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.

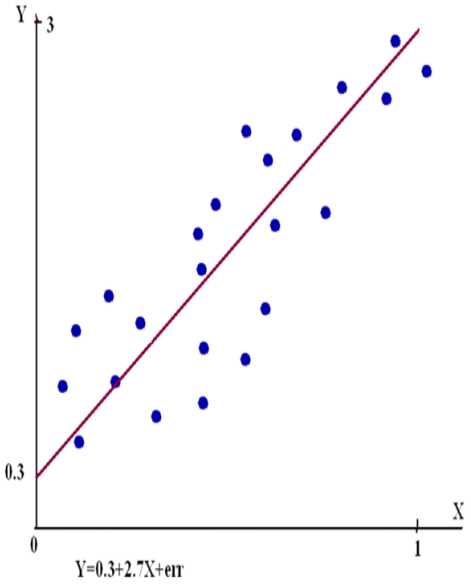

Рис. 1.1

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 .

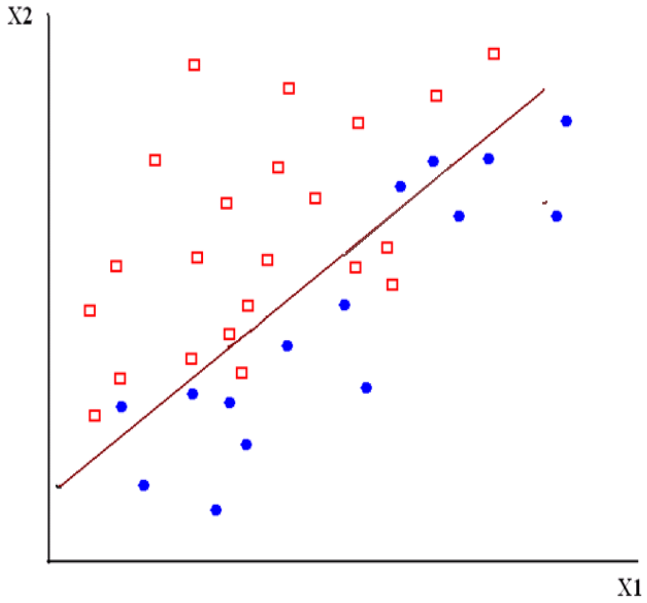

Рис. 1.2

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.

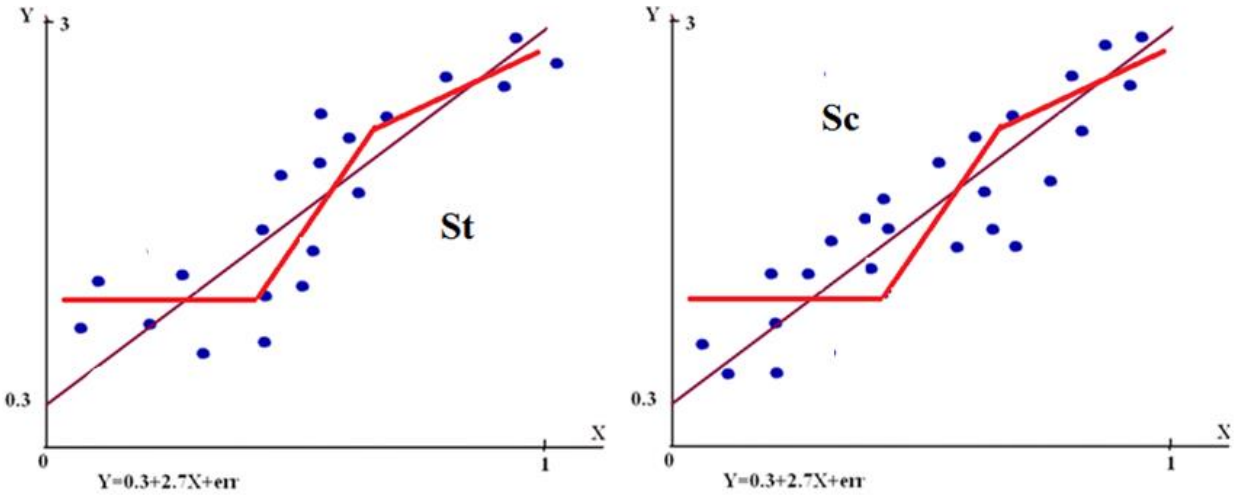

Рис. 1.4

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.

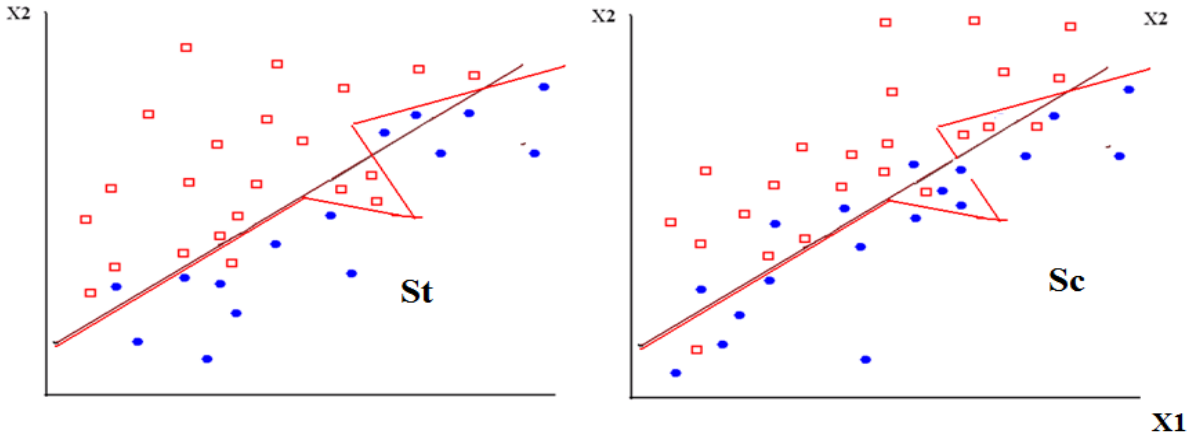

Рис. 1.5

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

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

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

$Cov(Y, X \mid \tilde{S}_t) = \frac{1}{m}\sum_{j=1}^{m}(y_j - \bar{y})(x_j - \bar{x})$ является выборочной ковариацией переменных Y и X , $D(X \mid \tilde{S}_t) = \frac{1}{m}\sum_{j=1}^{m}(x_j - \bar{x})^2$ - 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 $\mathbf{X}_s=\begin{pmatrix}\hat{x}_{11}^s.\ldots\ldots\hat{x}_{1n}^s\\ \ldots\ldots\ldots\ldots\ldots\\ \hat{x}_{j1}\ldots\ldots\hat{x}_{jn}\\ \ldots\ldots\ldots\ldots\ldots\\ \hat{x}_{m1}\ldots\ldots\hat{x}_{mn}\end{pmatrix}$ - 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

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

……

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

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

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

.

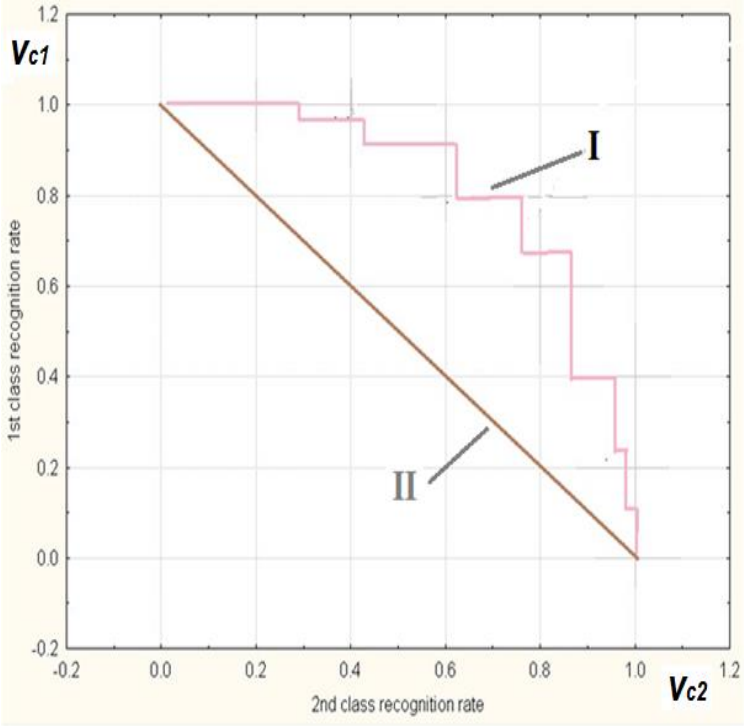

Рис 3.1. Ломаная (I) соединяет точки на двумерной плоскости в декартовой системе координат, которые являются соседними в ряду (1.1) .

Однако сохранение высокого значения v_{c1} при высоких значениях v_{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.

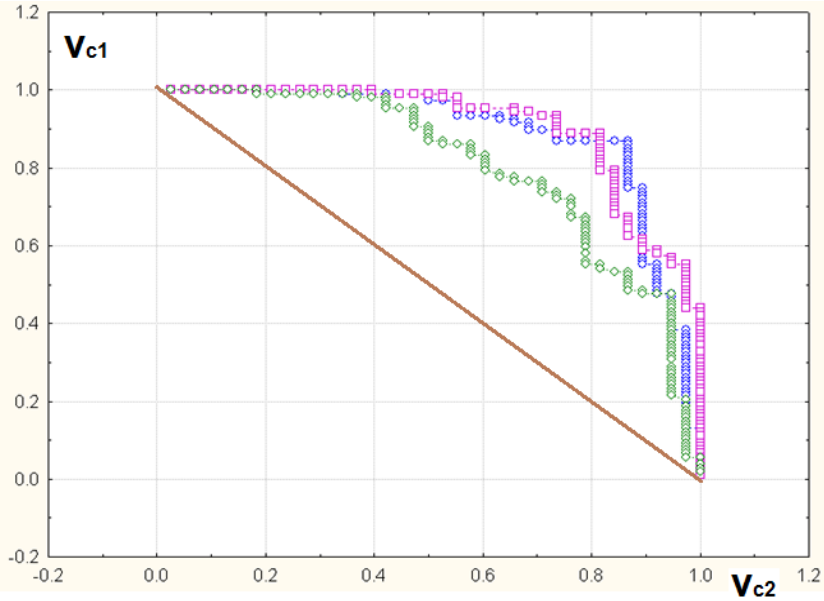

Рис. 3.2 Сравнение трёх метод распознавания с помощью

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.

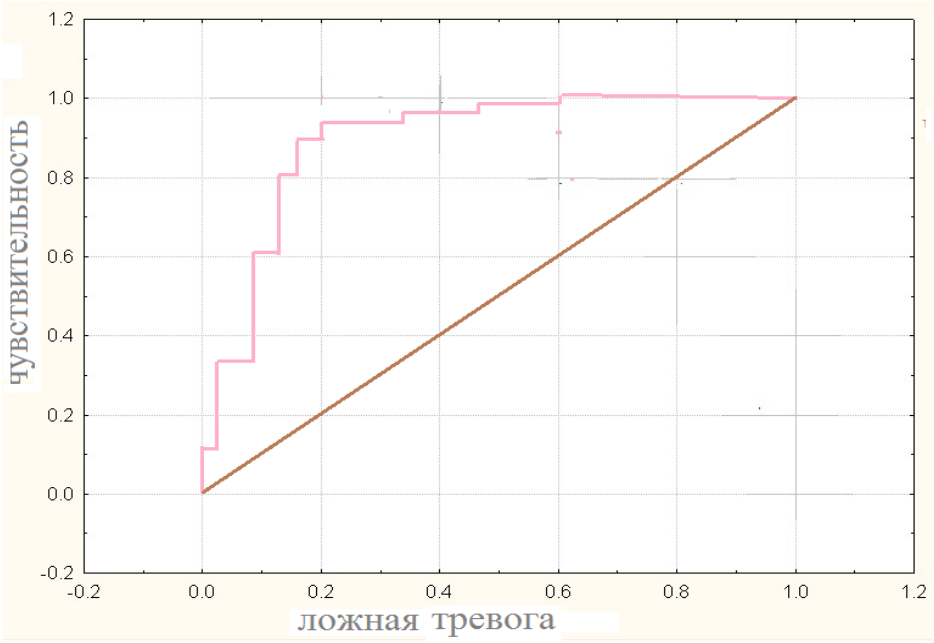

Рис. 3.3 Вид ROC кривой в координатах чувствительность (ось Y) и ложная тревога (ось X)

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

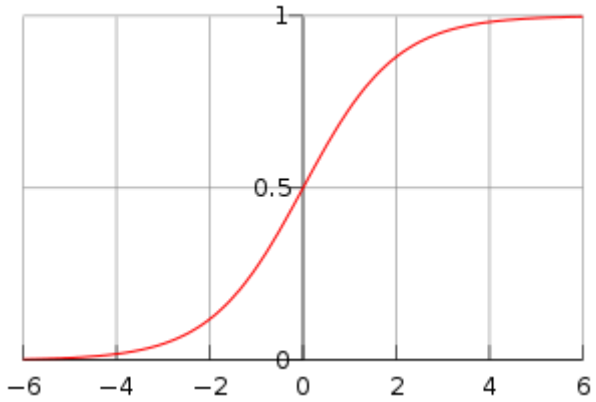

Рис.

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].