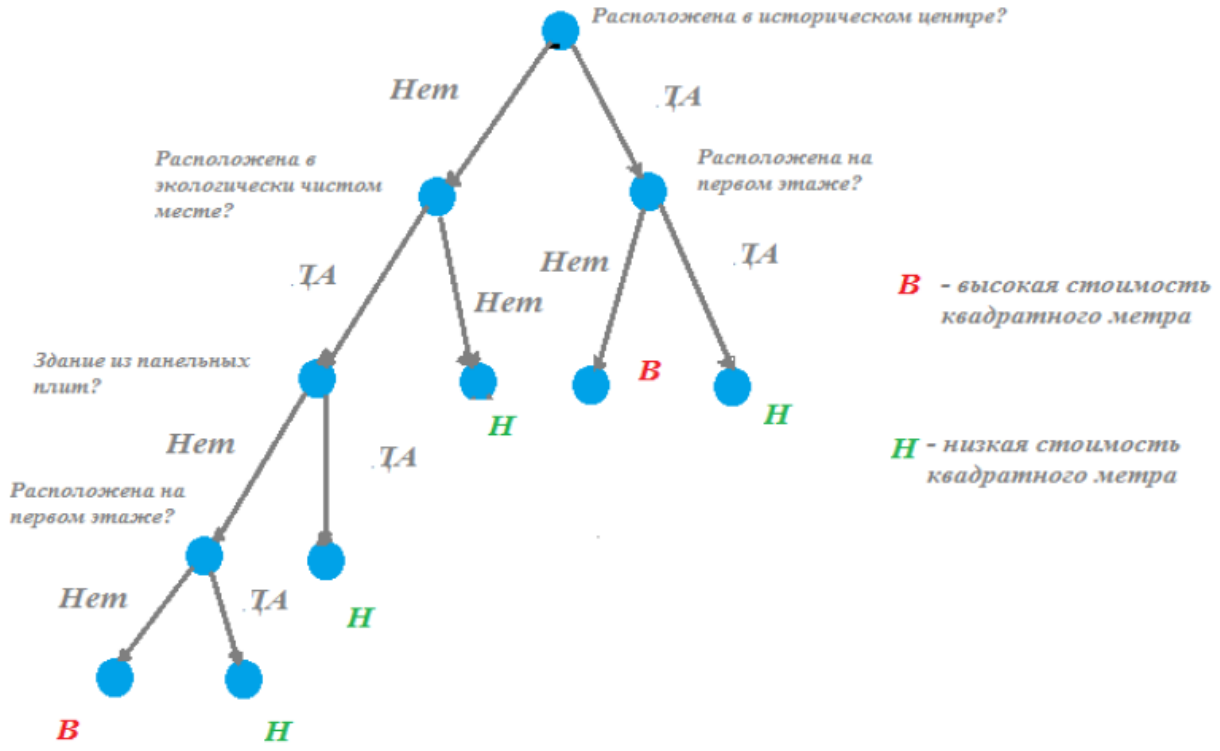


Рис. 4. Изображена структура решающего дерева, оценивающего стоимость квадратного метра жилых помещений. Для простоты выделяются два уровня стоимости – высокий и низкий.

Схеме принятия решений, изображённой на рисунке 1, соответствует связный ориентированный ациклический граф – ориентированное дерево. Дерево включает в себя корневую вершину, инцидентную только выходящим рёбрами, внутренние вершины, инцидентную одному входящему ребру и нескольким выходящим, и листья – концевые

Каждой из вершин дерева за исключением листьев соответствует некоторый вопрос, подразумевающий несколько вариантов ответов, соответствующих выходящим рёбрам. В зависимости от выбранного варианта ответа осуществляется переход к вершине следующего уровня. Концевым вершинам поставлены в соответствие метки, указывающие на отнесение распознаваемого объекта к одному из классов. Решающее дерево называется бинарным, если каждая внутренняя или корневая вершина инцидентна только двум выходящим рёбрам. Бинарные деревья удобно использовать в моделях машинного обучения.

Распознавание с помощью решающих деревьев. Предположим, что бинарное дерево **T** используется для распознавания объектов, описываемых набором признаков $X_1,\ldots,X_n$. Каждой вершине ν дерева **T** ставится в соответствие предикат, касающийся значения одного из признаков. Непрерывному признаку X_j соответствует предикат вида $"X_j \geq \delta_j^{\nu}"$, где δ_j^{ν} - некоторый пороговый параметр. Выбор одного из двух, выходящих из вершины ν рёбер производится в зависимости от значения предиката. Категориальному признаку X_j, принимающему значения из множества $M_{j'} = \{a_1^{j'},\ldots,a_{r(j')}^{j'}\}$ ставится в соответствие предикат вида $"X_{j'} \in M_{j'}^{\nu 1}"$, где $M_{j'}^{\nu 1}$ является элементом дихотомического разбиения $\{M_{j'}^{\nu 1}, M_{j'}^{\nu 2}\}$ множества $M_{j'}$. Выбор одного из двух, выходящих из вершины ν рёбер производится в зависимости от значения предиката. Процесс распознавания заканчивается при достижении концевой вершины (листа). Объект относится классу согласно метке, поставленной в соответствие данному листу.

Обучение решающих деревьев. Рассмотрим задачу распознавания с классами $K_1,\ldots,K_L$. Обучение алгоритма решающее дерево производится по обучающей выборке $\tilde{S}_t$ и включает в себя поиск оптимальных пороговых параметров или оптимальных дихотомических разбиений для признаков $X_1,\ldots,X_n$. При этом поиск производится исходя из требования снижения среднего индекса неоднородности в выборках, порождаемых искомым дихотомическим разбиением обучающей выборки $\tilde{S}_t$.

Индексы неоднородности вычисляется для произвольной выборки $\tilde{S}$, содержащей объекты из классов $K_1,\ldots,K_L$.

При этом используется несколько видов индексов, включая:

- энтропийный индекс неоднородности,

- индекс Джини,

- индекс ошибочной классификации.

Энтропийный индекс неоднородности вычисляется по формуле

$$\gamma_e(\tilde{S}) = -\sum_{i=1}^{L} P_i \ln(P_i),$$

где P_i - доля объектов класса в выборке $\tilde{S}$. При этом принимается, что $0\ln(0) = 0$. Наибольшее значение $\gamma_e(\tilde{S})$ принимает при равенстве долей классов. Наименьшее значение $\gamma_e(\tilde{S})$ достигается при принадлежности всех объектов одному классу. Индекс Джини вычисляется по формуле

$$\gamma_g(\tilde{S}) = 1 - \sum_{i=1}^{L} P_i^2 .$$

Индекс ошибочной классификации вычисляется по формуле

$$\gamma_m(\tilde{S}) = 1 - \max_{i\in\{1,..,L\}}(P_i).$$

Нетрудно понять, что индексы (2) и (3) также достигают минимального значения при принадлежности всех объектов обучающей выборке одному классу. Предположим, что в методе обучения используется индекс неоднородности $\gamma_*(\tilde{S})$. Для оценки эффективности разбиения обучающей выборки $\tilde{S}_t$ на непересекающиеся подвыборки $\tilde{S}_t^l$ и $\tilde{S}_t^r$ используется уменьшение среднего индекса неоднородности в $\tilde{S}_t^l$ и $\tilde{S}_t^r$ по отношению к $\tilde{S}_t$. Данное уменьшение вычисляется по формуле

$$\Delta(\gamma_*, \tilde{S}_t) = \gamma_*(\tilde{S}_t) - P_l\gamma_*(\tilde{S}_t^l) - P_r\gamma_*(\tilde{S}_t^r),$$

где P_l и P_r являются долями $\tilde{S}_t^l$ и $\tilde{S}_t^r$ в полной обучающей выборке $\tilde{S}_t$.

На первом этапе обучения бинарного решающего дерева ищется оптимальный предикат соответствующий корневой вершине. С этой целью оптимальные разбиения строятся для каждого из признаков из набора $X_1,\ldots,X_n$. Выбирается признак $X_{i_{\max}}$ с максимальным значением индекса $\Delta(\gamma_*,\tilde{S}_t)$. Подвыбороки $\tilde{S}_t^l$ и $\tilde{S}_t^r$, задаваемые оптимальным предикатом для $X_{i_{\max}}$ оцениваются с помощью критерия остановки. В качестве критерия остановки может быть использован простейший критерий достижения полной однородности по одному из классов. В случае, если какая-нибудь из выборок $\tilde{S}_t^*$ удовлетворяет критерию остановки, то соответствующая вершина дерева объявляется концевой и для неё вычисляется метка класса. В случае, если выборка $\tilde{S}_t^*$ не удовлетворяет критерию остановки, то формируется новая внутренняя вершина, для которой процесс построения дерева продолжается. Однако вместо обучающей выборки $\tilde{S}_t$ используется соответствующая вновь образованной внутренней вершине v выборка $\tilde{S}_v$, которая равна $\tilde{S}_t^*$. Для данной выборки производятся те же самые построения, которые на начальном этапе проводились для обучающей выборки $\tilde{S}_t$. Обучение может проводиться до тех пор, пока все вновь построенные вершины не окажутся однородными по классам. Такое дерево может быть построено всегда, когда обучающая выборка не содержит объектов с одним и тем же значениям каждого из признаков, принадлежащих разным классам. Однако абсолютная точность на обучающей выборке не всегда приводить к высокой обобщающей способности в результате эффекта переобучения.

Одним из способов достижения более высокой обобщающей способности является использования критериев остановки, позволяющих остановит процесс построения дерева до того, как будет достигнута полная однородность концевых вершин.

Рассмотри несколько таких критериев.

1. *Критерий остановки по минимальному допустимому числу объектов в выборках, соответствующих концевым вершинам.*

2. *Критерий остановки по минимально допустимой величине индекса* $\Delta(\gamma_*,\tilde{S}_t)$. Предположим, что некоторой вершине v соответствует выборка $\tilde{S}_v$, для которой найдены оптимальный признак вместе с оптимальным предикатом, задающим разбиение $\{\tilde{S}_v^l,\tilde{S}_v^r\}$. Вершина v считается внутренней, если индекс $\Delta(\gamma_*,\tilde{S}_t)$ превысил пороговое значение τ и считается концевой в противном случае.

3. Критерий остановки по точности на контрольной выборке. Исходная выборка данных

случайным образом разбивается на обучающую выборку $\tilde{S}_t$ и контрольную выборку $\tilde{S}_c$. Выборка $\tilde{S}_t$ используется для построения бинарного решающего дерева. Предположим, что некоторой вершине v соответствует выборка $\tilde{S}_v$, для которой найдены оптимальный признак вместе с оптимальным предикатом, задающим разбиение $\{\tilde{S}_v^l, \tilde{S}_v^r\}$.

На контрольной выборке $\tilde{S}_c$ производится сравнение эффективность распознающей способности деревьев $\mathbf{T}_v$ и $\mathbf{T}_v^{++}$.

Деревья $\mathbf{T}_v$ и $\mathbf{T}_v^{++}$ включает все вершины и рёбра, построенные до построения вершины v. В дереве $\mathbf{T}_v$ вершина v считается концевой. В дереве $\mathbf{T}_v^{++}$ вершина v считается внутренней, а концевыми считаются вершины, соответствующие подвыборкам $\tilde{S}_v^l$ и $\tilde{S}_v^r$.

Распознающая способность деревьев $\mathbf{T}_v$ и $\mathbf{T}_v^{++}$ сравнивается на контрольной выборке $\tilde{S}_c$. В том, случае если распознающая способность $\mathbf{T}_v^{++}$ превосходит распознающую способность $\mathbf{T}_v$ все дальнейшие построения исходят из того, что вершина v является концевой. В противном случае производится исследование $\tilde{S}_v^l$ и $\tilde{S}_v^r$.

4.Статистический критерий. Заранее фиксируется пороговый уровень значимости (P<0.05,p<0.01 или p<0.001). Предположим, что нам требуется оценить, является ли концевой вершина, для которой найдены оптимальный признак вместе с оптимальным предикатом, задающим разбиение $\{\tilde{S}_v^l, \tilde{S}_v^r\}$. Исследуется статистическая достоверность различий между содержанием объектов распознаваемых классов в подвыборках $\tilde{S}_v^l$ и $\tilde{S}_v^r$. Для этих целей может быть использованы известные статистический критерий: Хи-квадрат и другие критерии. По выборкам $\tilde{S}_v^l$ и $\tilde{S}_v^r$ рассчитывается статистика критерия и устанавливается соответствующее р-значение. В том случае, если полученное р-значение оказывается меньше заранее фиксированного уровня значимости вершина v считается внутренней. В противном случае вершина v считается концевой.

Использование критериев ранней остановки не всегда позволяет адекватно оценить необходимую глубину дерева. Слишком ранняя остановка ветвления может привести к потере информативных предикатов, которые могут быть на самом деле найдены только при достаточно большой глубине ветвления.

В связи с этим нередко целесообразным оказывается построение сначала полного дерева, которое затем уменьшается до оптимального с точки зрения достижения максимальной обучающей способности размера путём объединения некоторых концевых вершин. Такой процесс в литературе принято называть «pruning» («подрезка»).\\

При подрезке дерева может быть использован критерий целесообразности объединения двух вершин, основанный на сравнение на контрольной выборке точности распознавания до и после проведения «подрезки».

Ещё один способ оптимизации обобщающей способности деревьев основан на учёте при «подрезке» дерева до некоторой внутренней вершины v одновременно увеличения точности разделения классов на обучающей выборке и увеличения сложности, которые возникают благодаря ветвлению из v.

При этом прирост сложности, связанный с ветвлением из вершины v, может быть оценён через число листьев в поддереве $\mathbf{T}_v^{sub}$ полного решающего дерева с корневой вершиной v. Следует отметить, что рост сложности является штрафующим фактором, компенсирующим прирост точности разделения на обучающей выборке с помощью включения поддерева $\mathbf{T}_v^{sub}$ в решающее дерево. Разработан целый ряд эвристических критериев, которые позволяют оценить целесообразность включения $\mathbf{T}_v^{sub}$. Данные критерии учитывают одновременно сложность и разделяющую способность.

4.3.2 Решающие леса

В результате многочисленных экспериментов было установлено, что точность нередко значительно возрастает, если вместо отдельных решающих деревьев использовать коллективы (ансамбли) решающих деревьев, которые принято называть решающими лесами. Коллективное решение вычисляется по результатам распознавания отдельными членами ансамбля. В методах решающих лесов в качестве членов ансамблей принято использовать решающих деревьев, которые строятся по искусственно сгенерированным обучающим выборок, статистически сходных с исходной обучающей выборкой.

Получили распространение процедуры построения решающих лесов «бэггинг» и «бустинг»., основанные на различных способах генерации «искусственных» выборок из исходной обучающей выборки.

В методе «бэггинг» (bagging) каждая искусственная случайная выборка является выборкой с возвращениями из исходной обучающей выборки $\tilde{S}_t = \{s_1, \ldots, s_m\}$, также содержащей m объектов. Подобный способ генерации выборок

называют методом «бутрэп» (bootstrap). Название bagging является сокращённым и происходит от полного названия «бутстрэп агрегирование» (Bootstrap Aggregating). Отметим, что искусственная выборка состоит только из объектов исходной обучающей выборки $\tilde{S}_t$. Однако некоторые объекты $\tilde{S}_t$ могут встречаться искусственной выборке по нескольку раз, а некоторые могут вообще отсутствовать.

Для построения коллективного решения может быть использован простейшее решающее правило голосования по большинству: объект относится к тому классу, в который его отнесло большинство деревьев, формирующих лес.

Основной идеей метода бустинг (boosting) является пошаговое наращивание ансамбля деревьев. При этом на каждом шаге к ансамблю присоединяется алгоритм, который был обучен по выборке, искусственно сгенерированной из исходной обучающей выборки $\tilde{S}_t$. В отличие от метода «бэггинг», простая выборка с возвращениями, предполагающая равновероятность всех объектов $\tilde{S}_t$, используется для обучения только на первом шаге. На каждом последующем шаге k объекты в искусственные выборки выбираются с учётом вероятностей, приписанных объектам исходной выборки $\tilde{S}_t$. Последнее распределение вероятностей вычисляется с учётом результатов классификации с помощью ансамбля, использованного на предыдущем шаге. При этом объектам, которые на предыдущем шаге были классифицированы неверно приписываются более высокие веса.

Существуют различные варианты реализации схемы «бустинг», зависящие от способа вычисления вероятностей, приписываемых объектам $\tilde{S}_t$. а также способов вычисления коллективного решения. Одной из наиболее известных вариантов метода «бустинг» является метод Adaptive Boosting (AdaBoost).

4.4 Комбинаторно-логические методы, основанные на принципе частичной прецедентности

Многие прикладные задачи распознавания могут быть успешно решены с помощью методов, основанных на принципе частичной прецедентности. Данный принцип подразумевает поиск по обучающей выборке фрагментов описаний, позволяющих с разной степенью точности разделить распознаваемые классы $K_1,\ldots,K_L$. Распознаваемый объект оценивается по совокупности найденных фрагментов. Одной из первых реализаций принципа частичной прецедентности является тестовый алгоритм, предложенный в 1966 году. Данный алгоритм основан на понятии тупикового теста. Исходный вариант тестового алгоритма предназначен для распознавания объектов, описываемых с помощью бинарных или категориальных признаков $X_1,\ldots,X_n$. Иными словами $X_i \in \{a_1^i,\ldots,a_{k(i)}^i\}$, $i=1,\ldots,n$. Пусть обучающая выборка $\tilde{S}_t$ содержит объекты из классов $K_1,\ldots,K_L$. При этом общее число объектов равно m.

Выборке $\tilde{S}_t$ ставится в соответствие таблица $\mathbf{T}_{nml}$. В строке j таблицы $\mathbf{T}_{nml}$ находятся значения признаков $X_1,\ldots,X_n$ на объекте s_j.

Определение 1. Тестом таблицы $\mathbf{T}_{nml}$ называется такая совокупность столбцов $\{i_1,\ldots,i_r\}$, что для произвольной пары строк s' и s'', соответствующих объектам из разных классов, существует такой столбец i^* из множества $\{i_1,\ldots,i_r\}$, что значения на пересечении i^* со строками s' и s'' различны.

Иными словами набор признаков считается тестом, если описания любых двух объектов из разных классов отличаются хотя бы по одному из признаков, входящих в тест.

Определение 2. Тест $T=\{i_1,\ldots,i_r\}$ называется тупиковым, если никакое его отличное от T подмножество (собственное подмножество) тестом не является

На этапе обучения ищется множество всевозможных тупиковых тестов $\tilde{T}(\tilde{S}_t)$ для таблицы $\mathbf{T}_{nml}$. Предположим что нам требуется распознать объект s_* с векторным описанием $(x_{*1},\ldots,x_{*n})$. Выделим в векторном описании фрагмент

$(x_{i_1},\ldots,x_{i_r})$, соответствующий тесту T из множества $\tilde{T}(\tilde{S}_t)$. Фрагмент $(x_{i_1},\ldots,x_{i_r})$ сравнивается с множеством фрагментов строк $(x_{ji_1}^T,\ldots,x_{ji_r}^T)$ таблицы $\mathbf{T}_{nml}$,

соответствующих классу K_l : $\{(x_{ji_1}^T,\ldots,x_{ji_r}^T) \mid s_j \in K_l\}$ $(x_{i_1},\ldots,x_{i_r})$. В

случаях, когда выполняются равенства $x_{ji_1}^T = x_{*i_1},\ldots,x_{ji_r}^T = x_{*i_r}$

фиксируем полное совпадение. Обозначим число полных совпадений распознаваемого объекта s_* с объектами K_l из $\tilde{S}_t$ через $G_l(T,s_*)$.

Оценка объекта s_* за класс K_l вычисляется по формуле:

$$\gamma_l(s_*) = \frac{1}{m_l} \sum_{T \in \tilde{T}(\tilde{S}_t)} G_l(T,s_*) \ ,$$

где m_l - число объектов обучающей выборки из класса K_l . Классификация объекта может производится с помощью по вектору оценок $[\gamma_1(s_*),\ldots,\gamma_L(s_*)]$ с помощью стандартного решающего правила, т.е. объект относится в тот класс, оценка за который максимальна

Задача о поиске всевозможных тупиковых тестов сводится к известной задаче комбинаторного анализа о поиске всевозможных тупиковых покрытий элементам.

Нахождение всех тупиковых тестов является сложной комбинаторной задачей. Однако эффективные алгоритмы поиска разработаны для некоторых типов таблиц. При решении практических задач эффективен подход , основанный на вычислении только части тупиковых тестов.

Другим известным классом алгоритмов распознавания , основанным на принципе частичной прецедентности, являются алгоритмы типа КОРА. В отличие от тестового алгоритма, где в качестве информативных элементов используются несжимаемые наборы признаков – тупиковые тесты, в алгоритмах типа КОРА в качестве информативных элементов используются несжимаемые фрагменты описаний эталонных объектов обучающей выборки.

Определение 3.Пусть $(x_{v1},\ldots,x_{vn})$ - признаковое описание объекта $s_v \in K_l$. Набор $(x_{vj_1},\ldots,x_{vj_r})$ называется представительным набором для класса K_l , если для произвольной строки K_l таблицы $\mathbf{T}_{nml}$ соответствующей объекту

$s_u \notin K_i$ такое, что существует такое j' из множества $\{j_1,\ldots,j_r\}$, что $x_{vj'} \neq x_{uj'}$.

Определение 4. Представительный набор называется тупиковым, если никакое его собственное подмножество представительным набором не является.

На этапе обучения для каждого из классов $K_1,\ldots,K_L$ по таблице $\mathbf{T}_{nml}$ ищется множество всевозможных тупиковых представительных наборов. Обозначим через $\tilde{V}_l$ - множество всевозможных представительных наборов для класса K_l. Предположим, что нам требуется распознать объект s_* с описанием $(x_{*1},\ldots,x_{*n})$. Пусть $v=(x_{ui_1},\ldots,x_{ui_r})$ - представительный набор. Функция $\mathrm{B}(s_*,v)$ равна 1, если $(x_{*i_1}=x_{vi_1},\ldots,x_{*i_r}=x_{vi_r})$, и $\mathrm{B}(s_*,v)$ равна 0 в противном случае.

Оценка s_* за класс K_l вычисляется по формуле

$$\Gamma_l(s_*)=\frac{1}{|\tilde{V}_i|}\sum_{u\in\tilde{V}_i}\mathrm{B}(s_*,u)\ .$$

Первоначальные варианты тестового алгоритма и алгоритма типа КОРА были разработаны для бинарных или категориальных переменных. Они не могут быть напрямую использованы в задачах с признаками, принимающими значения из интервалов вещественной оси. Для того, чтобы обеспечить возможность работы с подобной информацией могут быть использованы два подхода**.**

а) Первый подход основан на разбиении области возможных значений каждого вещественнозначного признака на k связных подмножеств (интервалов, полуинтервалов, отрезков). Значению признака, принадлежащего элементу j разбиения присваивается само значение j. Разбиение оптимизируется с целью достижения максимального разделения классов. Выбирается такое число элементов разбиения k, при котором достигается максимальная точность распознавания.

Другой подход основан на модификации понятий теста и представительного набора с использованием пороговых параметров $\varepsilon_1,\ldots,\varepsilon_n$, которые задаются для признаков $X_1,\ldots,X_n$.

Определение 5. Тестом таблицы $\mathbf{T}_{nml}$ называется такая совокупность столбцов $\{i_1,\ldots,i_r\}$, что для произвольной пары строк s' и s'', соответствующих объектам из разных классов, существует такой столбец i^* из множества $\{i_1,\ldots,i_r\}$, что абсолютная

величина разницы значений, стоящих на пересечении i^* со строками s' и s'' превышает ε_{i^*} .

Аналогичным образом вводится модифицированное определение представительного набора**.**

Главным требованием при выборе ε - порогов является достижение максимальной отделимости объектов разных классов при сохранении сходства внутри классов.

Поиск тупиковых тестов и тупиковых представительных наборов при модифицированных определениях аналогичен их поиску в первоначальных вариантах методов.

Тестовый алгоритм и алгоритм с представительными наборами являются частью более общей конструкции, основанной на принципе частичной прецедентности и носящей название алгоритмов вычисления оценок.

Существует много вариантов моделей данного типа. Причём конкретный вид модели определяется выбранными способами задания различных её элементов. Рассмотрим основные составляющие модели

Задание системы опорных множеств. Под опорными множествами модели АВО понимается наборы признаков, по которым осуществляется сравнение распознаваемых и эталонных объектов. Примером системы опорных множеств является множество тупиковых тестов. Система опорных множеств Ω_A некоторого алгоритма A может задаваться через систему подмножеств множества $\{1,\ldots,n\}$ или через систему характеристических бинарных векторов.

Каждому подмножеству $\{1,\ldots,n\}$ может быть сопоставлен бинарный вектор размерности n. Пусть $\{i_1,\ldots,i_k\} \subseteq \{1,\ldots,n\}$. Тогда $\{i_1,\ldots,i_k\}$ сопоставляется вектор $\boldsymbol{\omega} = \{\omega_1,\ldots,\omega_n\}$, все компоненты которого равны 0 кроме равных 1 компонент $\omega_{i_1},\ldots,\omega_{i_k}$. Теоретические исследования свойств тупиковых тестов для случайных бинарных таблиц показали, что характеристические векторы для почти всех тупиковых тестов имеют асимптотически (при неограниченном возрастании размерности таблицы обучения) одну и ту же длину.

Данный результат является обоснованием выбора в качестве системы опорных векторов всевозможных наборов, включающих фиксированное число признаков k или

$\Omega_A = \{\boldsymbol{\omega} : |\boldsymbol{\omega}| = k\}$. Оптимальное значение k находится в процессе обучения или задаётся экспертом. Другой часто используемой системе опорных множеств соответствует множество всех подмножеств $\{1,\ldots,n\}$ за исключением пустого множества.

Задание функции близости. Близость между объектами по опорноным множествам задаётся аналогично тому, как задаётся близость между объектами по тупиковым тестам или представительным наборам. Пусть набор номеров $\{i_1,\ldots,i_k\}$ соответствует характеристическому вектору $\boldsymbol{\omega}$.

Фрагмент $(x_{\mu i_1},\ldots,x_{\mu i_k})$ описания $(x_{\mu 1},\ldots,x_{\mu n})$ объекта $\mathbf{x}_\mu$ называется $\boldsymbol{\omega}$ - частью объекта s_μ. Под функцией близости $\mathbf{B}_{\boldsymbol{\omega}}(s_\mu, s_\upsilon)$ понимается функция от соответствующих $\boldsymbol{\omega}$ -частей сравниваемых объектов, принимающая значения 1 (объекты близки) или 0 (объекты удалены).

Функции близости обычно задаются с помощью пороговых параметров $(\varepsilon_1,\ldots,\varepsilon_n)$, характеризующих близость объектов по отдельным признакам.

Примеры функций близости.

1) $\mathbf{B}_{\boldsymbol{\omega}}(s_\mu, s_\upsilon) = 1$, если при произвольном $i \in \{1,\ldots,n\}$, при котором $\omega_i = 1$, всегда выполняется неравенство $|x_{\mu i} - x_{\nu i}| < \varepsilon_i$. $\mathbf{B}_{\boldsymbol{\omega}}(s_\mu, s_\upsilon) = 0$, если существует такое $i' \in \{1,\ldots,n\}$, что одновременно $\omega_{i'} = 1$ и $|x_{\mu i'} - x_{\nu i'}| > \varepsilon_{i'}$.

2) Пусть ε - скалярный пороговый параметр. Функция $\mathbf{B}_{\boldsymbol{\omega}}(s_\mu, s_\upsilon) = 1$, если выполняется неравенство $\sum_{i=1}^{n} \omega_i |x_{\mu i} - x_{\nu i}| < \varepsilon$. В противном случае $\mathbf{B}_{\boldsymbol{\omega}}(s_\mu, s_\upsilon) = 0$.

Важным элементом АВО является оценка близости распознаваемого объекта s_* к эталону s_μ по заданной ω - части. Данная оценка близости, которая будет обозначаться $G(\boldsymbol{\omega}, s_*, s_\mu)$, формируется на основе введённых ранее функций близости и, возможно, дополнительных параметров. Приведём примеры функций близости различного уровня сложности:

А) $G(\boldsymbol{\omega}, s_*, s_\mu) = \mathrm{B}_{\boldsymbol{\omega}}(s_*, s_\mu)$,

B) $G(\boldsymbol{\omega}, s_*, s_\mu) = p_{\boldsymbol{\omega}} \mathrm{B}_{\boldsymbol{\omega}}(s_*, s_\mu)$, где $p_{\boldsymbol{\omega}}$ - параметр, характеризующий информативность опорного множества с характеристическим вектором $\boldsymbol{\omega}$.

C) $G(\boldsymbol{\omega}, s_*, s_\mu) = \gamma_\mu (\sum_{i=1}^{n} p_i \omega_i) \mathrm{B}_{\boldsymbol{\omega}}(s_*, s_\mu)$, где γ_μ - параметр, характеризующий информативность эталона s_μ, параметры $p_1, \ldots, p_n$ характеризуют информативность отдельных признаков.

Оценка объекта s_* за класс K_l при фиксированном характеристическом векторе $\boldsymbol{\omega}$ может вычисляться как среднее значение близости s_* к эталонным объектам из класса K_l: $\Gamma_l(\boldsymbol{\omega}, s_*) = \frac{1}{m_l} \sum_{s_\mu \in K_l} G(\boldsymbol{\omega}, s_*, s_\mu)$, где m_l - число объектов обучающей выборки из класса K_l.

Общая оценка s_* за класс K_l вычисляется как сумма оценок $\Gamma_l(\boldsymbol{\omega}, s_*)$ по опорным множествам из системы Ω_A:

$$\hat{\boldsymbol{\Gamma}}_l(s_*) = \sum_{\boldsymbol{\omega} \in \Omega_A} \Gamma_l(\boldsymbol{\omega}, s_*) \qquad (1)$$

Наряду с формулой (1) используется формула

$$\hat{\boldsymbol{\Gamma}}_l(s_*) = w_l \sum_{\boldsymbol{\omega} \in \Omega_A} \Gamma_l(\boldsymbol{\omega}, s_*) \qquad (2)$$

Использование взвешивающих параметров $w_1, \ldots, w_L$ позволяет регулировать доли правильно распознанных объектов $K_1, \ldots, K_L$.

Прямое вычисление оценок за классы по формулам (1) и (2) в случаях, когда в качестве систем опорных множеств используются наборы с фиксированным числом признаков или всевозможные наборы признаков, оказывается практически невозможным при сколь либо высокой размерности признакового пространства из-за необходимости вычисления огромного числа значений функций близости.

Однако при равенстве весов всех признаков существуют эффективные формулы для вычисления оценок по формуле (1). Предположим, что оценки близости распознаваемого объекта s_* к эталону s_μ по заданной $\boldsymbol{\omega}$ - части вычисляются по формуле (А).

Тогда оценка по формуле (1) принимает вид $\hat{\boldsymbol{\Gamma}}_l(s_*) = \sum_{s_\mu \in K_l} \sum_{\boldsymbol{\omega} \in \Omega_A} \mathrm{B}_{\boldsymbol{\omega}}(s_*, s_\mu)$

Рассмотрим сумму $\sum_{\omega\in\Omega_A} \mathrm{B}_{\omega}(s_*, s_\mu$. Предположим, что общее число признаков, по которым объект s_* близок объекту s_μ равно $d(s_*, s_\mu)$. Иными словами $d(s_*, s_\mu) = |D(s_*, s_\mu)|$, где $D(s_*, s_\mu) = \{i : | x_{*i} - x_{\mu i} | < \varepsilon_i\}$. Очевидно функция близости $\mathrm{B}_{\omega}(s_*, s_\mu)$ равна 1 тогда и только тогда, когда опорное множество, задаваемое характеристическим вектором ω, полностью входит в множество $D(s_*, s_\mu)$. Во всех остальных случаях $\mathrm{B}_{\omega}(s_*, s_\mu) = 0$.

Предположим, что система опорных множеств удовлетворяет условию $\Omega_{\mathrm{A}} = \{\omega : |\omega| = k\}$. Очевидно, что число опорных множеств, удовлетворяющих условию $\mathrm{B}_{\omega}(s_*, s_\mu) = 1$, равно $\mathrm{C}^k_{d(s_*, s_\mu)}$. Откуда следует, что $\sum_{\omega\in\Omega_A} \mathrm{B}_{\omega}(s_*, s_\mu) = \mathrm{C}^k_{d(s_*, s_\mu)}$. Следовательно оценка по формуле (1) может быть записана в виде

$$\hat{\Gamma}_l(s_*) = \frac{1}{m_l} \sum_{s_\mu \in K_l} \gamma_\mu \mathrm{C}^k_{d(s_*, s_\mu)} \ . \qquad (3)$$

Предположим, что система Ω_A включает в себя всевозможные опорные множества. В этом случае число опорных множеств в Ω_A, удовлетворяющих условию $\mathrm{B}_{\omega}(s_*, s_\mu) = 1$ равно $2^{d(s_*, s_\mu)}$-1. Следовательно оценка по формуле (1) может быть записана в виде

$$\hat{\Gamma}_l(s_*) = \frac{1}{m_l} \sum_{s_\mu \in K_l} \gamma_\mu [2^{d(s_*, s_\mu)} \text{-} 1]$$

Обучение алгоритмов вычисления оценок. Для обучения алгоритмов АВО в общем случае может быть использован тот же самый подход, который используется для обучения в методе «Линейная машина». Предположим, что решается задача обучения алгоритмов

для распознавания объектов , принадлежащих классам $K_1,\ldots,K_L$. При правильного распознавания объекта $s_j \in K_l$ должна выполняться система неравенств

$$\tilde{\mathbf{\Gamma}}_l(s_l) > \tilde{\mathbf{\Gamma}}_{l'}(s_j), \text{ где } l' \in \{1,\ldots,L\} \setminus l .$$

В наиболее общем из приведённых выше вариантов модели АВО обучение может быть сведено к поиску максимальной совместной подсистемы системы неравенств

$$\forall s_j \in K_l \cap \tilde{S}_t \qquad (4)$$
$$\frac{1}{m_l}\sum_{S_\mu \in K_l} \gamma_\mu \left(\sum_{t=1}^{n} p_t \omega_t\right) \mathrm{B}_{\boldsymbol{\omega}}(s_j, s_\mu) > \frac{1}{m_{l'}}\sum_{S_\nu \in K_{l'}} \gamma_\nu \left(\sum_{t=1}^{n} p_t \omega_t\right) \mathrm{B}_{\boldsymbol{\omega}}(s_j, s_\nu)$$

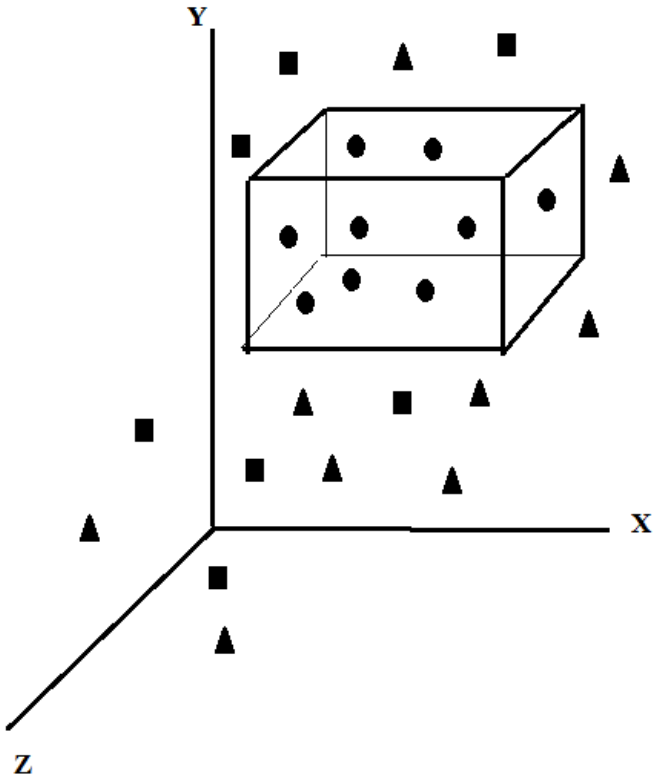
Поиск максимальной совместной подсистемы системы (2) может производиться с использованием эвристического релаксационного метода, аналогичного тому, что был использован при обучении алгоритма «Линейная машина».

Тестовый алгоритм был впервые предложен в работе [], где для решении задачи геологического прогнозирования.

4.5 Методы, основанные на голосовании по системам логических закономерностей

Одним из эффективных подходов к решению задач прогнозирования и распознавания является использование коллективных решений по системам закономерностей. Под закономерностью понимается распознающий или прогностический алгоритм, определённый на некоторой подобласти признакового пространства или связанный с некоторым подмножеством признаков.

В качестве примера закономерностей могут быть приведены представительные наборы, являющиеся по сути подмножествами признаковых описаний, характерных для одного из распознаваемых классов. Аналогом представительный наборов в задач с вещественнозначной информацией являются логические закономерности классов. Под логической закономерностью класса K_l понимается область признакового пространства,



имеющая ф орму гиперпараллелепипеда и содержащая только объекты из K_l ..

Рис. На рисунке представлена логическая закономерность, содержащая объекты класса K_1, обозначенные синим кружком, и не содержащая объектов обозначенных по другому классов K_2 и K_3

Математически логическая закономерность класса K_l, которую мы будем обозначать $r(l)$, описывается с помощью наборов предикатов вида

$$P_i[r(l)] = "a_i^{r(l)} \le x_i \le b_i^{r(l)}" , \qquad (1)$$

где $i = 1,\ldots,n$.

Напомним, что предикатом называется утверждение, принимающее значения «ИСТИНА» или «ЛОЖЬ» в зависимости от значений входящих в них переменных. Полностью логическая закономерность задаётся конъюнкцией предикатов вида (1):

$$\mathbf{P}[r(l)] = P_1[r(l)] \& \ldots \& P_n[r(l)] \qquad (2)$$

Очевидно, что множество векторов $(x_1,\ldots,x_n)$, для которых, как раз представляет собой гиперпараллелепипед в многомерном пространстве признаков. Не все признаки являются на самом деле существенными для закономерности $r(l)$. Для несущественного признака $x_{i'}$ отрезок $[a_{i'}^{r(l)}, b_{i'}^{r(l)}]$ совпадает с отрезком, из которого принимает значения признак $x_{i'}$.

На этапе обучения для каждого класса K_l строится множество логических закономерностей $\tilde{R}_l$. Границы $(a_i^{r(l)}, b_i^{r(l)})$ подбираются таким образом, чтобы равенство $\mathbf{P}[r(l)] =$«ИСТИНА» выполнялось бы на максимально большом числе объектов обучающей выборки из класса K_l и равенство $\mathbf{P}[r(l)] =$«ЛОЖЬ» выполнялось бы на всех объектах обучающей выборки из класса K_l. Наряду с полными логическими закономерностями, удовлетворяющими последним условиям, используются также частичные логические закономерности, для которых допускается попадание в них небольшой доли объектов чужих классов. Методы построения логических закономерностей подробно излагаются в работе [11], а также книге [9].

Предположим, что нам требуется распознать новый объект s^*. Для каждого класса K_l ищется число закономерностей в $\tilde{R}_l$, для которых $\mathbf{P}[r(l)] =$ «ИСТИНА». В качестве оценки за класс K_l используется доля таких закономерностей в $\tilde{R}_l$. Классификация s^* производится с помощью стандартного решающего правила. То есть объект относится в тот класс, оценка за который максимальна.

4.6 Метод мультимодельных статистически взвешенных синдромов

Метод мультимодельных статистически взвешенных синдромов является методом распознавания, основанном на принятии коллективных решений по системам синдромов. Под "синдромом" понимается такая область признакового пространства, в которой содержание объектов одного из классов, отличается от содержания объектов этого класса в обучающей выборке или по крайней мере в одной из соседних областях. Синдромы ищутся для каждого из распознаваемых классов с помощью построения оптимальных разбиений интервалов допустимых значений единичных признаков или совместных двумерных областей допустимых значений пар признаков. Пример синдромов, характеризующих разделение двух классов , приведён на рисунке 2

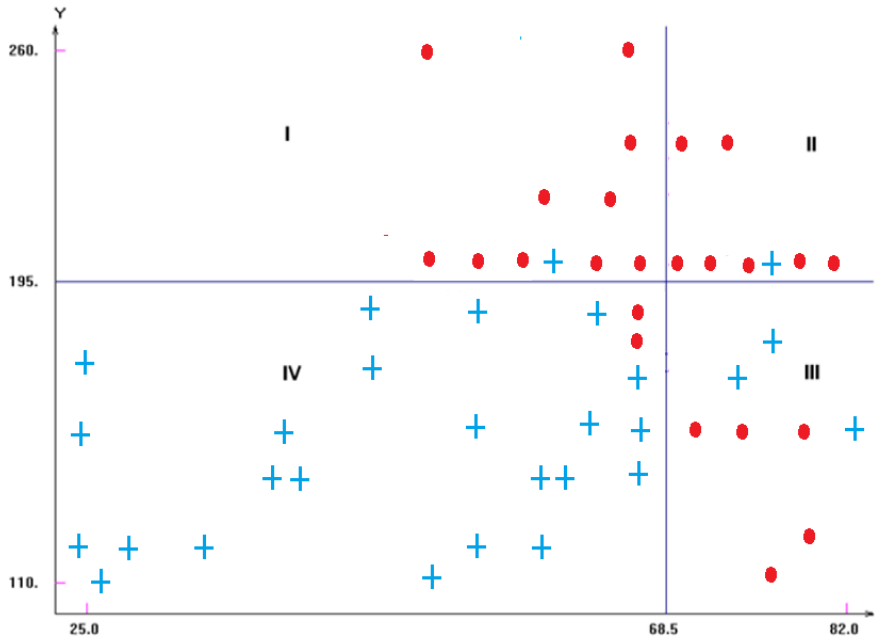


Рис. 2 Внутри синдромов I (верхний слева) и II (верхний справа) преобладают объекты класса K_1 , обозначенного ●. Внутри синдрома IV преобладают объекты класса K_2. обозначенные +.

Поиск синдромов производится с использованием четырёх семейств разбиений, имеющих различный уровень сложности. Примеры разбиений для каждого из семейств приведены на рисунке 3. Семейство I включает всевозможные разбиения интервалов допустимых значений отдельных признаков на два интервала с помощью одной

граничной точки. Семейство II включает всевозможные разбиения интервалов допустимых значений отдельных признаков на 3 интервала с помощью двух граничных точек. Семейство III включает всевозможные разбиения совместных двумерных областей допустимых значений пар признаков на 4 подобласти с помощью двух граничных точек (по одной точке для каждого из двух признаков).

Семейство IV включает всевозможные разбиения совместных двумерных областей допустимых значений пар признаков на 2 подобласти с помощью прямой граничной линии, произвольно ориентированной относительно координатных осей.

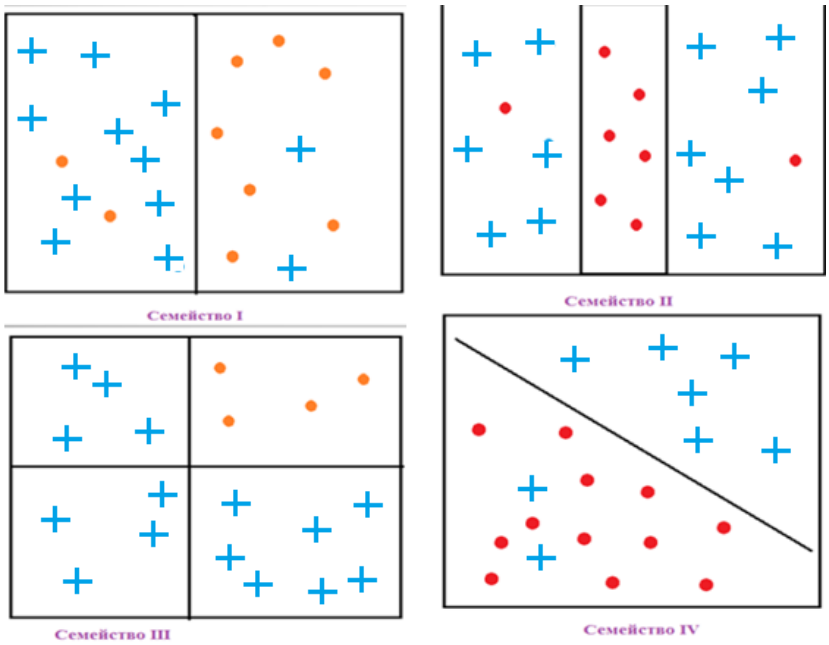


Рис 3. Примеры разбиений для каждого из четырёх семейств, используемых в методе СВС.

В ходе поиска выбирается разбиение с максимальным значением функционала качества. В различных вариантах метода используется два функционала качества, зависящих от обучающей выборки $\tilde{S}_t$, распознаваемого класса K_l, и разбиения R:

- интегральный $F_i(\tilde{S}_t, K_l, R)$;
- локальный $F_{loc}(\tilde{S}_t, K_l, R)$.

Обозначим через $q_1,\ldots,q_r$ элементы некоторого разбиения R. Пусть ν_0^l является долей объектов класса K_l в обучающей выборке $\tilde{S}_t$. ν_i^l - доля объектов K_l среди объектов, описания которых принадлежат элементу q_i, m_i - число объектов, описания которых принадлежат q_i. Интегральный функционал задаётся формулой

$F_i(\tilde{S}_t, K_l, R) = \frac{1}{\nu_0^l(1-\nu_0^l)} \sum_{i=1}^{r} (\nu_i^l - \nu_0^l)^2 m_i$. В то время как локальный функционал задаётся формулой $F_{loc}(\tilde{S}_t, K_l, R) = \max_{i=1,\ldots,r} \frac{(\nu_i^l - \nu_0^l)^2 m_i}{\nu_0^l(1-\nu_0^l)}$

Метод СВС, впервые предложенный в работе [13] был основан на использовании одномерных семейств разбиений. Позже была предложена модификация СВС –метод мультимодельные статистически взвешенных синдромов (МСВС) [25]. В методе МСВС наряду с одномерными семействами I и II используются также семейства III и IV. Синдромы, задаваемые некоторым оптимальным разбиением R^* включаются в финальный набор, используемый в дальнейшем для распознавания новых объектов, если R^* удовлетворяет специальному критерию. В методе СВС для поиска синдромов используется интегральный функционал $F_i(\tilde{S}_t, K_l, R)$. Для формирования финального набора используется простой критерий: все элементы оптимального разбиения R^* включаются в набор, если величина интегрального функционала $F_i(\tilde{S}_t, K_l, R^*)$ превышает задаваемый пользователем порог δ. Опыт решения прикладных задач показывает, что эффективность распознавания достигается при значениях δ, меняющихся от 2 до 10. Несколько более сложный критерий используется в методе МСВС. Для поиска синдромов используется локальный функционал $F_{loc}(\tilde{S}_t, K_l, R)$. Синдромы оптимального разбиения R^* включаются в финальный набор в случае выполнения неравенства $\kappa F_{loc}(\tilde{S}_t, K_l, R) > \delta$, где величина параметра κ зависит от сложности используемой модели. Эксперименты на прикладных задачах показали, что высокая эффективность достигается при $\kappa = 1$ для простейших разбиений из семейства I и $\kappa = 0.5$ для разбиений из семейства II-IV.

Предположим, что на этапе обучения для класса K_l найдено множество синдромов $\tilde{Q}_l$. Пусть описание $\mathbf{x}^*$ распознаваемого объекта s^* принадлежит синдромам $q_1, \ldots, q_r$ из множества $\tilde{Q}_l$. Оценка s^* за класс K_l вычисляется по формуле

$$\Gamma(s^*, K_l) = \frac{\sum_{i=1}^{r} w_i^l \nu_i^l}{\sum_{i=1}^{r} w_i^l},$$

где ν_i^l - доля объектов класса K_l в синдроме q_i , w_i^l - вес синдрома при классификации объектов класса K_l , который вычисляется по формуле $w_i^l = \frac{m_i}{m_i+1} \frac{1}{\nu_0^l(1-\nu_0^l)}$, где m_i - число объектов обучающей выборки, попавших в синдром q_i . Данная формула была получена в работе [] через максимизацию специального функционала, сходного с функционалом правдоподобия.

4.7 Метод опорных векторов.

4.7.1 Линейная разделимость.

Принцип максимизации зазора. Метод опорных векторов является универсальным методом распознавания, позволяющим наряду с линейными реализовывать также нелинейные решающие правила. Исходный вариант метода был предложен для задач с двумя распознаваемыми классами K_1 и K_2. В случаях, когда объекты разных классов в обучающей выборке линейно разделимы, обычно существует целая совокупность линейных поверхностей, осуществляющих такое разделение. На рисунке 1 представлены двумерные данные, где объекты двух классов могут быть раделены с помощью прямых A, B, C, D. Однако наша интуиция, подсказывает что наилучшей обобщающей способностью должна обладать разделяющая прямая F, одинаково удалённая от групп объектов из разных классов. Однако наша интуиция, подсказывает что наилучшей обобщающее

й способностью должна обладать разделяющая прямая F, одинаково удалённая от групп объектов из разных классов.

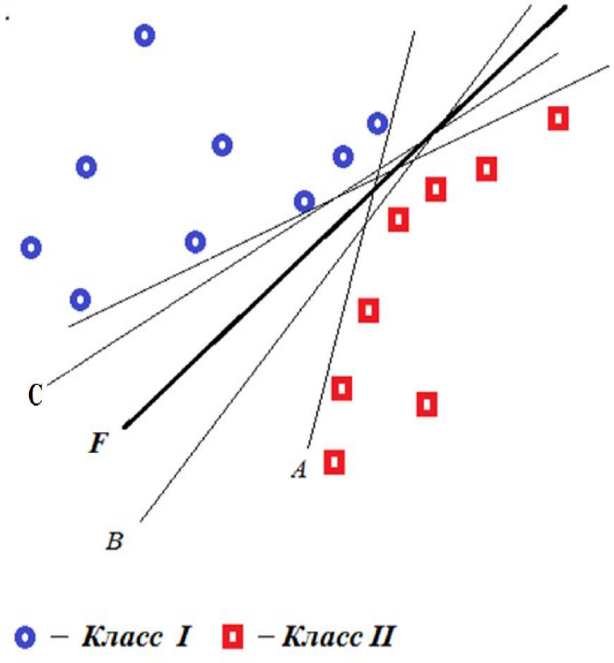


Рис. 1 Иллюстрируются различные варианты разделения классов K_1 и K_2.с помощью линейных границ.

Интуитивные представления об оптимальной разделимости формализует проведение разделяющей гиперплоскости посередине между двумя параллельными гиперплоскостями, каждая из которых отделяет объекты одного из классов. При этом две плоскости строятся таким образом, чтобы «зазор» между ними был бы максимальным.

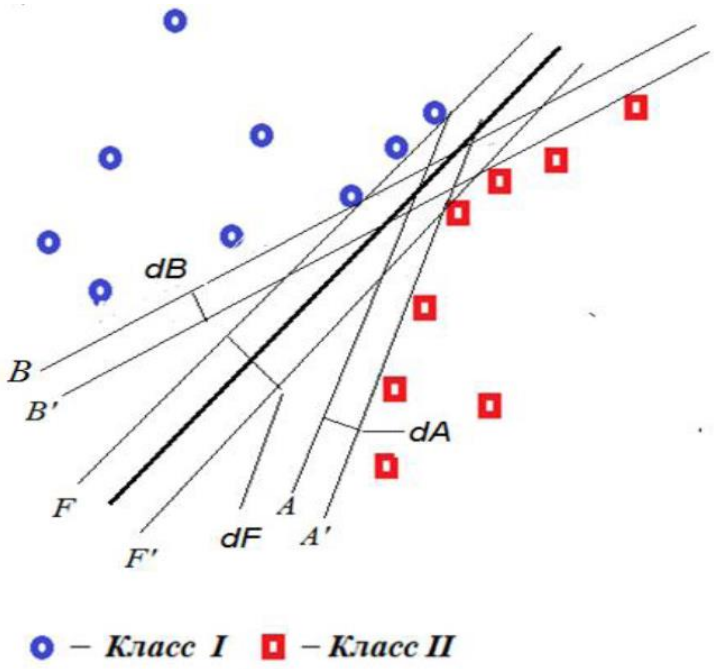


Рис. 1 Иллюстрируются разделение классов K_1 и K_2.с помощью линейных границ с испо льзованием концепции максимального «зазора».

Напомним, что пара параллельных гиперплоскостей P_1 и P_2 в многомерном пространстве $\mathbf{R}^n$ описывается с помощью уравнений:

$$(\mathsf{P}_1) \qquad \mathbf{w}\mathbf{x}^t = b_1, \tag{1}$$

$$(\mathsf{P}_2) \qquad \mathbf{w}\mathbf{x}^t = b_2,$$

где $\mathbf{w}$ является направляющим вектором для гиперплоскостей.

Пусть $\mathbf{z} = \lambda\mathbf{w}$, где λ - некоторое вещественное число. Нетрудно таким образом подобрать λ и b, чтобы система

$$(\mathsf{P}_1) \qquad \mathbf{z}\mathbf{x}^t = b + 1, \tag{2}$$

$$(\mathsf{P}_2) \qquad \mathbf{z}\mathbf{x}^t = b - 1,$$

Описывала те же самые гиперплоскости, что и система (1). Пусть точки $\mathbf{x}_1$ и $\mathbf{x}_2$ принадлежат плоскостям P_1 и P_2 соответственно. Расстояние (величина зазора) δ между гиперплоскостями P_1 и P_2 равно проекции разности $(\mathbf{x}_1 - \mathbf{x}_2)$ на направление $\mathbf{z}$, Данная проекция по определению равна $\frac{\mathbf{z}(\mathbf{x}_1^t - \mathbf{x}_2^t)}{|\mathbf{z}|}$. Однако согласно системе (2) $\frac{\mathbf{z}(\mathbf{x}_1^t - \mathbf{x}_2^t)}{|\mathbf{z}|} = \frac{2}{|\mathbf{z}|}$. Следовательно задача поиска двух максимально удалённых друг от друга параллельных гиперплоскостей, каждая из которых отделяет объекты одного из классов, может быть сведена к оптимизационной задаче с ограничениями.

$$\frac{2}{|\mathbf{z}|} \to \max \qquad (3)$$

$$\mathbf{z}\mathbf{x}_j^t \geq b + 1 \quad \text{при} \quad s_j \in \tilde{S}_t \cap K_1$$

$$\mathbf{z}\mathbf{x}_j^t \leq b - 1 \quad \text{при} \quad s_j \in \tilde{S}_t \cap K_2, \quad j = 1, \ldots, m.$$

При этом оптимизация производится по компонентам направляющего вектора $\mathbf{z} = (z_1, \ldots, z_n)$ и параметру сдвига b.

Введём обозначение: $\alpha_j = 1$ при $s_j \in \tilde{S}_t \cap K_1$ и $s_j \in \tilde{S}_t \cap K_2$. Учитывая, функция $\frac{2}{|\mathbf{z}|}$ монотонно возрастает с уменьшением $\sum_{i=1}^{n} z_i^2$, переходим от задачи (3) к задаче

$$\frac{1}{2}\sum_{i=1}^{n} z_i^2 \to \min \qquad (4)$$

$$\alpha_j(\mathbf{z}\mathbf{x}_j^t - b) \geq 1, \ j = 1, \ldots, m.$$

Задача (4) относится к хорошо изученному классу задач квадратичного программирования.

Решение задачи квадратичного программирования. Важным инструментом исследования экстремальных значений оптимизируемых функций при ограничениях является функция Лагранжа или лагранжиан, который для задачи (4) записывается в виде

$$\mathbf{L}(\mathbf{z},b,\boldsymbol{\lambda}) = \tfrac{1}{2}\sum_{i=1}^{n} z_i^2 - \sum_{j=1}^{m} \lambda_j[\alpha_j(\mathbf{z}\mathbf{x}_j^t - b) - 1],$$

где $(\lambda_1,\ldots,\lambda_m)$ являются неотрицательными вещественными, которые называются множителями Лагранжа.

Из известной теоремы Каруша-Куна-Такера (ККТ) следует, что для точки $(\mathbf{z}^*,b^*)$, в которой функция $\frac{1}{2}\sum_{i=1}^{n} z_i^2$ достигает своего минимума при ограничениях задачи (4), и некоторого вектора значений неотрицательных множителей Лагранжа $\boldsymbol{\lambda}^* = (\lambda_1^*,\ldots,\lambda_m^*)$ соблюдаются условия стационарности лагранжиана $\mathbf{L}(\mathbf{z},b,\boldsymbol{\lambda})$ по переменным $(\mathbf{z},b)$.

Также из теоремы ККТ следует необходимость выполнения m равенств, которые носят название условий дополняющей нежёсткости

$$\lambda_j^*[\alpha_j(\mathbf{z}^*\mathbf{x}_j^t - b) - 1] = 0, \quad j = 1,\ldots,m$$

Условия стационарности заключаются в выполнении n равенств

$$\left.\frac{\partial \mathbf{L}(\mathbf{z},b,\boldsymbol{\lambda})}{\partial z_i}\right|_{(\mathbf{z}^*,b^*,\boldsymbol{\lambda}^*)} = z_i^* - \sum_{j=1}^{m} \lambda_j^* \alpha_j x_{ji} = 0, i = 1,\ldots,n \qquad (5)$$

В векторной форме система (5) принимает вид

$$\mathbf{z}^* - \sum_{j=1}^{m} \lambda_j^* \alpha_j \mathbf{x}_j = 0.$$

Из условия стационарности также следует выполнение равенства

$$\left.\frac{\partial \mathbf{L}(\mathbf{z},b,\boldsymbol{\lambda})}{\partial b}\right|_{(\mathbf{z}^*,b^*,\boldsymbol{\lambda}^*)} = \sum_{j=1}^{m} \lambda_j^* \alpha_j = 0 \qquad (6)$$

Условия стационарности (5,6) для лагранжиана $\mathbf{L}(\mathbf{z}, b, \boldsymbol{\lambda})$ являются необходимыми условиями экстремума при ограничениях задачи (4).

Поиск оптимальных значений множителей Лагранжа. Предположим, что $(\mathbf{z}', b', \boldsymbol{\lambda})$ является некоторой точкой, в которой соблюдаются условия стационарности и соблюдаются ограничения задачи (4).

Нетрудно показать, воспользовавшись уравнениями (5,6), что лагранжиан в точке $(\mathbf{z}', b', \boldsymbol{\lambda})$ может быть записан в виде

$$\mathbf{L}(\mathbf{z}', b', \boldsymbol{\lambda}) = g(\boldsymbol{\lambda}) = \sum_{j=1}^{m} \lambda_j - \tfrac{1}{2} \sum_{j'=1}^{m} \sum_{j''=1}^{m} \lambda_{j'} \lambda_{j''} \alpha_{j'} \alpha_{j''} (\mathbf{x}_{j'} \mathbf{x}_{j''}^t).$$

Отметим, что в силу соблюдения ограничений задачи (4) и неотрицательности множителей Лагранжа в точке $(\mathbf{z}', b', \boldsymbol{\lambda})$ выполняется неравенство

$$g(\boldsymbol{\lambda}) \le \tfrac{1}{2} \sum_{i=1}^{n} (z_i')^2 \le \tfrac{1}{2} \sum_{i=1}^{n} (z_i^*)^2$$

Из условий дополняющей нежёсткости следует, что в точке $(\mathbf{z}^*, b^*, \boldsymbol{\lambda}^*)$ справедливо равенство

$$\mathbf{L}(\mathbf{z}^*, b^*, \boldsymbol{\lambda}^*) = \tfrac{1}{2} \sum_{i=1}^{n} (z_i^*)^2 - \sum_{j=1}^{m} \lambda_j^* [\alpha_j (\mathbf{z}^* \mathbf{x}_j^t - b^*) - 1] = \tfrac{1}{2} \sum_{i=1}^{n} (z_i^*)^2$$

Таким образом максимум $g(\boldsymbol{\lambda})$ равен $\tfrac{1}{2} \sum_{i=1}^{n} (z_i^*)^2$ и достигается при $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$.

Таким образом, оптимальные значения неотрицательных множителей Лагранжа $(\lambda_1^*, \ldots, \lambda_m^*)$ могут быть найдены как решение оптимизационной задачи, которая называется квадратичного программирования, двойственной по отношению к задаче (4):

$$\sum_{j=1}^{m} \lambda_j - \tfrac{1}{2} \sum_{j'=1}^{m} \sum_{j''=1}^{m} \lambda_{j'} \lambda_{j''} \alpha_{j'} \alpha_{j''} (\mathbf{x}_{j'} \mathbf{x}_{j''}^t) \to \max \qquad (7)$$

$$\sum_{j=1}^{m} \lambda_j \alpha_j = 0$$

$$\lambda_j \ge 0,\ j = 1, \ldots, m$$

Пусть $(\hat{\lambda}_1,\ldots,\hat{\lambda}_m)$ - решение задачи (7) Направляющий вектор оптимальной разделяющей гиперплоскости находится по формуле $\mathbf{z}^* = \sum_{j=1}^{m} \hat{\lambda}_j \alpha_j \mathbf{x}_j$.

То есть направляющий вектор разделяющей гиперплоскости является линейной комбинацией векторных описаний объектов обучающей выборки, для которых значения соответствующих оптимальных множителей Лагранжа отличны от 0. Такие векторные описания принято называть опорными векторами. Пусть

$$J_0 = \{ j = 1,\ldots,m \,|\, [\alpha_j (\mathbf{z}^* \mathbf{x}_j^t - b) - 1] \neq 0 \}$$

Из условий дополняющей нежёсткости видно, при $j \in J_0$ обязательно должно выполняться равенство $\hat{\lambda}_j = 0$. Поэтому векторное описание $\mathbf{x}_j$ соответствующего объекта обучающей выборки является опорным вектором, если j не принадлежит J_0. Оценка параметра сдвига $\hat{b}$ находится из ограничения, соответствующего произвольному опорному вектору.

Распознавание новых объектов. Классификация нового распознаваемого объекта s с описанием $\mathbf{x}$ вычисляется согласно знаку выражения

$$g(\mathbf{x}) = \sum_{j=1}^{m} \hat{\lambda}_j \alpha_j (\mathbf{x}_j \mathbf{x}^t) - \hat{b}$$

Объект s относится к классу K_1, если $g(\mathbf{x}) > 0$ и объект s относится к классу K_2 в противном случае.

4.7.2 Случай отсутствия линейной разделимости

Существенным недостатком рассмотренного варианта метода опорных векторов является требование линейной разделимости классов. Однако данный недостаток может быть легко преодолён с помощью следующей модификации, основанной на использовании дополнительного вектора неотрицательных переменных $(\zeta_1,\ldots,\zeta_m)$.

Требования об отделимости классов из задачи (3) заменяются более мягкими требованиями:

$$\mathbf{z}\mathbf{x}_j^t \geq b + 1 - \zeta_j \quad \text{при} \quad s_j \in \tilde{S}_t \cap K_1$$

$$\mathbf{z}\mathbf{x}_j^t \le b - 1 + \zeta_j \quad \text{при} \quad s_j \in \tilde{S}_t \cap K_2, \quad j = 1,\ldots,m.$$

При этом выдвигается требование минимальности суммы $\sum_{j=1}^{m} \zeta_j$. Поиск оптимальных параметров разделяющей гиперплоскости при отсутствии линейной разделимости таким образом сводится к решению задачи квадратично программирования

$$\frac{1}{2}\sum_{i=1}^{n} z_i^2 + C\sum_{j=1}^{m} \zeta_j \to \min$$

$$\alpha_j(\mathbf{z}\mathbf{x}_j^t - b) \ge 1 - \zeta_j,\ \zeta_j \ge 0,\ ,\ j = 1,\ldots,m,$$

Положительная константа C является открытым параметром алгоритма. Иными словами оптимальное значение C подбирается пользователем.

Пусть $\boldsymbol{\lambda}^* = (\lambda_1^*,\ldots,\lambda_m^*)$ - вектор множителей Лагранжа, соответствующих ограничениям $\alpha_j(\mathbf{z}\mathbf{x}_j^t - b) \ge 1 - \zeta_j$;

$\boldsymbol{\eta}^* = (\eta_1^*,\ldots,\eta_m^*)$ - вектор множителей Лагранжа, соответствующих ограничениям $\zeta_j \ge 0,\ ,\ j = 1,\ldots,m$;

Из теоремы ККТ следует, что для точки $(\mathbf{z}^*, b^*, \boldsymbol{\zeta}^*)$, в которой функция $\frac{1}{2}\sum_{i=1}^{n} z_i^2 + C\sum_{j=1}^{m} \zeta_j$ достигает своего минимума при ограничениях задачи (4), и некоторых векторов значений неотрицательных множителей Лагранжа $\boldsymbol{\lambda}^*$ и $\boldsymbol{\eta}^*$ соблюдаются условия стационарности лагранжиана

$$\mathbf{L}(\mathbf{z},\boldsymbol{\zeta},b,\boldsymbol{\lambda},\boldsymbol{\eta}) = \tfrac{1}{2}\sum_{i=1}^{n} z_i^2 + C\sum_{j=1}^{m} \zeta_j - \sum_{j=1}^{m} \lambda_j[\alpha_j(\mathbf{z}\mathbf{x}_j^t - b) - 1 - \zeta_j] + \sum_{j=1}^{m} \eta_j\zeta_j$$

по переменным $(\mathbf{z}, \boldsymbol{\eta}, b)$.

Данные условия записываются в виде

$$\left.\frac{\mathbf{L}(\mathbf{z},\boldsymbol{\zeta},b,\boldsymbol{\lambda},\boldsymbol{\eta})}{\partial z_i}\right|_{(\mathbf{z}^*,b^*,\boldsymbol{\zeta}^*)} = z_i^* - \sum_{j=1}^{m} \lambda_j^*\alpha_j x_{ji} = 0, \quad i = 1,\ldots,n;$$

$$\left.\frac{\partial \mathbf{L}(\mathbf{z},b,\boldsymbol{\lambda})}{\partial b}\right|_{(\mathbf{z}^*,b^*,\boldsymbol{\lambda}^*)} = \sum_{j=1}^{m} \lambda_j^* \alpha_j = 0;$$

$$\left.\frac{\partial \mathbf{L}(\mathbf{z},b,\boldsymbol{\lambda})}{\partial \zeta_j}\right|_{(\mathbf{z}^*,b^*,\boldsymbol{\lambda}^*)} = C - \lambda_j - \eta_j = 0, \quad j = 1,\ldots,m$$

Также из теоремы ККТ следует необходимость выполнения m равенств, которые носят название условий дополняющей нежёсткости

$$\lambda_j^*[\alpha_j(\mathbf{z}^*\mathbf{x}_j^t - b) - 1 + \zeta_j] = 0, \eta_j\zeta_j, \ j = 1,\ldots,m., \quad j = 1,\ldots,m$$

Оптимальные значения множителей $(\lambda_1^*,\ldots,\lambda_m^*)$ могут быть найдены как решение двойственной задачи квадратичного программирования.

$$\sum_{j=1}^{m} \lambda_j - \tfrac{1}{2}\sum_{j'=1}^{m}\sum_{j''=1}^{m} \lambda_{j'}\lambda_{j''}\alpha_{j'}\alpha_{j''}(\mathbf{x}_{j'}\mathbf{x}_{j''}^t) \to \max \qquad (7)$$

$$\sum_{j=1}^{m} \lambda_j \alpha_j = 0$$

$$C \geq \lambda_j \geq 0, \ j = 1,\ldots,m$$

Как и в случае линейной разделимости направляющий вектор оптимальной разделяющей гиперплоскости находится по формуле $\mathbf{z}^* = \sum_{j=1}^{m} \lambda_j^* \alpha_j \mathbf{x}_j$. Из условий $C - \lambda_j - \eta_j = 0$ и $\zeta_j\eta_j = 0$ следует что $\eta_j > 0$ и $\zeta_j = 0$ при $0 < \lambda_j < C$.

Также как и в случае существования линейной разделимости параметра сдвига $\hat{b}$ находится из ограничения, соответствующего произвольному опорному вектору $\mathbf{x}_j$. Действительно, из условий дополняющей нежёсткости и и следующего из них равенства $\zeta_j = 0$ следует выполнение равенства $\alpha_j(\mathbf{z}^*\mathbf{x}_j^t - \hat{b}) - 1 = 0$, эквивалентного равенству $\hat{b} = \mathbf{z}^*\mathbf{x}_j^t - \alpha_j$.

Распознавание нового объекта s производится по его описанию $\mathbf{x}$ также как и в случае линейно разделимых классов с помощью решающего правила (8) по величине распознающей функции $g(\mathbf{x})$.

4.7.3 Построение оптимальных нелинейных разделяющих поверхностей с помощью метода опорных векторов.

Предположим что в исходном признаковом пространстве эффективное линейное разделение отсутствует. Однако может существовать такое евклидово пространство H_y и такое отображение Φ из области пространства $\mathbf{R}^n$ содержащей описания распознаваемых объектов, в пространство H_y, что образы объектов обучающей выборки из классов K_1 и K_2 оказываются разделимыми с помощью некоторой гиперплоскости P_y. Пусть $\{\mathbf{y}_1,\ldots,\mathbf{y}_m\}$ -образы в пространстве H_y векторов описаний объектов обучающей выборки $\{\mathbf{y}_1,\ldots,\mathbf{y}_m\}$.

Линейная разделимость означает существование решения аналога задачи квадратичного программирования (4) для пространства H_y, которое сводится к решению двойственной задачи

$$\sum_{j=1}^{m}\lambda_j - \tfrac{1}{2}\sum_{j'=1}^{m}\sum_{j''=1}^{m}\lambda_{j'}\lambda_{j''}\alpha_{j'}\alpha_{j''}(\mathbf{y}_{j'}\mathbf{y}_{j''}^t) \to \max$$

$$\sum_{j=1}^{m}\lambda_j\alpha_j = 0$$

$$\lambda_j \geq 0,\ j=1,\ldots,m$$

Отметим, что необходимость полного восстановления преобразования $\Phi(\mathbf{x})$ для поиска всех коэффициентов задачи квадратичного программирования (13) отсутствует. Достаточно найти функцию, связывающую скалярное произведение $(\mathbf{y}_{j'}\mathbf{y}_{j''}^t)$ с векторами $\mathbf{x}_{j'}$ и $\mathbf{x}_{j'}$, где $\mathbf{y}_{j'}=\Phi(\mathbf{x}_{j'})$ и $\mathbf{y}_{j''}=\Phi(\mathbf{x}_{j''})$.

Такую функцию мы далее будем называть потенциальной и обозначать $K(\mathbf{x}_{j'},\mathbf{x}_{j''})$. Можно подобрать потенциальную функцию таким образом, чтобы решение (13) было оптимальным. При этом поиск оптимальной потенциальной функции может производится внутри некоторого заранее заданного семейства. Например, потенциальную функцию можно задать с помощью простого сдвига $K(\mathbf{x}_{j'},\mathbf{x}_{j''})=\mathbf{x}_{j'}\mathbf{x}_{j''}^t+\theta$. Решение, полученное путём замены скалярных произведений на потенциальные функции, может

рассматриваться как построении линейной разделяющей поверхности в трансформированном пространстве, если удаётся доказать существование отображения $\Phi(\mathbf{x})$, для которого при произвольных $\mathbf{x}'$ и $\mathbf{x}''$ из $\mathbf{R}^n$ выполняется равенство

$$K(\mathbf{x}', \mathbf{x}'') = F(\mathbf{x}')F^t(\mathbf{x}'')$$

\Существование преобразования $\Phi(\mathbf{x})$, для которого выполняется равенство (15), было показано для неотрицательных симметричных потенциальных функций вида

$$K(\mathbf{x}', \mathbf{x}'') = [\mathbf{x}'(\mathbf{x}'')^t]^d + q,$$

где d -целое число, q -вещественная константа.

Существование преобразования $\Phi(\mathbf{x})$ с выполнением равенство (15) доказано также для ядровых функции типа гауссианы

$$K(\mathbf{x}', \mathbf{x}'') = \frac{1}{\sqrt{2p}d} e^{-\frac{(\mathbf{x}' - \mathbf{x}'')^2}{2d^2}},$$

где σ - вещественная неотрицательная константа (размер ядра). Поскольку в общем случае преобразование является нелинейным, то прообразом в пространстве $\mathbf{R}^n$ линейной разделяющей гиперплоскости, существующей в пространстве H_y, может оказаться нелинейная поверхность.

Для большого числа прикладных задач линейная разделимость является недостижимой. Поэтому выбор ядровой функции может производиться из требования о минимальности числа ошибок в смысле задачи квадратичного програмирования (9). На практике подбор ядровых функций и их параметров производится исходя из требования достижения максимальной обобщающей способности, которая оценивается с помощью скользящего контроля или оценок на контрольной выборке. Опыт решения прикладных задач показывает, что высокая эффективность распознавания достигается при выборе в качестве ядровой функции гауссианы.

Прототипом метода опорных векторов явился метод «Обобщенный портрет», разработанный В.Н.Вапником и А.Я.Червоненкисом [1] . В современном варианте метод был предложен в работе [21]. Подробное описание метода появилось в работе [18] в 1998 году. .В настоящее время метод опорных векторов является одним из наиболее распространённым в мире средством решения задач распознавания, высокая эффективность которого подтверждается практикой. В связи с этим были предложены подходы, использующие основные принципы метода опорных векторов для решения задач регрессионного анализа.

Литература

[1] Вапник В.Н., Червоненкис А.Я. Теория распознавания образов (статистические проблемы обучения). М.: Наука. 1974. - 416 с.

[3] Воронцов К.В. (Курс лекций). www.machinelearning.ru

[4] Докукин А.А., Сенько О.В. Оптимальные выпуклые корректирующие процедуры в задачах высокой размерности Журнал вычислительной математики и математической физики. 2011. Т. 51. № 9. С. 1751-1760.

[5] А.М. Дубров, В.С.Мхитарян, Л.И.Трошин Многомерные статистические методы: Учебник, - М.: Финансы и статистика, 2000, - 352с.

[6] Дмитриев А.Н., Журавлев Ю.И., Кренделев Ф.П., О математических принципах классификации предметов и явлений. Сб. "Дискретный анализ". Вып. 7. Новосибирск, ИМ СО АН СССР. 1966. С. 3-11.

[7] Донской В.И. Алгоритмические модели обучения классификации: обоснование, сравнение, выбор. –Симферополь, «ДИАЙПИ», 2014,-227 с.

[8] Дюкова Е.В. Алгоритмы распознавания типа “Кора”: сложность реализации и метрические свойства// Распознавание, классификация, прогноз (матем. методы и их применение). М.: Наука, 1989. Вып.2. С. 99-125.

[9] Журавлев Ю.И., Никифоров В.В. Алгоритмы распознавания, основанные на вычислении оценок // Кибернетика. 1971. №3. С. 1-11.

[10] Журавлев Ю.И., ИЗБРАННЫЕ НАУЧНЫЕ ТРУДЫ. - М.: Издательство Магистр, 1998. - 420 с.

[11] Журавлев Ю.И., Рязанов В.В., Сенько О.В. «Распознавание». Математические методы. Программная система. Практические применения. — М.: Фазис, 2006. 159 с.

[

[12] Кузнецов В.А., Сенько О.В., Кузнецова А.В. и др. Распознавание нечетких систем по методу статистически взвешенных синдромов и его применение для иммуногематологической характеристики нормы и хронической патологии. // Химическая физика. 1996. Т.15. N 1. С.81-100.

[13] Лбов Г.С., Старцева Н.С. Логические решающие правила и вопросы статистической устойчивости решений.Новосибирск: Изд-во Ин-та математики, 1999, 212 с.

[14] Мерков А.Б. Распознавание образов: Введение в методы статистического обучения. **.** ***M***.: Едиториал УРСС, 2011. — 256 с.

[15] А.С. Потапов. Распознавание образов и машинное восприятие. Общий подход на основе принципа минимальной длины описаний. –Спб.: Политехника, 2007, -548 с.

[16] Рязанов В.В. Логические закономерности в задачах распознавания (параметрический подход)//*Ж. вычисл. матем. и матем. физ., 2007, том 47, номер 10, страницы 1793–1808*

[17] L. Breiman Bagging predictors. Machine learning, 24, 123-140, 1996.

[18] Chris. J.C. Burges A Tutorial on Support Vector Machines for Pattern Recognition. Kluwer Academic Publishers, Boston. Manufactured in The Netherlands. Appeared in: Data Mining and Knowledge Discovery 2, 121-167, 1998.

[19] T. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting and randomization. Machine Learning, 40(2):139–157, 2000.\

[20] Kuncheva L.I. Combining Pattern Classifiers. Methods and Algorithms. — Wiley Interscience, New Jersey, 2004.

[21] Cortes, C.; Vapnik, V. (1995). "Support-vector networks". //*Machine Learning* **20** (3): 273

[22] Senko O.B., Kuznetsova A.B. A recognition method based on collective decision making using systems of regularities of various types. Pattern Recognition and Image Analysis. 2010. V. 20. № 2. P. 152–162.