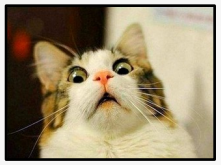
# Лекция 5: методы восстановления плотности распределения

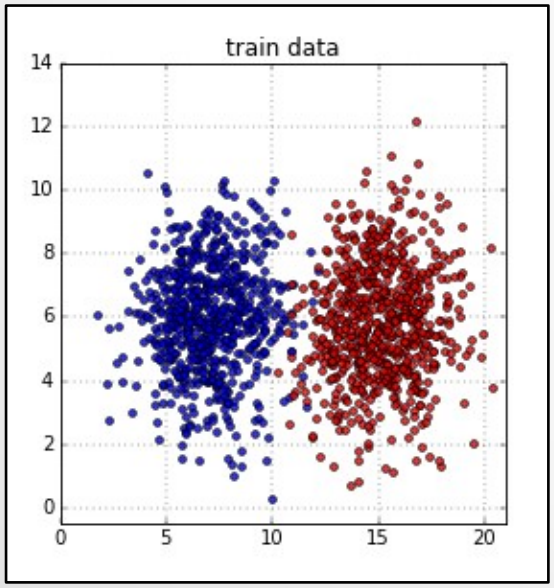
Евгений Борисов

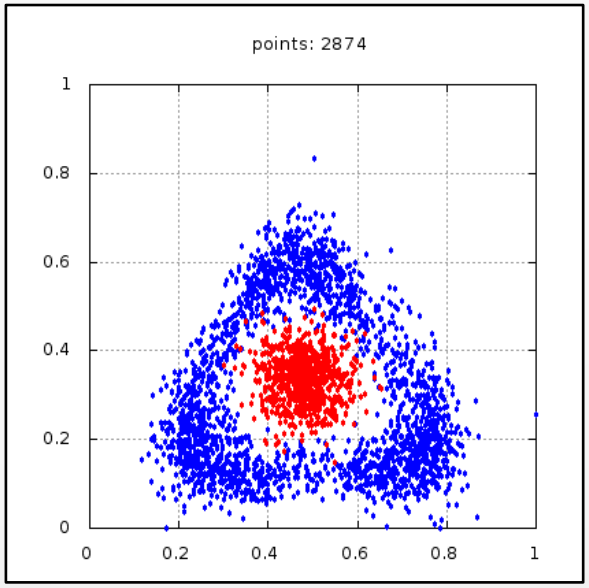
четверг, 18 октября 2018 г.

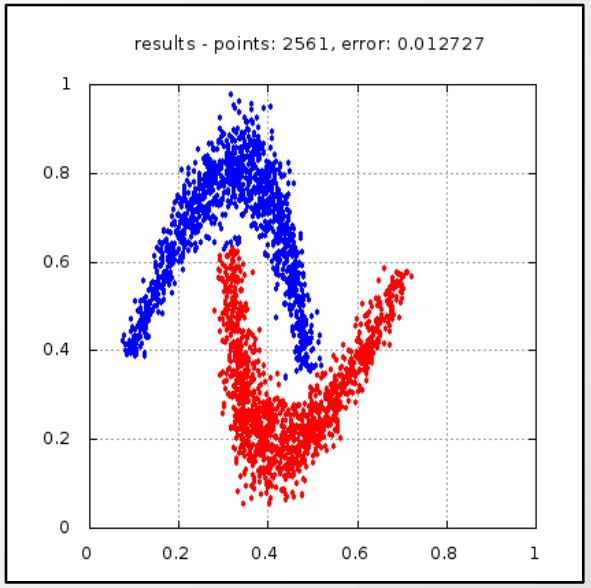
# Классификатор

разделения объектов на классы

**Детектор котов:**  → вектор-признак → есть/нет



# Восстановление плотности

$$a(x) = \underset{y \in Y}{argmax}\, \lambda_y P(y)\, p(x|y)$$

## Байесовский классификатор

$\lambda_y$ - *потеря для объектов y*

*P(y) - априорная вероятность класса y*
*(доля примеров класса y,*
*пропорция классов должна соответствовать)*

*p(x|y) - ф-ция правдоподобия класса y (плотность)*

# Восстановление плотности

**подходы к оценке плотности распределения**:

- непараметрический
- параметрический
- смеси распределений

# Восстановление плотности

параметрический подход к оцениванию плотности

$$\hat{p}(x) = \varphi(x, \theta)$$

# Восстановление плотности

параметрический подход к оценке плотности

$$\hat{p}(x) = \varphi(x, \theta)$$

смеси распределений

$$\hat{p}(x) = \sum_{j=1}^{k} w_j \varphi_j(x, \theta_j)$$

# Восстановление плотности

параметрический подход к оцениванию плотности

$$\hat{p}(x) = \varphi(x, \theta)$$

смеси распределений

$$\hat{p}(x) = \sum_{j=1}^{k} w_j \varphi_j(x, \theta_j)$$

НЕпараметрический подход к оцениванию плотности

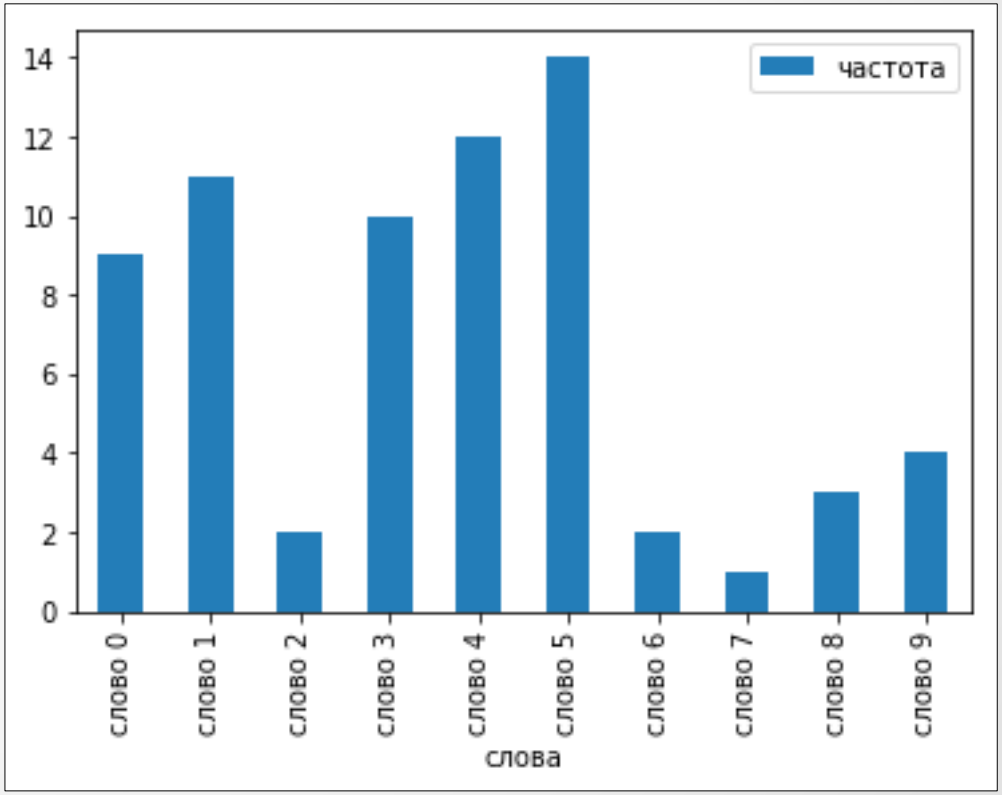
$$\hat{p}(x) = \frac{1}{m V(h)} \sum_{j=1}^{m} K\left(\frac{\rho(x, x_j)}{h}\right)$$

# Восстановление плотности

***НЕпараметрический подход***

дискретный случай (гистограмма)

$$\hat{p}(x) = \frac{1}{m} \sum_{i=1}^{m} [x = x_i]$$



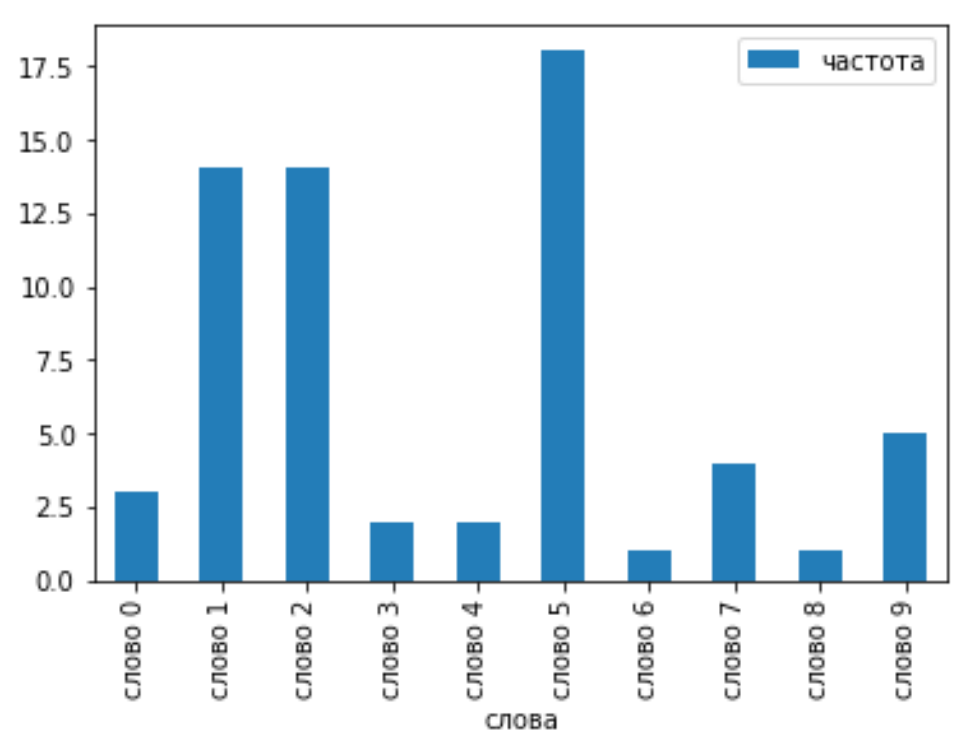

***пример:*** распределение повторов слов в тексте

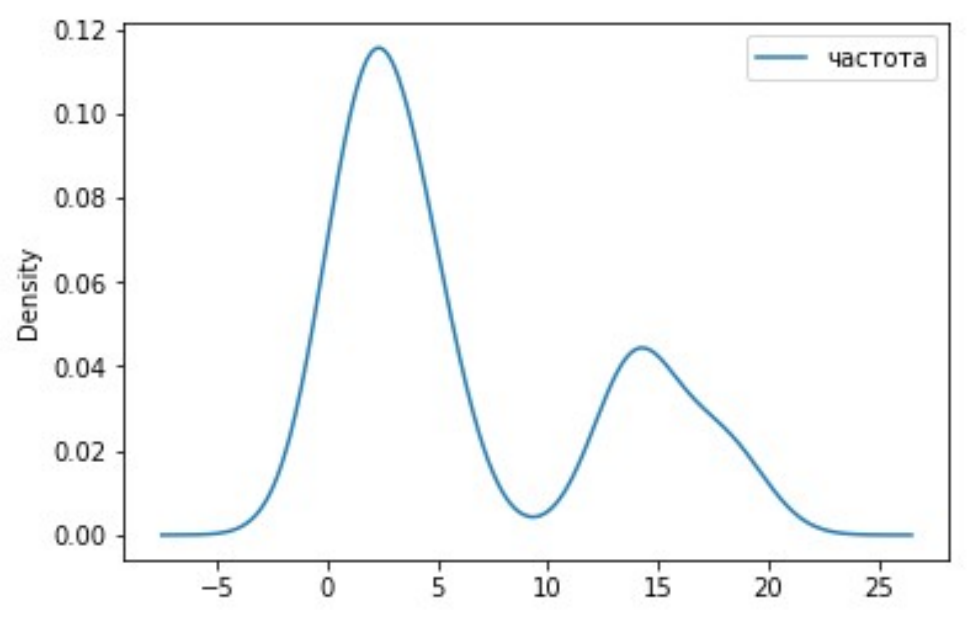
# Восстановление плотности

***непараметрические методы***

*непрерывный случай: доля объектов попавших в окно ширины h*

$$\hat{p}(x) = \frac{1}{mh} \sum_{i=1}^{m} \frac{1}{2} \left[ \frac{|x - x_i|}{h} < 1 \right]$$



# Восстановление плотности

***непараметрические методы:***

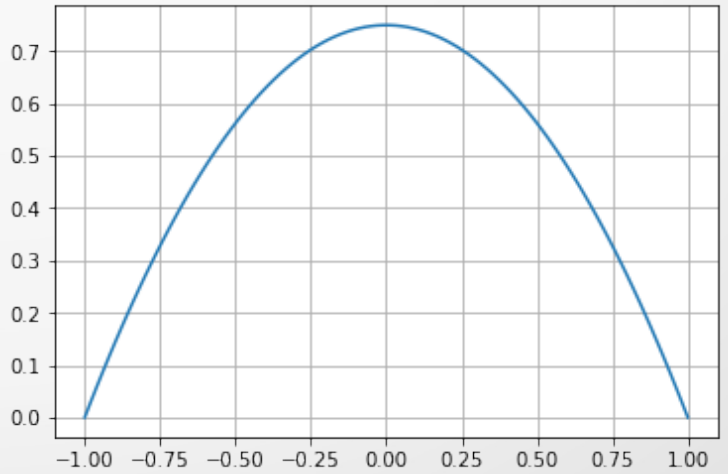
**оценка плотности Парзена-Розенблата**

K(r) - ядро
$\rho(x_1,x_2)$ - мера на X
V(h) - нормирующий множитель

$$\hat{p}(x)=\frac{1}{m\,V(h)}\sum_{i=1}^{m}K\left(\frac{\rho(x,x_i)}{h}\right)$$

*ядро Епанечникова:*

$$K(r)=\frac{3}{4}(1-r^2);|r|\le 1$$



# Восстановление плотности

Байесовский классификатор: метод Парзеновского окна

$$a(x)=\underset{y\in Y}{argmax}\,\lambda_y P(y)p(x|y)$$

$$a(x,X^l,h)=\underset{y\in Y}{argmax}\,\lambda_y P(y)\frac{1}{l_y}\sum_{i:\,y=y_i}K\left(\frac{\rho(x,x_i)}{h}\right)$$

# Восстановление плотности

***непараметрические методы:***

*выбор оптимального размера Парзеновского окна* ***h***

**методом скользящего контроля (Leave One Out, LOO)**

$$LOO(h, X) = \sum_{i=1}^{l} \left[ a(x_i, \{X \setminus x_i\}, h) \neq y_i \right] \rightarrow \min_h$$

параметр **h** выбираем перебором

для разных значений **h**
проверяем суммарную ошибку на учебном множестве

при этом
из учебного набора **X** удаляется текущий (проверяемый) пример

# Восстановление плотности

***параметрический подход***

$$\hat{p}(x) = \varphi(x, \theta)$$

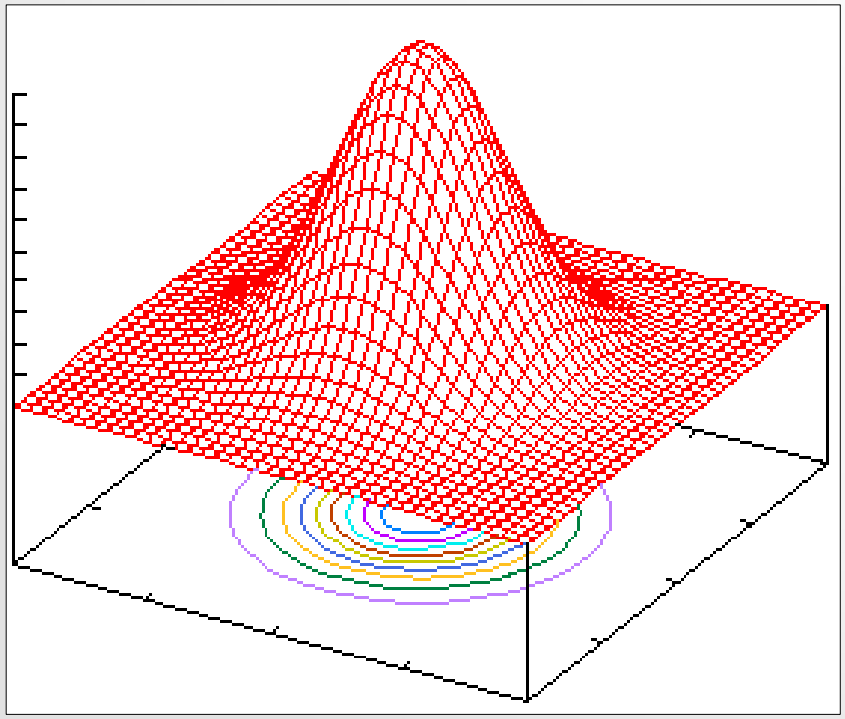
принцип максимума правдоподобия

$$L(\theta, X) = \sum_{i=1}^{m} \ln \varphi(x_i, \theta) \rightarrow \max_{\theta}$$

# Восстановление плотности

**параметрический подход**:
допустим - **p(x)** это нормальная плотность

$$p(x) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)}{\sqrt{(2\pi)^n \det \Sigma}}$$



*n-мерная нормальная плотность*

# Восстановление плотности

параметры оценки максимального правдоподобия
для n-мерной гауссовской плотности имеют следующий вид

мат.ожидание

$$\hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} x_i$$

матрица ковариаций

$$\hat{\Sigma} = \frac{1}{m} \sum_{i=1}^{m} (x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

# Восстановление плотности

**Теорема:** параметры оценки максимального правдоподобия для n-мерных гауссовских плотностей классов y имеют следующий вид

$$\hat{\mu}_y = \frac{1}{l_y} \sum_{i:\, y = y_i} x_i \qquad \hat{\Sigma}_y = \frac{1}{l_y} \sum_{i:\, y = y_i} (x_i - \hat{\mu}_y)(x_i - \hat{\mu}_y)^T$$

**Байесовский классификатор:** квадратичный дискриминант

$$a(x) = \underset{y \in Y}{argmax} \left( \ln(\lambda_y P_y) - (x - \hat{\mu}_y)^T \hat{\Sigma}_y^{-1} (x - \hat{\mu}_y) - \frac{1}{2} \ln(det\, \hat{\Sigma}_y) \right)$$

# Восстановление плотности

**Дополнение**:
если матрицы ковариаций классов равны
то параметры оценки плотности имеют следующий вид

$$\hat{\mu}_y = \frac{1}{l_y} \sum_{i:\, y = y_i} x_i \qquad \hat{\Sigma} = \frac{1}{l} \sum_{i=1}^{l} (x_i - \hat{\mu}_{yi})(x_i - \hat{\mu}_{yi})^T$$

**Байесовский классификатор:** линейный дискриминант Фишера

$$a(x) = \underset{y \in Y}{argmax} \left( \ln(\lambda_y P_y) - \frac{1}{2} \hat{\mu}_y^T \hat{\Sigma}^{-1} \hat{\mu}_y + x^T \hat{\Sigma}^{-1} \hat{\mu}_y \right)$$

# Восстановление плотности

***смеси распределений***

$$\hat{p}(x)=\sum_{j=1}^{k} w_j \varphi_j(x,\theta_j); \qquad \sum_{j=1}^{k} w_j=1; \qquad w_j \geq 0$$

# Восстановление плотности

**смеси распределений:**

**EM (expectation–maximization algorithm)**

# Восстановление плотности

**смеси распределений: RBF**

# Восстановление плотности

**Пример:** детектор новых объектов для неподвижных камер

# Классификатор: литература

git clone https://github.com/mechanoid5/ml_lectorium.git

К.В. Воронцов Байесовская теория классификации и методы восстановления плотности. - Курс "Машинное обучение" ШАД Яндекс 2014

Борисов Е.С. Байесовский классификатор. *http://mechanoid.kiev.ua/ml-bayes.html*

Борисов Е.С. Восстановление смеси плотностей распределений с помощью EM-алгоритма. http://mechanoid.kiev.ua/ml-em-base.html

# Классификатор: почти последний слайд...

**Вопросы ?**

# Классификатор: практика

**источники данных для экспериментов**


sklearn.datasets

UCI Repository

kaggle