

# Логистическая регрессия. Дискриминантный анализ. Отбор признаков.

Владимир Агеев, 622 гр.

3 октября, 2017г.

# Задача классификации

Дано:

- Множество объектов  $X \in \mathbb{R}^p$ ;
- Множество ответов  $Y, Y_i \in \mathcal{G}$ .

Задача:

- По выборке  $X^N = (\mathbf{x}_i, y_i)_{i=1}^N$  построить классификатор  $a : \mathbb{R}^p \rightarrow \mathcal{G}$ , который по новому вектору признаков  $X$  предскажет отметку класса, т.е.  $a(\mathbf{x}) \in \mathcal{G}$ ;
- На  $a(\mathbf{x})$  должна достигаться минимальная ошибка классификации в некотором смысле;
- Хотим знать вероятности принадлежности объекта к тому или иному классу.

# Задача классификации

На генеральном языке:

- $\xi \in \mathbb{R}^p$  – случайный вектор признаков;
- $\eta \in \mathcal{G} = \{G_k\}_{k=1}^K$  – дискретная случайная величина, класс;
- $P(\xi, \eta)$  – их совместное распределение.

Дано: выборка  $(x_i, y_i)_{i=1}^N$  –  $N$  реализаций случайных векторов  $(\xi, \eta)$ .

По выборке необходимо построить классифицирующую функцию

$$a : \mathbb{R}^p \rightarrow \mathcal{G}.$$

Мера ошибки предсказания (функция потерь):

$$\mathbf{L} = (\lambda_{ij})_{i,j=1}^K, \quad K = \text{card}(\mathcal{G}), \quad \lambda_{ij} \text{ – цена ошибки.}$$

Средний риск:  $R(a) = \mathbb{E}(\mathbf{L}(\eta, a(\xi)))$ .

Задача:  $R(a) \rightarrow \min_a$ .

# Байесовский классификатор

Средний риск:

$$R(a) = \mathbb{E}(\mathbf{L}(\eta, a(\xi))) = \mathbb{E}_{\xi} \sum_{k=1}^K \mathbf{L}(G_k, a(\xi)) P(G_k | \xi).$$

Функция классификации:

$$a(\mathbf{x}) = \arg \min_{G \in \mathcal{G}} \sum_{k=1}^K \mathbf{L}(G_k, G) P(G_k | \xi = \mathbf{x}).$$

Подставим сюда 0-1 функцию потерь

$$a(\mathbf{x}) = \arg \min_{G \in \mathcal{G}} (1 - P(G | \xi = \mathbf{x})).$$

Равносильно

$$a(\mathbf{x}) = \arg \max_{G \in \mathcal{G}} P(G | \xi = \mathbf{x}) = \arg \max_{G \in \mathcal{G}} P(G) P(\xi | \eta = G).$$

Это решение называется байесовским классификатором.

# Дискриминантный анализ

Для построения байесовского классификатора, необходимо знать апостериорные вероятности  $P(G \mid \xi = \mathbf{x})$ . Обозначим

- $p_k(\mathbf{x}) = P(\xi = \mathbf{x} \mid \eta = G_k)$  условные плотности классов;
- $\pi_k = P(\eta = G_k)$  – априорные вероятности,  $\sum_{k=1}^K \pi_k = 1$ .

По теореме Байеса

$$P(\eta = G_k \mid \xi = \mathbf{x}) = \frac{p_k(\mathbf{x})\pi_k}{\sum_{i=1}^K p_i(\mathbf{x})\pi_i}.$$

Возникает вопрос: откуда брать априорные вероятности?

- Брать равновероятные:  $\pi_i = \frac{1}{K}$ ;
- Брать пропорционально объемам классов  $\pi_i = \frac{n_i}{N}$ ;
- Соответственно имеющейся информации.

# Дискриминантный анализ

## QDA

Предположим, что  $P(\xi \mid \eta = G_k) = \mathcal{N}_p(\mu_k, \Sigma_k)$ , его плотность

$$p_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} e^{-\frac{1}{2}(\mathbf{x} - \mu_k)^T \Sigma_k^{-1} (\mathbf{x} - \mu_k)}.$$

Подставим плотности в байесовский классификатор

$$\begin{aligned} a(\mathbf{x}) &= \arg \max_{i \in 1 \dots K} \pi_i p_i(\mathbf{x}) = \arg \max_{i \in 1 \dots K} \log(\pi_i p_i(\mathbf{x})) = \\ &= \arg \max_{i \in 1 \dots K} \left( -\frac{1}{2}(\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i) - \frac{1}{2} \log(|\Sigma_i|) + \log(\pi_i) \right) = \\ &= \arg \max_{i \in 1 \dots K} g_i(\mathbf{x}). \end{aligned}$$

Заметим, что получившийся классификатор квадратично зависит от  $\mathbf{x}$ , отсюда название quadratic discriminant analysis.

# Дискриминантный анализ

## LDA

Пусть распределения классов  $P(\xi \mid \eta = G_k) = \mathcal{N}_p(\mu_k, \Sigma)$ .

Упростим классификатор:

$$\begin{aligned} a(\mathbf{x}) &= \arg \max_{i \in 1 \dots K} \left( -\frac{1}{2} (\mathbf{x} - \mu_i)^T \Sigma^{-1} (\mathbf{x} - \mu_i) - \frac{1}{2} \log(|\Sigma|) + \log(\pi_i) \right) = \\ &= \arg \max_{i \in 1 \dots K} \left( -\frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \mu_i^T \Sigma^{-1} \mathbf{x} + \log(\pi_i) \right) = \arg \max_{i \in 1 \dots K} \delta_i(\mathbf{x}). \end{aligned}$$

Зависимость от  $\mathbf{x}$  линейна.

Разделяющая два класса гиперплоскость:

$$\begin{aligned} \{\mathbf{x} : \delta_i(\mathbf{x}) = \delta_j(\mathbf{x})\} &= \\ = \{\mathbf{x} : -\frac{1}{2} (\mu_i - \mu_j)^T \Sigma^{-1} (\mu_i + \mu_j) + (\mu_i - \mu_j)^T \Sigma^{-1} \mathbf{x} + \log(\pi_i/\pi_j) = 0\}. \end{aligned}$$

От соотношения между априорными вероятностями зависит положение границы относительно классов (к какому она ближе).

# Дискриминантный анализ

## Оценка параметров

ОМП параметров нормальных плотностей:

- Среднее

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j: y_j = G_i} \mathbf{x}_j;$$

- Ковариационная матрица класса

$$\hat{\Sigma}_i = \frac{1}{n_i - 1} \sum_{j: y_j = G_i} (\mathbf{x}_j - \hat{\mu}_i)^T (\mathbf{x}_j - \hat{\mu}_i);$$

- Pooled ковариационная матрица

$$\hat{\Sigma} = \sum_{j=1}^K \frac{n_j - 1}{N - K} \hat{\Sigma}_j.$$



# Дискриминантный анализ

## Возвращение к вероятностям

Если нам необходимо получить вероятности отношения  $\mathbf{x}$  к классу  $i$ , то, вычислив  $\hat{\delta}_i(\mathbf{x})$ , можно вычислить

$$\hat{P}(\eta = G_i \mid \boldsymbol{\xi} = \mathbf{x}) = \frac{e^{\hat{\delta}_i(\mathbf{x})}}{\sum_{j=1}^K e^{\hat{\delta}_j(\mathbf{x})}},$$

подставив оценки  $\widehat{\pi_i p_i(\mathbf{x})}$  в формулу байеса для  $P(\eta = G_k \mid \boldsymbol{\xi} = \mathbf{x})$ . Далее относим  $\mathbf{x}$  к тому классу, для которого оценка максимальна.

# Дискриминантный анализ

## Regularized Discriminant Analysis

Компромис между LDA и QDA.

- Regularized Discriminant Analysis. Рассматривается матрица

$$\hat{\Sigma}_i(\alpha) = \alpha \hat{\Sigma}_i + (1 - \alpha) \hat{\Sigma},$$

где  $\hat{\Sigma}$  – pooled ковариационная матрица. Здесь  $\alpha \in [0, 1]$  порождает континуум моделей между LDA и QDA, выбирается скользящим контролем.

- Модифицируем pooled ковариационную матрицу и рассмотрим

$$\hat{\Sigma}(\gamma) = \gamma \hat{\Sigma} + (1 - \gamma) \sigma^2 \mathbf{I}_p,$$

где  $\gamma$  определяет вид ковариационной матрицы и выбирается скользящим контролем.

Пришли к выбору ковариационных матриц  $\hat{\Sigma}_i(\alpha, \gamma)$  скользящим контролем.

# Дискриминантный анализ

## Проблемы с ковариационной матрицей

Матрица  $\hat{\Sigma}_i$  окажется вырожденной если признаки линейно зависимы, и может оказаться плохо обусловлена.

Способы разрешения этой проблемы:

- Регуляризация ковариационной матрицы. Предлагается обращаться

$$\hat{\Sigma}_i + \tau \mathbf{I}_p,$$

где  $\tau$  выбрано по скользящему контролю.

- Нормальный наивный байесовский классификатор. Пусть

$$p_i(x) = \prod_{j=1}^p p_{ij}(x_j), \quad p_{ij}(x_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{(x_j - \mu_{ij})^2}{2\sigma_{ij}^2}}.$$

Классифицирующая функция:

$$\delta_i(x) = -\frac{1}{2} \sum_{j=1}^p \frac{(x_j - \mu_{ij})^2}{2\sigma_{ij}^2} + \log(\pi_i).$$

# Дискриминантный анализ

## Канонические переменные

Пусть  $\zeta = A\xi$ , тогда  $P(\zeta \mid \eta = G_k) = \mathcal{N}_p(A^T \mu_k, A^T \Sigma_k A)$ .

На выборочном языке:  $Z = A^T \mathbf{X}$ .

Выборочная дисперсия  $Z$  имеет вид

$$A^T \mathbf{T} A = A^T (\mathbf{W} + \mathbf{B}) A = A^T \mathbf{W} A + A^T \mathbf{B} A,$$

где

- $\mathbf{T}$  – total covariance matrix  $\mathbf{X}$ ,
- первое слагаемое – оценка внутригрупповых отклонений,
- второе слагаемое – оценка межгрупповых отклонений.

# Дискриминантный анализ

## Канонические переменные

Выборочная дисперсия  $Z$  имеет вид

$$A^T T A = A^T (W + B) A = A^T W A + A^T B A,$$

Обобщенная задача на собственные числа и собственные вектора:

$$\frac{A^T B A}{A^T W A} \rightarrow \max_A.$$

Пусть  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$  – собственные числа  $B^{-1}W$ , а  $A_1, \dots, A_d$  – собственные вектора. Причем  $A_i^T W A_j = 0$ . Максимум равен  $\lambda_1$  и достигается на  $A_1$ .

$$\max_{A, A \perp A_1} \frac{A^T B A}{A^T W A} = \lambda_2$$

достигается на  $A_2$  и так далее.

Вектора  $A_i$  – канонические коэффициенты, новые признаки  $Z_i$  – канонические переменные,  $Z_i$  ортогональны в обычном смысле.

# Дискриминантный анализ

## Значимость канонических переменных

Сколько канонических переменных нам окажется достаточно взять?

$H_0 : A_i, i = \ell, \dots, d$  не описывают отличия.

Введем статистику  $\Lambda$  — *prime*:

$$\Lambda_\ell^p = \prod_{i=\ell}^d \frac{1}{1 + \lambda_i}.$$

Тогда гипотезу выше можно переформулировать так

$$H_0 : \Lambda_\ell^p = 1 \Leftrightarrow \lambda_\ell = \dots = \lambda_d = 0 \Leftrightarrow \text{rank} \mathbf{B} = \ell - 1.$$

Критерий:

$$t = \Lambda_\ell^p \sim \Lambda_{\nu_{\mathbf{B}} + (\ell - 1), \nu_{\mathbf{W}} - (\ell - 1)}.$$

# Дискриминантный анализ

## Последовательный дискриминантный анализ

Какие признаки следует исключить?

- Имеют большой коэффициент множественной корреляции  $R^2 = R^2_{pooled}(\xi_i; \{\xi_j \mid j \neq i\})$ ;
- Не влияют на качество разделения. Введем статистику

$$\begin{aligned}(Partial\Lambda)_i &= \Lambda(X_i \mid X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_p) = \\ &= \frac{\Lambda(X_1, \dots, X_p)}{\Lambda(X_i \mid X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_p)}.\end{aligned}$$

Гипотеза:

$$H_0 : (Partial\Lambda)_i = 1.$$

Критерий:

$$F_i = \frac{1 - (Partial\Lambda)_i / \nu_B}{(Partial\Lambda)_i / (\nu_W - p + 1)} \sim F_{\nu_B, \nu_W - p + 1}.$$

Далее жадным образом отбираем признаки.

# Логистическая регрессия

## Постановка задачи

Задача: по выборке построить функцию классификации  $a(\xi)$ , на которой достигается минимум среднего риска

$$R(a) = \mathbb{E}_{\xi} \sum_{k=1}^K \mathbf{L}(G_k, a(\xi)) P(G_k | \xi) \rightarrow \min_a.$$

Выше был получен оптимальный байесовский классификатор, на котором достигается минимум  $R(a)$ :

$$a(x) = \arg \max_{G \in \mathcal{G}} P(G | \xi).$$

Далее рассмотрим логистическую регрессию как метод оценки апостериорных вероятностей  $P(G | \xi)$ .



# Логистическая регрессия

## Модель

Модель задается системой

$$\log \frac{P(\eta = G_i \mid \xi = \mathbf{x})}{P(\eta = G_K \mid \xi = \mathbf{x})} = \beta_{i0} + \beta_i^T \mathbf{x}, \quad i = 1, \dots, K - 1.$$

То есть границы между классами линейны.

Перейдем от логитов к вероятностям, их сумма будет равна единице

$$P(\eta = G_i \mid \xi = \mathbf{x}) = \frac{e^{\beta_{i0} + \beta_i^T \mathbf{x}}}{1 + \sum_{k=1}^{K-1} e^{\beta_{k0} + \beta_k^T \mathbf{x}}}, \quad i = 1, \dots, K - 1,$$

$$P(\eta = G_K \mid \xi = \mathbf{x}) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\beta_{k0} + \beta_k^T \mathbf{x}}}.$$

# Логистическая регрессия

## Метод максимального правдоподобия

Воспользуемся ММП для оценки параметров условного распределения. Логарифм функции правдоподобия

$$\ell(\theta) = \sum_{i=1}^N \log P(\eta = G_k \mid \xi = \mathbf{x}_i; \theta), \quad \theta = (\beta_{10}, \beta_1^T, \dots, \beta_{(K-1)0}, \beta_{K-1}^T).$$

Случай двух классов  $\mathcal{G} = \{0, 1\}$ . Тогда

$$\ell(\beta) = \sum_{i=1}^N (y_i \beta^T \mathbf{x}_i - \log(1 + e^{\beta^T \mathbf{x}_i})), \quad \beta = \{\beta_{10}, \beta_1\}.$$

Обозначим  $p(\mathbf{x}, \theta) = P(\eta = 0 \mid \xi = \mathbf{x}; \theta)$  и  $1 - p(\mathbf{x}, \theta) = P(\eta = 1 \mid \xi = \mathbf{x}; \theta)$ . Приравниваем производные к нулю, получаем систему из  $p + 1$  уравнения

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^N \mathbf{x}_i (y_i - p(\mathbf{x}_i; \beta)) = 0.$$

# Задача классификации

## Минимизация аппроксимированного эмпирического риска

На языке ML рассмотрим задачу классификации,  $Y = \{-1, +1\}$ .

Введем некоторые понятия:

- $a(x, \beta) = \text{sign } f(x, \beta)$  – семейство классификаторов;
- $M_i(\beta) = y_i f(x_i, \beta)$  – отступ объекта  $x_i$ ;
- $\mathcal{L}(M_i(\beta))$  – монотонно невозрастающая функция потерь, мажорирующая  $[M < 0]$ .

Задача поиска классификатора  $a(x, \beta)$  сводится к задаче

$$Q(\beta, \mathbf{X}) = \sum_{i=1}^N [M_i(\beta) < 0] \leq \sum_{i=1}^N \mathcal{L}(M_i(\beta)) \rightarrow \min_{\beta}.$$

Положив  $\mathcal{L}(M_i(\theta)) = -\log P(x_i, y_i; \theta)$  получаем эквивалентность с задачей максимизации правдоподобия.

# Логистическая регрессия

## Минимизация аппроксимированного эмпирического риска

В логистической регрессии минимизируется аппроксимация:

$$Q(\beta) = \sum_{i=1}^N \log(1 + e^{-y_i \beta^T x_i}) \rightarrow \min_{\beta},$$

то есть функция потерь имеет вид  $\mathcal{L}(M_i(\theta)) = \log(1 + e^{-y_i \beta^T x_i})$ .

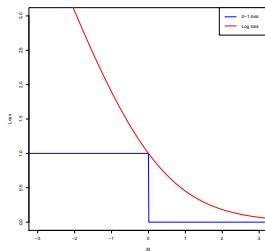


Рис.: Логистическая функция потерь

# Логистическая регрессия

## Метод Ньютона-Рафсона

Используем метод Ньютона-Рафсона для решения системы

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^N \mathbf{x}_i (y_i - p(\mathbf{x}_i; \beta)) = 0.$$

Гессиан логарифма правдоподобия

$$\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} = - \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^T p(\mathbf{x}_i; \beta) (1 - p(\mathbf{x}_i; \beta)) = 0.$$

$\beta^{old}$  – начальное приближение  $\beta$ , итерация алгоритма:

$$\beta^{new} = \beta^{old} - \left( \frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} \right)^{-1} \frac{\partial \ell(\beta)}{\partial \beta},$$

где производные вычисляются в точке  $\beta^{old}$ .

# Логистическая регрессия

## Итерация алгоритма Н-Р как IRLS

Пусть  $\mathbf{y}$  – ответы  $y_i$ ,  $\mathbf{X}$  – матрица данных,  $\mathbf{p} = (p(\mathbf{x}_i; \beta^{old}))$ ,  $\mathbf{W} = \text{diag}\{w_1, \dots, w_n\}$ , где  $w_i = p(\mathbf{x}_i; \beta^{old})(1 - p(\mathbf{x}_i; \beta^{old}))$ . Тогда

$$\frac{\partial \ell(\beta)}{\partial \beta} = \mathbf{X}^T(\mathbf{y} - \mathbf{p}), \quad \frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} = -\mathbf{X}^T \mathbf{W} \mathbf{X}.$$

Перепишем шаг алгоритма Ньютона-Рафсона

$$\beta^{new} = \beta^{old} + (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T(\mathbf{y} - \mathbf{p}) = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{z}.$$

Получили взвешенную регрессию, где в качестве ответа выступает

$$\mathbf{z} = \mathbf{X} \beta^{old} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p}).$$

Получили IRLS, на каждом шаге решается задача

$$\beta^{new} = \arg \min_{\beta} (\mathbf{z} - \mathbf{X} \beta)^T \mathbf{W} (\mathbf{z} - \mathbf{X} \beta).$$

В качестве  $\beta^{old}$  можно взять оценки линейной регрессии или  $\beta^{old} = 0$ .

# Логистическая регрессия

## Регуляризация

LASSO:

$$\max_{\beta_0, \beta} \sum_{i=1}^N \left( y_i(\beta_0 + \beta^T \mathbf{x}_i) - \log(1 + e^{\beta_0 + \beta^T \mathbf{x}_i}) \right) - \lambda \sum_{j=1}^p |\beta_j|.$$

Ridge Regression:

$$\max_{\beta_0, \beta} \sum_{i=1}^N \left( y_i(\beta_0 + \beta^T \mathbf{x}_i) - \log(1 + e^{\beta_0 + \beta^T \mathbf{x}_i}) \right) - \lambda \sum_{j=1}^p \beta_j^2.$$

Для этого можно снова использовать алгоритм Ньютона-Рафсона,  $\lambda$  выбирается с помощью скользящего контроля.

## LR vs. LDA

Посмотрим на log-posterior odds между классами  $i$  и  $K$  в случае LDA:

$$\begin{aligned}\log \frac{P(\eta = G_i \mid \boldsymbol{\xi} = \mathbf{x})}{P(\eta = G_K \mid \boldsymbol{\xi} = \mathbf{x})} &= \\ &= -\frac{1}{2}(\mu_i - \mu_K)^T \Sigma^{-1}(\mu_i + \mu_K) + (\mu_i - \mu_K)^T \Sigma^{-1} \mathbf{x} + \log(\pi_i / \pi_K) = \\ &= \alpha_{i0} + \alpha_i^T \mathbf{x}.\end{aligned}$$

С другой стороны, линейная логистическая регрессия имеет линейные логиты по построению

$$\log \frac{P(\eta = G_i \mid \boldsymbol{\xi} = \mathbf{x})}{P(\eta = G_K \mid \boldsymbol{\xi} = \mathbf{x})} = \beta_{i0} + \beta_i^T \mathbf{x}.$$

Модели выглядят очень похоже. Различие заключается в том как оцениваются линейные коэффициенты.



## LR vs. LDA

Выпишем совместную плотность  $\xi$  и  $\eta$

$$P(\xi = \mathbf{x}, \eta = G_i) = P(\mathbf{x})P(\eta = G_i | \xi = \mathbf{x}).$$

И в линейном дискриминантном анализе, и в логистической регрессии второй множитель выражается как

$$P(\eta = G_i | \xi = \mathbf{x}) = \frac{e^{\beta_{i0} + \beta_i^T \mathbf{x}}}{1 + \sum_{k=1}^{K-1} e^{\beta_{k0} + \beta_k^T \mathbf{x}}}.$$

В логистической регрессии  $P(\mathbf{x})$  – произвольная плотность, а параметры  $P(\eta = G_i | \xi = \mathbf{x})$  оцениваются максимизацией условного правдоподобия (discriminative learning). Решается задача

$$\ell(\theta) = \sum_{i=1}^N \log P(\eta = G_k | \xi = \mathbf{x}_i; \theta) \rightarrow \max_{\theta}.$$

## LR vs. LDA

В LDA максимизируется полноценный логарифм функции правдоподобия совместной плотности

$$P(\mathbf{x}, \eta = G_i) = \phi(\mathbf{x}; \mu_i, \Sigma) \pi_i,$$

где  $\phi(\mathbf{x}; \mu_i, \Sigma)$  - плотность нормального распределения (generative learning).

Решается задача

$$\ell(\mu_i, \Sigma) = \sum_{i=1}^N \phi(\mathbf{x}_i; \mu_i, \Sigma) \pi_i \rightarrow \max_{\mu_i, \Sigma}.$$

При этом  $P(\mathbf{x})$  – это смесь распределений

$$P(\mathbf{x}) = \sum_{i=1}^K \phi(\mathbf{x}; \mu_i, \Sigma) \pi_i.$$

# LR vs. LDA

Некоторые плюсы и минусы методов:

- Предположение о нормальности распределения дает нам больше информации о параметрах, отсюда меньше дисперсия оценок;
- Точки, далекие от разделяющей плоскости (у которых в логистической регрессии вес будет меньше), влияют на оценку ковариационной матрицы. LDA не является робастным по отношению к выбросам;
- В логистической регрессии более гибкая модель;
- В логистической регрессии требуется оценка меньшего числа параметров;
- Алгоритм Ньютона-Рафсона требует обращения матрицы на каждом шаге.

# Непараметрическая классификация

## Непараметрическая оценка плотности

### Локальная непараметрическая оценка Парзена-Розенבלата

$$\hat{p}_h(z) = \frac{1}{N} \sum_{i=1}^N \prod_{j=1}^p \frac{1}{h_j} K\left(\frac{z_j - x_{ij}}{h_j}\right),$$

где

- $K(x)$  – ядро, четная и нормированная функция  $\int K(x)dx = 1$ ;
- $h > 0$  – ширина окна, выбирается с помощью скользящего контроля (LOO).

Если

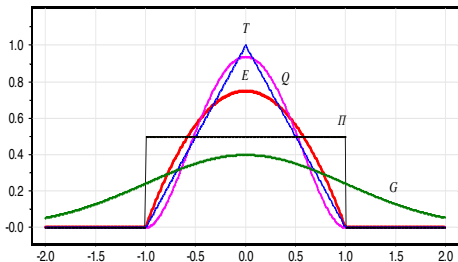
- $K(x)$  – непрерывно,  $\int K(x)^2 dx < \infty$ ,
- последовательность  $h_N$  такая, что  $\lim_{N \rightarrow \infty} h_N = 0$ ,  $\lim_{N \rightarrow \infty} N h_N = \infty$ ,

тогда  $\hat{p}_{h_m}(z) \rightarrow p(x)$  п.в. при  $N \rightarrow \infty$ .

# Непараметрическая классификация

## Непараметрическая оценка плотности

Выбор ядра не влияет на качество оценки, но определяет гладкость функции  $\hat{\rho}_h$  и влияет на эффективность вычислений.



- E – Епанечикова
- Q – Квартическое
- T – Треугольное
- G – Гауссовское
- П – Прямоугольное