

МАТЕМАТИЧЕСКИЕ ОСНОВЫ ТЕОРИИ ПРОГНОЗИРОВАНИЯ

Лектор

Сенько Олег Валентинович

Лекция 4

Байесовское обучение

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

Байесовские методы обучения основаны на аппроксимации условных вероятностей классов в точках признакового пространства с использованием формулы Байеса.

Байесовское обучение

Рассмотрим задачу распознавания классов K_1, \dots, K_L

Формула Байеса позволяет рассчитать условные вероятности классов в точке признакового пространства:

$$\mathbf{P}(K_i | \mathbf{x}) = \frac{f_i(\mathbf{x})\mathbf{P}(K_i)}{\sum_{j=1}^L f_j(\mathbf{x})\mathbf{P}(K_j)}, \text{ где } f_i(\mathbf{x}) \text{ - плотность}$$

распределения вероятности для класса K_i ;

$\mathbf{P}(K_i)$ - вероятность класса K_i безотносительно к признаковым описаниям (априорная вероятность).

Байесовское обучение

При этом в качестве оценок априорных вероятностей

$\mathbf{P}(K_1), \dots, \mathbf{P}(K_L)$ могут быть взяты доли объектов соответствующих классов в обучающей выборке.

Плотности вероятностей $f_1(\mathbf{x}), \dots, f_L(\mathbf{x})$ восстанавливаются исходя из предположения об их принадлежности фиксированному типу распределения.

Чаще всего используется многомерное нормальное распределения.

Байесовское обучение

Плотность данного распределения в общем виде представляется выражением

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

где $\boldsymbol{\mu}$ - математическое ожидание вектора признаков \mathbf{x} ;

Σ - матрица ковариаций признаков X_1, \dots, X_n ;

$|\Sigma|$ - детерминант матрицы Σ .

Байесовское обучение

Для построения распознающего алгоритма достаточно

оценить вектора математических ожиданий $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_L$ и

матрицы ковариаций $\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_L$ для классов K_1, \dots, K_L

соответственно.

Оценка $\boldsymbol{\mu}_i$ вычисляется как среднее значение векторов

признаков по объектам обучающей выборки из класса K_i :

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{m_i} \sum_{s_j \in \tilde{S}_t \cap K_i} \mathbf{x}_j, \text{ где } m_i - \text{число объектов класса } K_i$$

в обучающей выборке.

Байесовское обучение

Элемент матрицы ковариаций для класса K_i вычисляется по формуле

$$\sigma_{kk'}^i = \frac{1}{m_i} \sum_{s_j \in \tilde{S}_i \cap K_i} (x_{jk} - \mu_k^i)(x_{jk'} - \mu_{k'}^i), \quad k, k' \in \{1, \dots, n\}, \quad \text{где}$$

μ_k^i - k -ая компонента вектора $\boldsymbol{\mu}^i$. Матрицу ковариации, состоящую из элементов $\sigma_{kk'}^i$ обозначим $\hat{\Sigma}_i$

Очевидно, что согласно формуле Байеса максимум $\mathbf{P}(K_i | \mathbf{x})$

достигается для тех же самых классов для которых максимально

произведение $f_i(\mathbf{x})\mathbf{P}(K_i)$.

Байесовское обучение

Очевидно, что для байесовской классификации может

использоваться также натуральный логарифм $\ln[f_i(\mathbf{x})\mathbf{P}(K_i)]$

который согласно вышеизложенному может быть оценён

выражением $g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}\hat{\Sigma}_i\mathbf{x}^t) + \mathbf{w}_i\mathbf{x}^t + g_i^0$, где

$$\mathbf{w}_i = \hat{\boldsymbol{\mu}}_i\hat{\Sigma}_i^{-1}$$

$$g_i^0 = -\frac{1}{2}(\hat{\boldsymbol{\mu}}_i\hat{\Sigma}_i^{-1}\hat{\boldsymbol{\mu}}_i^t) - \frac{1}{2}\ln(|\hat{\Sigma}_i|) + \ln(v_i) - \frac{n}{2}\ln(2\pi) \quad - \text{не}$$

зависящее от \mathbf{x} слагаемое;

v_i - доля объектов класса K_i в обучающей выборке.

Байесовское обучение

Таким образом объект с признаковым описанием \mathbf{x} будет отнесён построенной выше аппроксимацией байесовского классификатора к классу, для которого оценка $g_i(\mathbf{x})$ 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.

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 $\frac{(\mathbf{w}\mathbf{x}^t)}{|\mathbf{w}|}$.

Линейный дискриминант Фишера

В качестве меры различий проекций классов на \mathbf{w} используется функционал

- $$\Phi(\mathbf{w}) = \frac{[\hat{X}_{w1}(\mathbf{w}) - \hat{X}_{w2}(\mathbf{w})]^2}{\hat{d}_1(\mathbf{w}) + \hat{d}_2(\mathbf{w})}, \text{ где}$$
- $$\hat{X}_{wi}(\mathbf{w}) = \frac{1}{m_i} \sum_{s_j \in \tilde{S}_t \cap K_i} \frac{(\mathbf{w} \mathbf{x}_j^t)}{|\mathbf{w}|} - \text{среднее значение проекции}$$

векторов, описывающих объекты из класса $K_i, i \in \{1, 2\}$

Линейный дискриминант Фишера

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

дисперсия проекций векторов, описывающих объекты из класса.

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.

Линейный дискриминант Фишера

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

$\mathbf{w}^t = \hat{\Sigma}_{12}^{-1}(\hat{\boldsymbol{\mu}}_1^t - \hat{\boldsymbol{\mu}}_2^t)$, 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 $\hat{\mathbf{w}}^t = \hat{\Sigma}_{12}^{-1}(\hat{\boldsymbol{\mu}}_1^t - \hat{\boldsymbol{\mu}}_2^t)$ ($\hat{\boldsymbol{\mu}}_i = \frac{1}{m_i} \sum_{s_j \in \tilde{S}_i \cap K_i} \mathbf{x}_j$)

Распознавание нового объекта s_* по признаковому описанию

\mathbf{x}_* производится по величине проекции $\gamma(\mathbf{x}_*) = \frac{(\mathbf{w}\mathbf{x}_*)}{|\mathbf{w}|}$

помощью простого порогового правила

При $\gamma(\mathbf{x}_*) \geq b$ объект s_* относится к классу K_1 и s_* относится к классу K_2 в противном случае.

Линейный дискриминант Фишера

Граничный параметр b 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 b средней проекции объектов обучающей выборки на \mathbf{w} .

Линейный дискриминант Фишера

Метод ЛДФ легко обобщается на случай с несколькими классами.

При этом исходная задача распознавания классов K_1, \dots, K_L сводится к последовательности задач с двумя классами K'_1 и K'_2 :

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.

Линейный дискриминант Фишера

В результате получается набор из L направлений $\mathbf{w}_1, \dots, \mathbf{w}_L$

При распознавании нового объекта s_* по признаковому описанию \mathbf{x}_* вычисляются проекции на $\mathbf{w}_1, \dots, \mathbf{w}_L$

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

Распознаваемый объект относится к тому классу,

соответствующему максимальной величине проекции.

Распознавание может производиться также по величинам

$$[\gamma_1(\mathbf{x}_*) - b_1], \dots, [\gamma_L(\mathbf{x}_*) - b_L]$$

Логистическая регрессия

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

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



Рис. 1

Логистическая регрессия

В методе логистическая регрессия связь условной вероятности класса K с прогностическими признаками осуществляются через переменную z , которая задаётся как линейная комбинация признаков:

$$z = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$$

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

$$\mathbf{P}(K | \mathbf{x}) = \frac{1}{e^{-\beta_0 - \beta_1 x_{*1} - \dots - \beta_n x_{*n}} + 1} = \frac{e^{\beta_0 + \beta_1 x_{*1} + \dots + \beta_n x_{*n}}}{e^{\beta_0 + \beta_1 x_{*1} + \dots + \beta_n x_{*n}} + 1} \quad (1)$$

Логистическая регрессия

Оценки регрессионных параметров $\beta_0, \beta_1, \dots, \beta_n$ могут быть вычислены по обучающей выборке с помощью различных вариантов метода максимального правдоподобия. Предположим, что объекты обучающей выборки сосредоточены в точках признакового пространства $\{\mathbf{x}_1, \dots, \mathbf{x}_r\}$. При этом распределение объектов обучающей выборка по точкам задаётся с помощью набора пар $\{(m_1, k_1), \dots, (m_r, k_r)\}$, где m_i - общее число объектов в точке \mathbf{x}_i , k_i - число объектов класса K в точке \mathbf{x}_i .

Логистическая регрессия

Вероятность данной конфигурации подчиняется распределению

Бернулли. Введём обозначение $p(\mathbf{x}) = \mathbf{P}(K | \mathbf{x})$. Функция

правдоподобия может быть записана в виде

$$L(\boldsymbol{\beta}, \mathbf{x}) = \prod_{i=1}^r C_{m_i}^{k_i} [p(\mathbf{x})]^{k_i} [1 - p(\mathbf{x})]^{m_i - k_i}$$

$$\frac{p(\mathbf{x})}{1 - p(\mathbf{x})} = e^{\beta_0 + \beta_1 x_{*1} + \dots + \beta_1 x_{*n}}$$

Принимая во внимание что

$$1 - p(\mathbf{x}) = \frac{1}{1 + e^{\beta_0 + \beta_1 x_{*1} + \dots + \beta_1 x_{*n}}}$$

Логистическая регрессия

- Получаем

$$L(\boldsymbol{\beta}, \mathbf{x}) = \prod_{i=1}^r C_{n_i}^{k_i} e^{k_i(\beta_0 + \beta_1 x_{*1} + \dots + \beta_1 x_{*n})} \left(\frac{1}{1 + e^{\beta_0 + \beta_1 x_{*1} + \dots + \beta_1 x_{*n}}} \right)^{m_i}$$

$$\begin{aligned} \ln L(\boldsymbol{\beta}, \mathbf{x}) &= \sum_{i=1}^r \ln(C_{n_i}^{k_i}) + \sum_{i=1}^r k_i(\beta_0 + \beta_1 x_{*1} + \dots + \beta_1 x_{*n}) + \\ &+ \sum_{i=1}^r m_i \ln \{ 1 / [1 + e^{(\beta_0 + \beta_1 x_{*1} + \dots + \beta_1 x_{*n})}] \} \end{aligned}$$

Метод k-ближайших соседей

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

Метод k-ближайших соседей

Окрестность V_k задаётся с помощью функции расстояния

$\rho(\mathbf{x}', \mathbf{x}'')$ заданной на декартовом произведении $\tilde{\mathbf{X}} \times \tilde{\mathbf{X}}$,

где $\tilde{\mathbf{X}}$ - область допустимых значений признаков

описаний. В качестве функции расстояния может быть

использована стандартная евклидова метрика

$$\rho_e(\mathbf{x}', \mathbf{x}'') = \sqrt{\sum_{i=1}^n \frac{1}{n} (x'_i - x''_i)^2}$$

Метод k -ближайших соседей

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 k векторных описаний, ближайших в смысле выбранной функции расстояний, к описанию \mathbf{x}^* распознаваемого объекта s^* .

Метод k-ближайших соседей

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.

Распознавание при заданной точности распознавания некоторых классов

- Байесовский классификатор обеспечивает максимальную общую точность распознавания. Однако при решении конкретных практических задач потери, связанные с неправильной классификацией объектов, принадлежащих к одному из классов, значительно превышают потери, связанные с неправильной классификацией объектов других классов. Для оптимизации потерь необходимо использование методов распознавания с учётом предпочтительной точности распознавания для некоторых классов.

Распознавание при заданной точности распознавания некоторых классов

Одним из возможных подходов является фиксирование порога для точности распознавания одного из классов.

Оптимальное решающее правило в задаче распознавания с двумя классами K_1 и K_2 , обеспечивающее максимальную точность распознавания K_2 при фиксированной точности распознавания K_1 , описывается критерием Неймана-Пирсона.

Распознавание при заданной точности для некоторых классов

Критерий Неймана-Пирсона

Максимальная точность распознавания K_2 при точности распознавания K_1 равной α обеспечивается

правилом: Объект с описанием \mathbf{x} относится в класс K_1 ,

если $\mathbf{P}(K_1 | \mathbf{x}) \geq \eta \mathbf{P}(K_2 | \mathbf{x})$, где параметр η

определяется из условия $\int_{\Theta} \mathbf{P}(K_1 | \mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \alpha$, а

$$\Theta = \{\mathbf{x} | \mathbf{P}(K_1 | \mathbf{x}) \geq \eta \mathbf{P}(K_2 | \mathbf{x})\}$$

$f(\mathbf{x})$ - плотность распределения $K_2 \cup K_1$ в точке \mathbf{x}

Распознавание при заданной точности для отдельных классов

Критерий Неймана-Пирсона может быть использован, если известны плотности распределения распознаваемых классов. Плотности могут быть восстановлены в рамках Байесовских методов обучения на основе гипотез о виде распределений.

Однако существуют эффективные средства регулирования точности распознавания при предпочтительности одного из классов, которые не требуют гипотез о виде распределения.

Распознавание при заданной точности для отдельных классов

Данные средства основаны на структуре распознающего алгоритма.

Каждый алгоритма распознавания классов

K_1, \dots, K_L 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 за классы K_1, \dots, K_L :

$$\{\gamma(s)_1, \dots, \gamma(s)_L\}$$

Распознавание при заданной точности для отдельных классов

Решающее правило C производит отнесение объекта s по вектору оценок $\{\gamma_1(s), \dots, \gamma_L(s)\}$ 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 к классу K_2 в противном случае.

Распознавание при заданной точности для отдельных классов

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 $C(\delta)$

ROC анализ

$C(\delta)$: 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 $C(\delta)$ 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 $p_{c2}(\delta, s_*)$.

ROC анализ

Зависимость между $p_{c_2}(\delta, s_*)$ и $p_{c_1}(\delta, s_*)$ может быть приближённо восстановлена по обучающей выборке \tilde{S}_t , включающей описания объектов $\{s_1, \dots, s_m\}$

Пусть $\begin{pmatrix} \gamma_1(s_1) \dots \gamma_1(s_m) \\ \gamma_2(s_1) \dots \gamma_2(s_m) \end{pmatrix}$ - матрица оценок за классы объектов \tilde{S}_t ,

$$[\gamma(s_1) = \gamma_1(s_1) - \gamma_2(s_1), \dots, \gamma(s_m) = \gamma_1(s_m) - \gamma_2(s_m)]$$

ROC анализ

Предположим, что величины $[\gamma(s_1), \dots, \gamma(s_m)]$ принимают r значений $(\Gamma_1, \dots, \Gamma_r)$. Рассмотрим r пороговых решающих правил $[C(\Gamma_1), \dots, C(\Gamma_r)]$. Для каждого из правил $C(\Gamma_i)$ вычислим $\nu_{c1}(\Gamma_i)$ - долю K_1 среди объектов обучающей выборки, удовлетворяющих условию $\gamma(s_*) \geq \Gamma_i$, и $\nu_{c2}(\Gamma_i)$ - долю K_2 среди объектов обучающей выборки, удовлетворяющих условию $\gamma(s_*) < \Gamma_i$. Отообразим результаты расчётов $\{[\nu_{c1}(\Gamma_1), \nu_{c2}(\Gamma_1)], \dots, [\nu_{c1}(\Gamma_r), \nu_{c2}(\Gamma_r)]\}$ как точки на в декартовой системе координат.

ROC анализ

Соединив точки отрезками прямых, получим ломаную линию (I), соединяющую точки (1,0) и (0,1). 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 v_{c1} и v_{c2} наиболее полно оценивает эффективность распознающего оператора R . Отметим, что v_{c2} постепенно убывает по мере роста v_{c1} .

ROC анализ

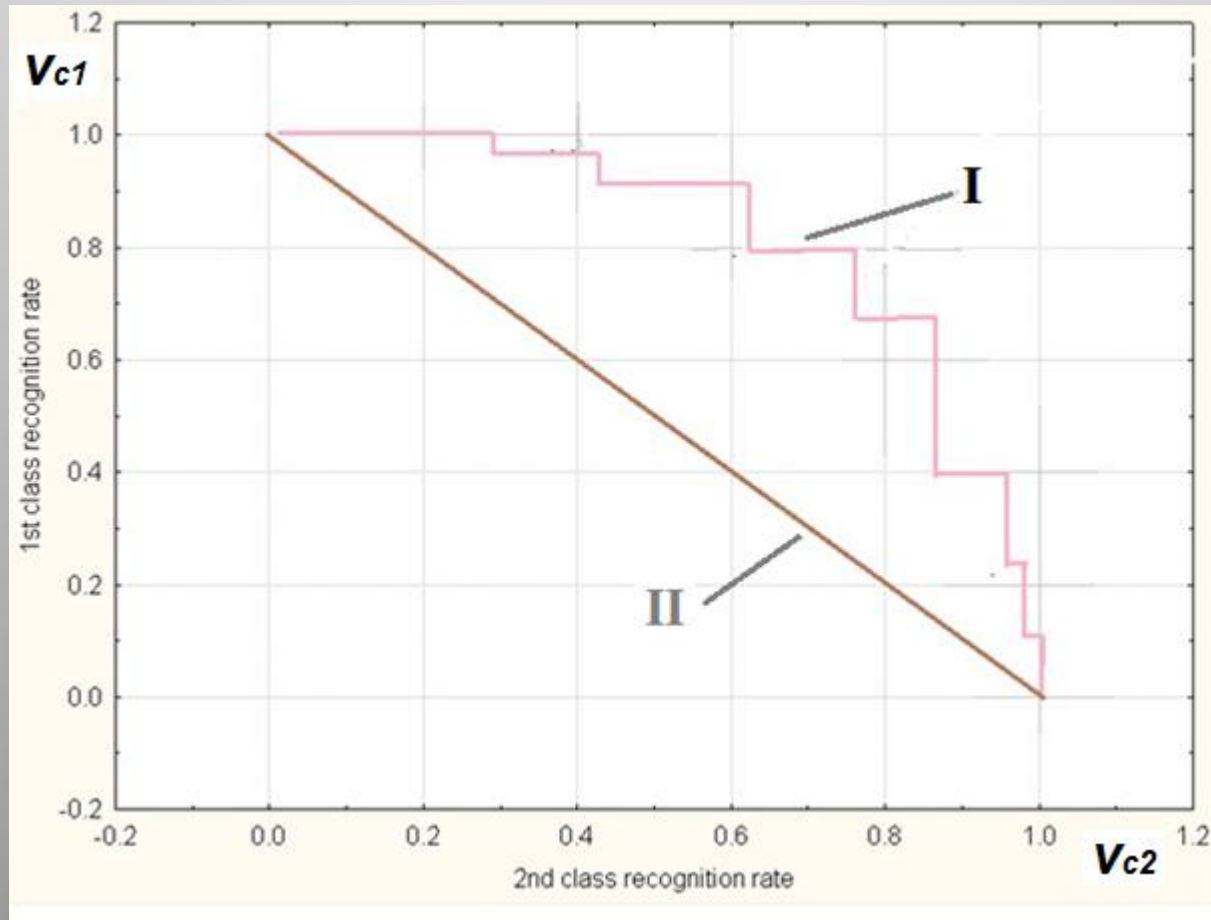


Рис. 2

ROC анализ

Однако сохранение высокого значения V_{c2} при высоких значениях V_{c1} соответствует существованию решающего правила, при котором точность распознавания обоих классов высока. Таким образом эффективному распознающему оператору соответствует близость линии I к прямой, связывающей точки (0,1) и (1,1). То есть наиболее высокой эффективности соответствует максимально большая площадь под линией I. Отсутствию распознающей способности соответствует близость к прямой II, связывающей точки (0, 1) и (1,0)

ROC анализ

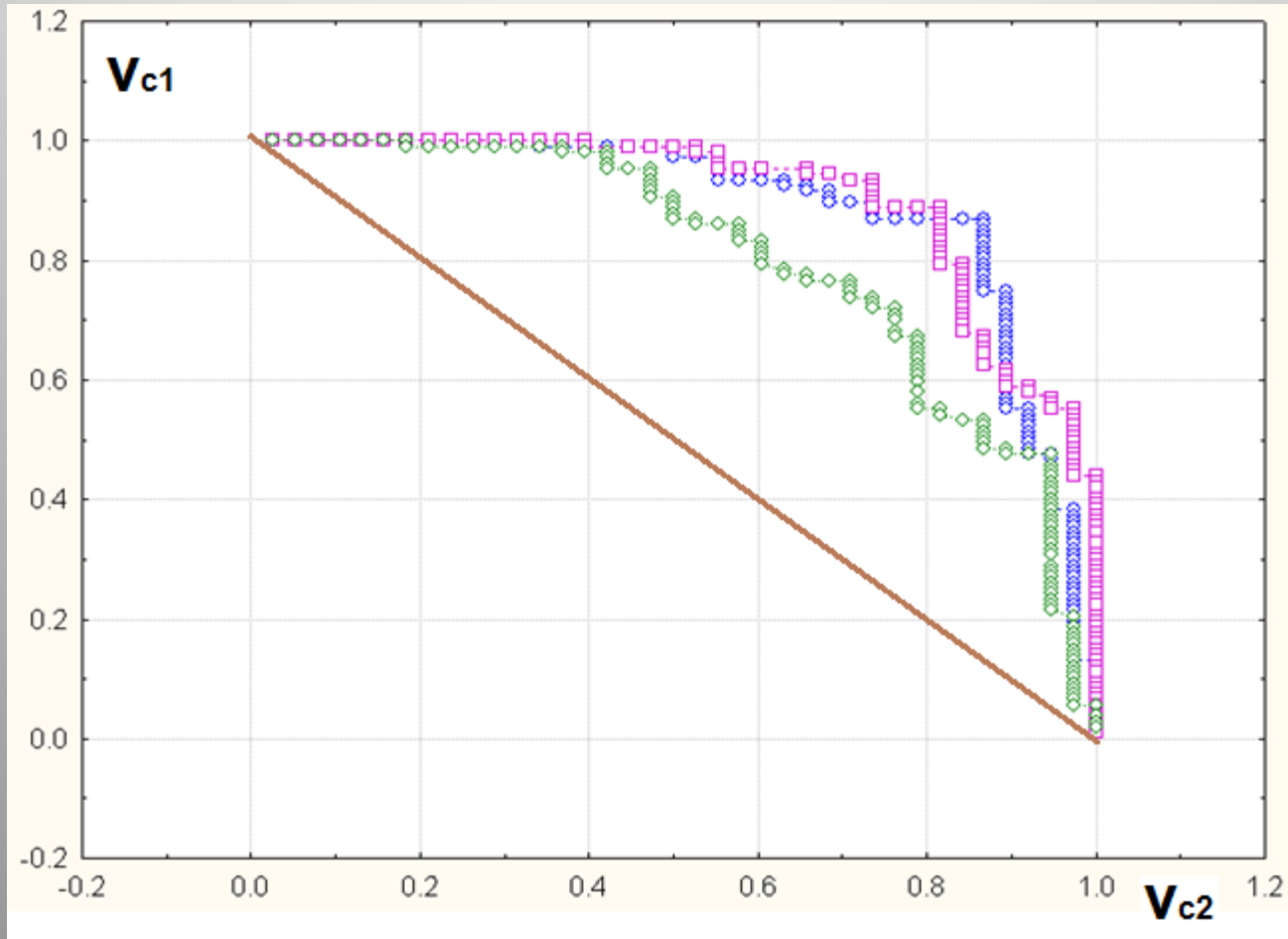


Рис. 3

ROC анализ

- На рисунке 3 сравниваются линии, характеризующие эффективность распознающих операторов, принадлежащих к трём методам распознавания, при решении задачи распознавания двух видов аутизма по психометрическим показателям .



Линейный дискриминант Фишера;



Метод опорных векторов



Статистически взвешенные синдромы

ROC анализ

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.

ROC анализ

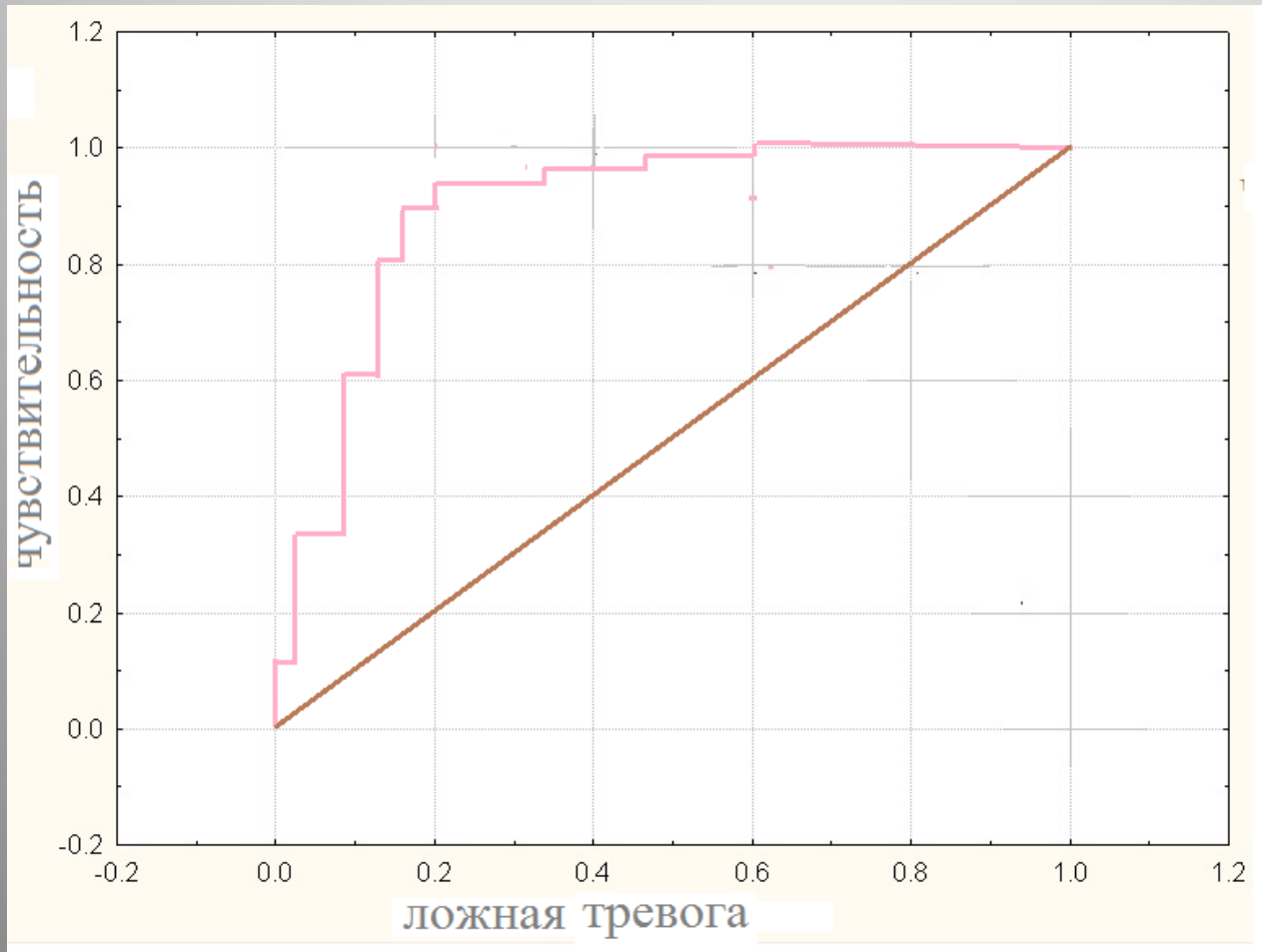


Рис. 4

ROC анализ

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

Линии, связывающих параметры «Чувствительность» и «Ложная тревога» принято называть **ROC-кривыми.**