Лекция 5:
методы восстановления плотности распределения

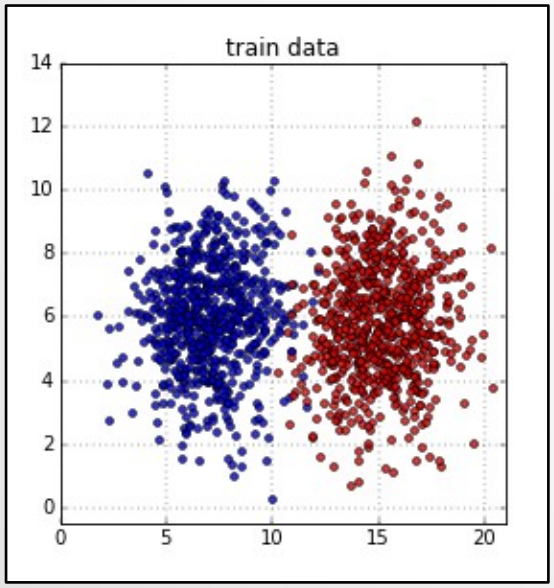
Евгений Борисов

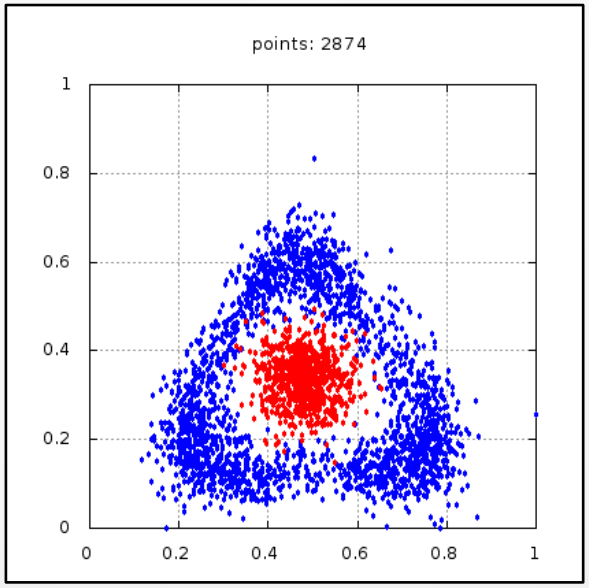
четверг, 18 октября 2018 г.

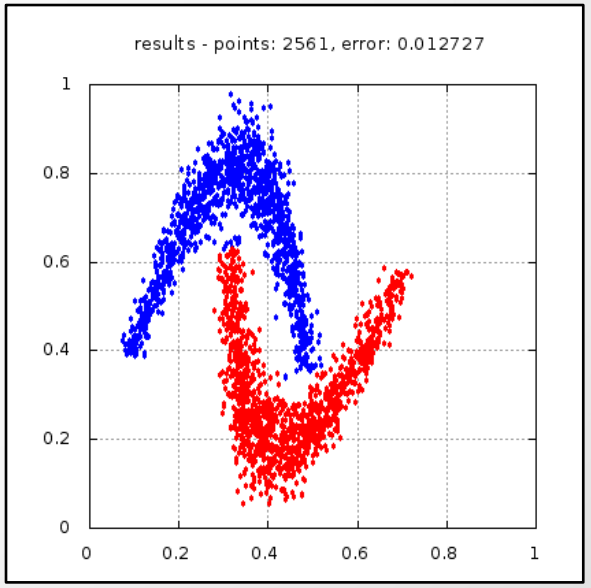
Классификатор

разделения объектов на классы

Детектор котов:  → вектор-признак → есть/нет



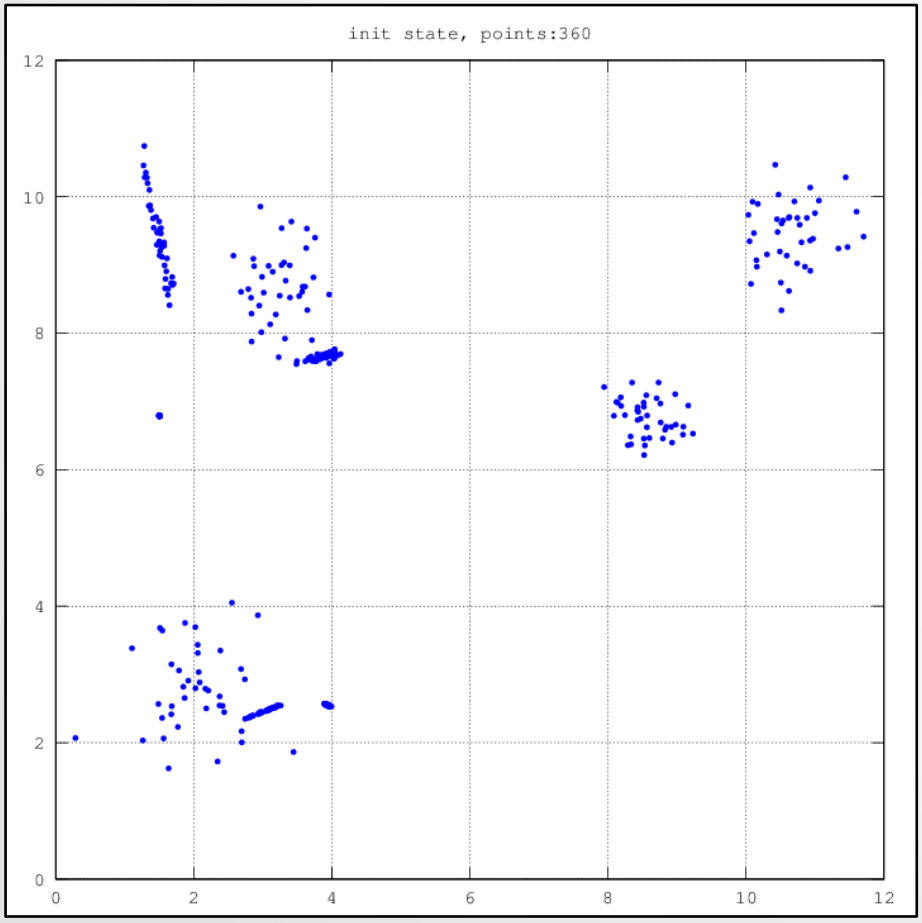


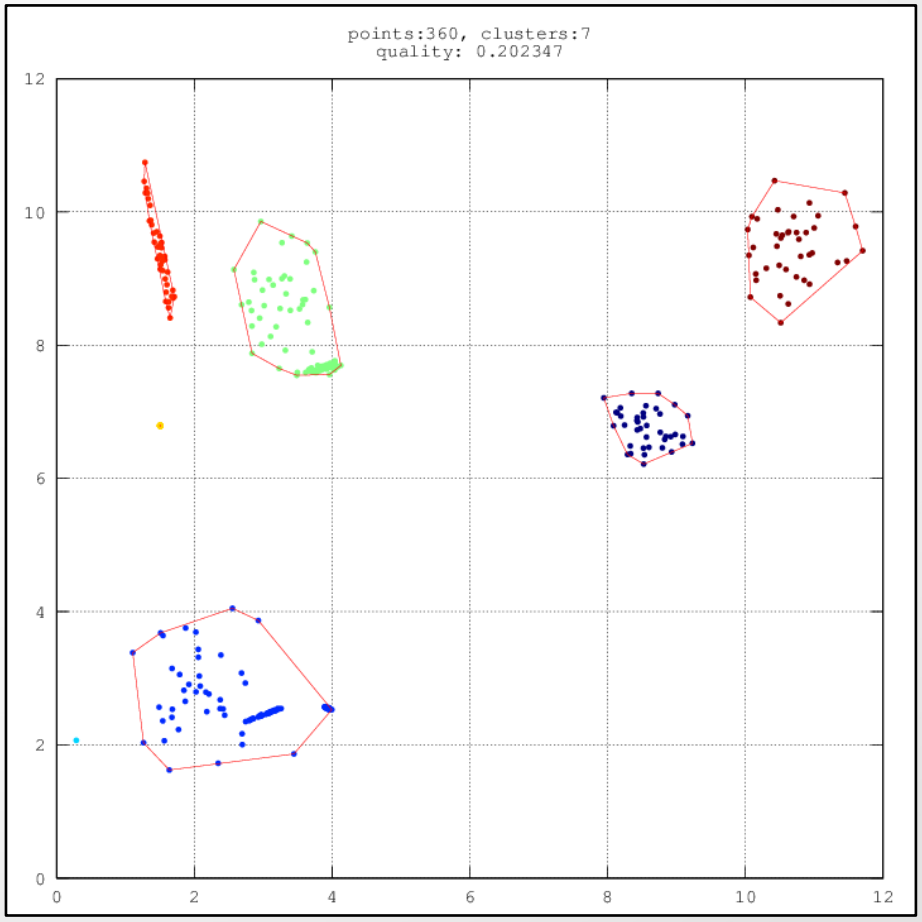


Кластеризатор

объединение схожих объектов в группы

Поиск похожих текстов: текст → признаки → группа





Восстановление плотности

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

Байесовский классификатор

λ_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:

- параметрический
- непараметрический
- смеси распределений

Восстановление плотности

подходы к оценке плотности распределения:

параметрический подход

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

НЕпараметрический подход

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

смеси распределений

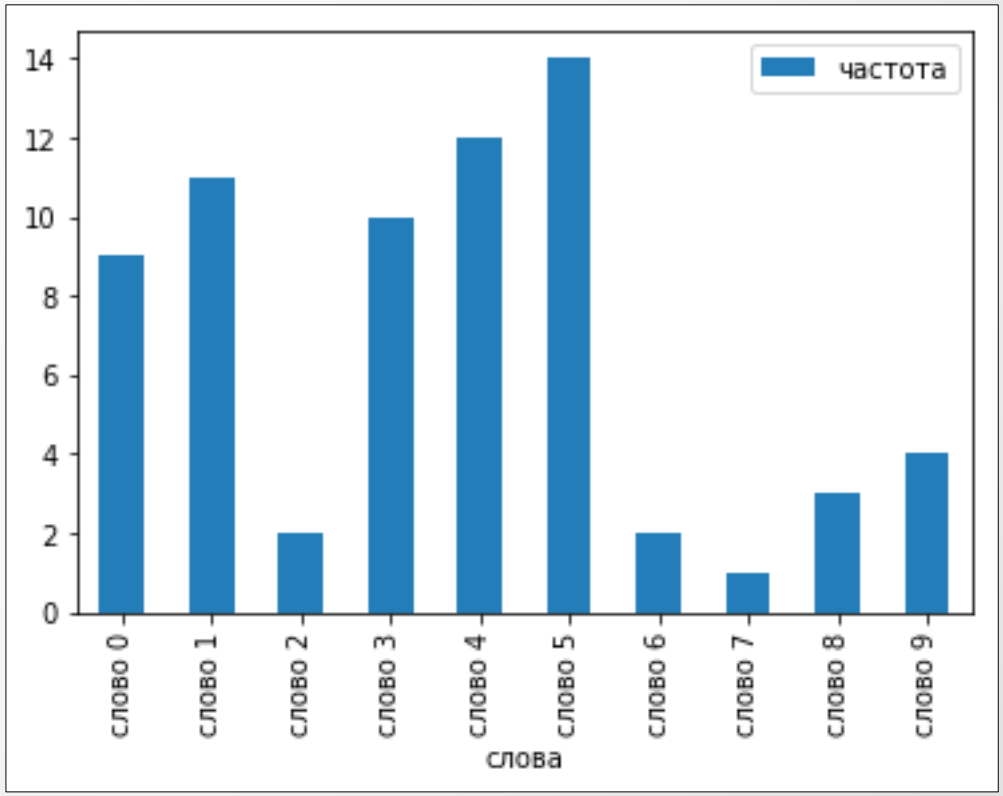
$$\hat{p}(x) = \sum_{j=1}^{k} w_j \varphi_j(x, \theta_j)$$

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

<u>непрерывный случай:</u> 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

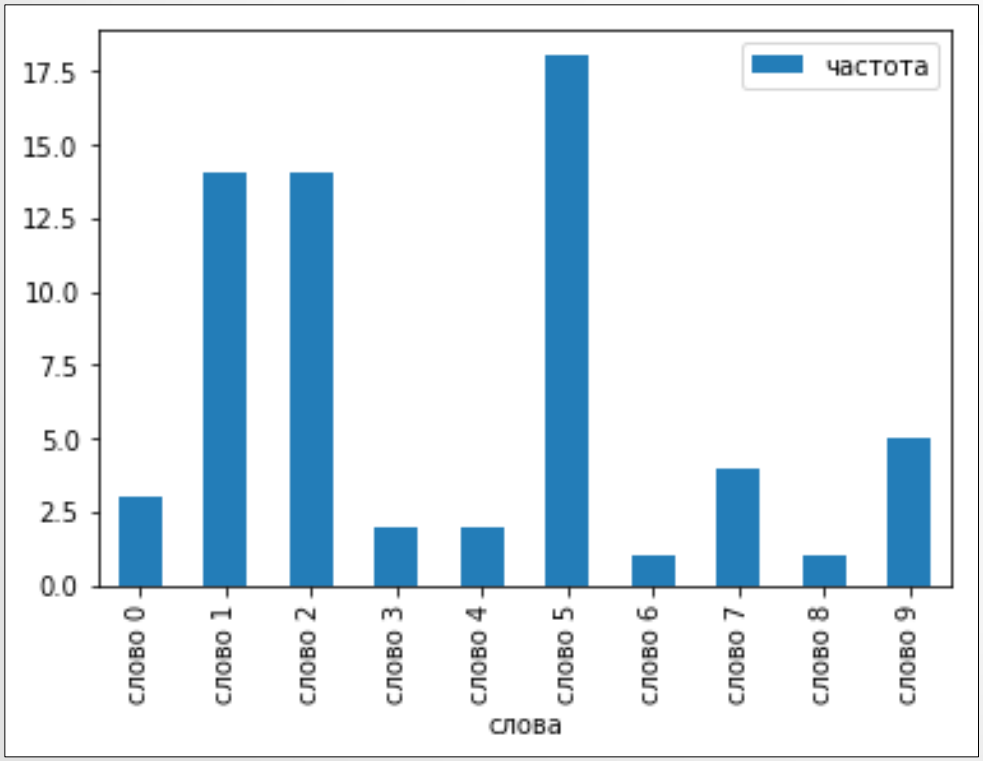
непараметрические методы:
оценка плотности Парзена-Розенблата

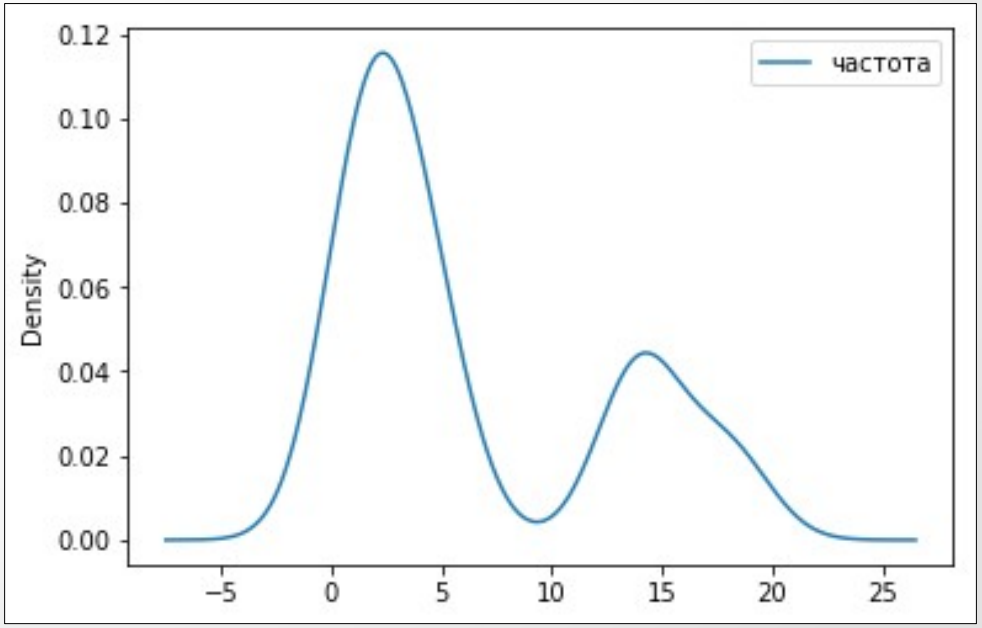
$$\hat{p}(x)=\frac{1}{m\,V(h)}\sum_{i=1}^{m}K\left(\frac{\rho(x,x_i)}{h}\right)$$

ядерное сглаживание (гистограммы)

KDE, Kernel Density Estimation

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

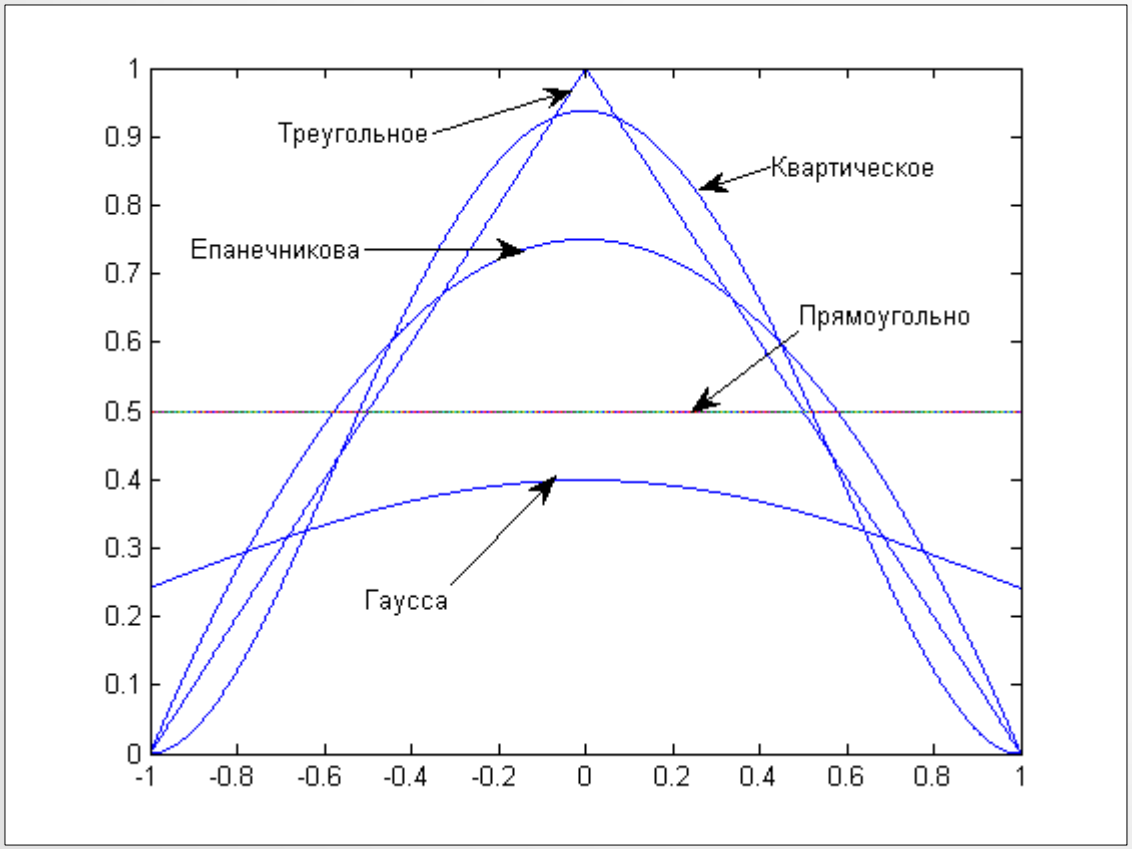




Восстановление плотности

непараметрические методы:
функции ядра для сглаживания гистограммы

KDE, Kernel Density Estimation



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

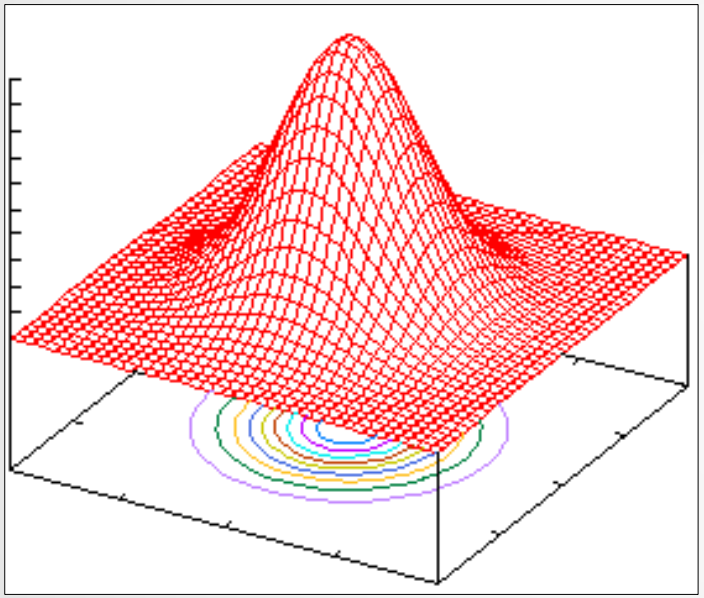
$$\hat{p}(x) = \varphi(x, \theta)$$

принцип максимума правдоподобия

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

Восстановление плотности

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



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

Восстановление плотности

n-мерная гауссовская плотность

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

оценки параметров для максимального правдоподобия 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

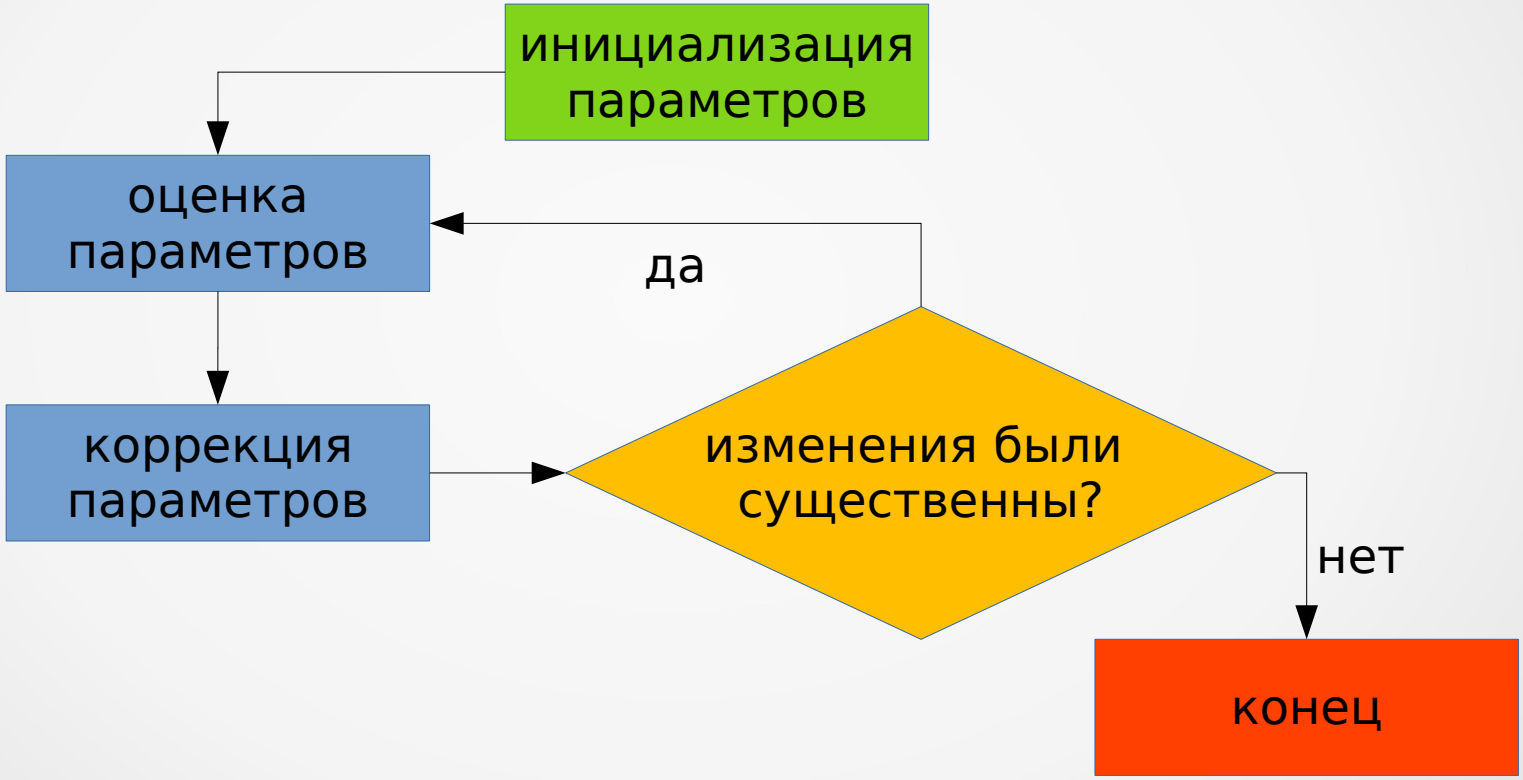
$$\hat{p}(x)=\sum_{j=1}^{k} w_j \varphi_j(x,\theta_j);$$

$$\sum_{j=1}^{k} w_j=1; \qquad w_j \geq 0$$

Восстановление плотности

EM (expectation-maximization algorithm):

базовый вариант алгоритма



Восстановление плотности

EM (expectation–maximization algorithm)

оценка

$$g_{ij} = \frac{w_j \varphi_j(x_i, \theta_j)}{\sum_{k=1}^{s} w_k \varphi_k(x_i, \theta_k)}$$

i=1...m
m - количество примеров X
s - количество компонент смеси

Восстановление плотности

EM (expectation–maximization algorithm)

оценка

$$g_{ij} = \frac{w_j \varphi_j(x_i, \theta_j)}{\sum_{k=1}^{s} w_k \varphi_k(x_i, \theta_k)}$$

коррекция

$$w_j = \frac{1}{m} \sum_{i=1}^{m} g_{ij}$$

$$\theta_j = \underset{\theta}{argmax} \sum_{i=1}^{m} g_{ij} \ln \varphi_j(x_i, \theta)$$

i=1...m
m - количество примеров X
s - количество компонент смеси

Восстановление плотности

EM (expectation–maximization algorithm)

оценка

$$g_{ij} = \frac{w_j \varphi_j(x_i, \theta_j)}{\sum_{k=1}^{s} w_k \varphi_k(x_i, \theta_k)}$$

коррекция

$$w_j = \frac{1}{m} \sum_{i=1}^{m} g_{ij}$$

$$\theta_j = \underset{\theta}{argmax} \sum_{i=1}^{m} g_{ij} \ln \varphi_j(x_i, \theta)$$

i=1...m
m - количество примеров X
s - количество компонент смеси

условие остановки: параметры не изменились

$$|g_{ij}(t-1) - g_{ij}(t)| < \delta ; 0 < \delta < 1$$

Восстановление плотности

EM для нормальной плотности

$$p(x)=\sum_k w_k N(x;\Sigma_k,\mu_k)$$

n-мерная гауссовская плотность

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

оценки параметров для максимального правдоподобия 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

EM для нормальной плотности $p(x)=\sum_k w_k N(x;\Sigma_k,\mu_k)$

оценка: $g_{ij}=\dfrac{w_j N(x_i;\Sigma_j,\mu_j)}{\sum_k w_k N(x_i;\Sigma_k,\mu_k)}$

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

задача: $\Sigma_j,\mu_j=\underset{\Sigma,\mu}{argmax}\sum_{i=1}^{m} g_{ij}\ln N(x_i;\Sigma,\mu)$

Восстановление плотности

EM для нормальной плотности $p(x)=\sum_k w_k N(x;\Sigma_k,\mu_k)$

оценка: $g_{ij}=\frac{w_j N(x_i;\Sigma_j,\mu_j)}{\sum_k w_k N(x_i;\Sigma_k,\mu_k)}$

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

задача: $\Sigma_j,\mu_j=\underset{\Sigma,\mu}{argmax}\sum_{i=1}^{m} g_{ij}\ln N(x_i;\Sigma,\mu)$

коррекция: $w_j=\frac{1}{m}\sum_{i=1}^{m} N_j$ $\quad N_j=\sum_{i=1}^{m} g_{ij}$

вес компоненты

Восстановление плотности

EM для нормальной плотности $p(x)=\sum_k w_k N(x;\Sigma_k,\mu_k)$

оценка: $g_{ij}=\frac{w_j N(x_i;\Sigma_j,\mu_j)}{\sum_k w_k N(x_i;\Sigma_k,\mu_k)}$

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

задача: $\Sigma_j,\mu_j=\underset{\Sigma,\mu}{argmax}\sum_{i=1}^m g_{ij}\ln N(x_i;\Sigma,\mu)$

коррекция:

$$w_j=\frac{1}{m}\sum_{i=1}^m N_j$$

вес компоненты

$$N_j=\sum_{i=1}^m g_{ij}$$

$$\mu_j=\frac{1}{N_j}\sum_{i=1}^m g_{ij}x_i$$

мат.ожидание компоненты

Восстановление плотности

EM для нормальной плотности $p(x)=\sum_k w_k N(x;\Sigma_k,\mu_k)$

оценка: $g_{ij}=\frac{w_j N(x_i;\Sigma_j,\mu_j)}{\sum_k w_k N(x_i;\Sigma_k,\mu_k)}$

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

задача:

$$\Sigma_j,\mu_j=\underset{\Sigma,\mu}{argmax}\sum_{i=1}^{m} g_{ij}\ln N(x_i;\Sigma,\mu)$$

коррекция:

$$w_j=\frac{1}{m}\sum_{i=1}^{m} N_j$$

вес компоненты

$$N_j=\sum_{i=1}^{m} g_{ij}$$

$$\mu_j=\frac{1}{N_j}\sum_{i=1}^{m} g_{ij}x_i$$

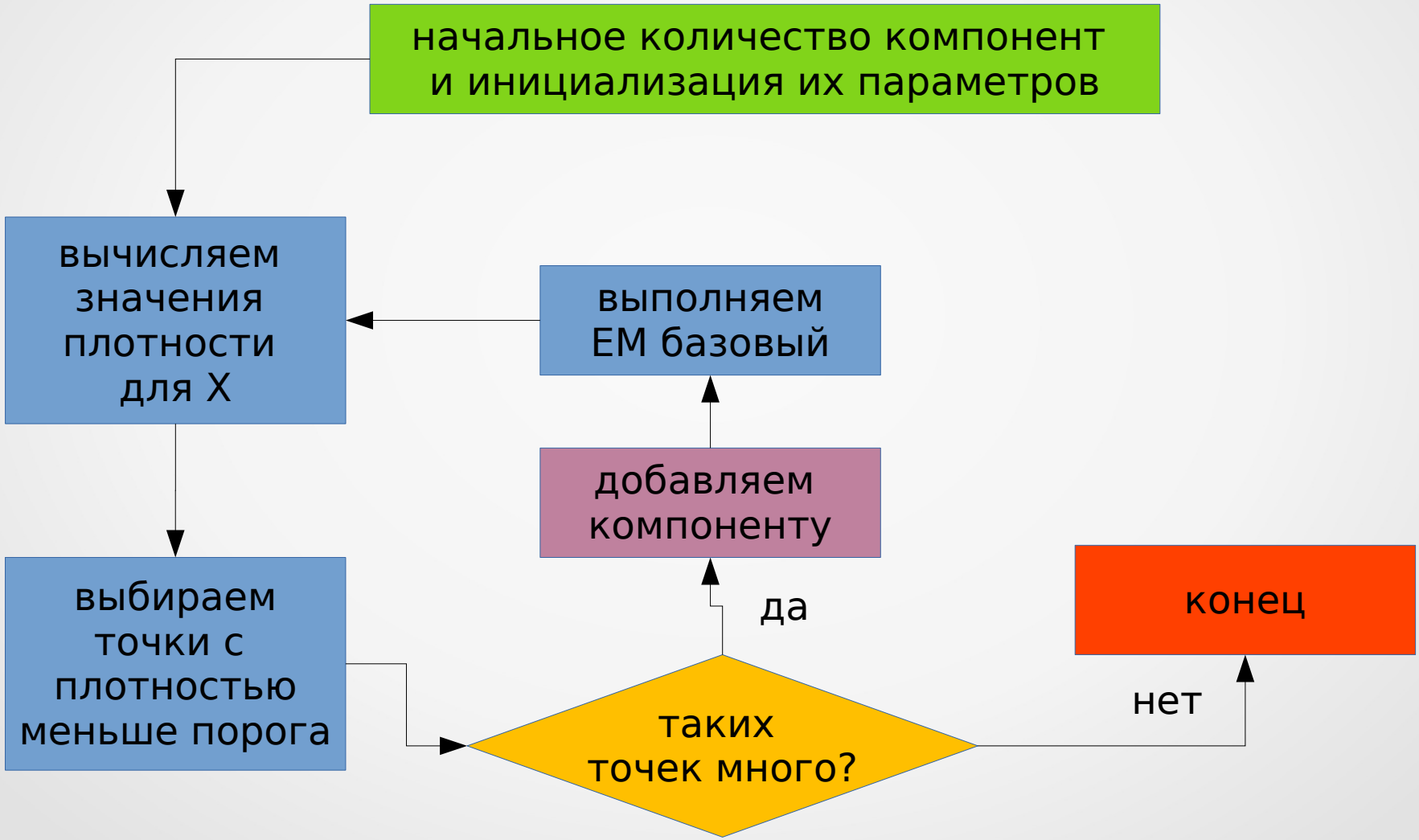
мат.ожидание компоненты

$$\Sigma_j=\frac{1}{cN_j}\sum_{i=1}^{m} g_{ij}(x_i-\mu_j)^T(x_i-\mu_j);\ 0<c\le 1$$

матрица ковариаций компоненты

Восстановление плотности

EM с последовательным добавлением компонент



Восстановление плотности

EM последовательное добавление компонент (для нормальной плотности)

начальные значения параметров первой компоненты смеси

$$w_1 = 1$$

вес компоненты

$$\Sigma_1 = cov(X)$$

матрица ковариаций компоненты

$$\mu_1 = \frac{1}{|X|} \sum X$$

мат.ожидание компоненты

Восстановление плотности

EM последовательное добавление компонент
(для нормальной плотности)

$X_{low} \subset X$ - точки с правдоподобием (значением смеси) ниже порога

начальные значения параметров новой компоненты смеси

$$w_{k+1} = \frac{|X_{low}|}{|X|}$$

вес компоненты

$$\Sigma_{k+1} = cov(X_{low})$$

матрица ковариаций компоненты

$$\mu_{k+1} = w_{k+1} \frac{1}{|X_{low}|} \sum X_{low}$$

мат.ожидание компоненты

Восстановление плотности

EM последовательное добавление компонент (для нормальной плотности)

$X_{low} \subset X$ - точки с правдоподобием (значением смеси) ниже порога

начальные значения параметров новой компоненты смеси

$$w_{k+1} = \frac{|X_{low}|}{|X|}$$

вес компоненты

$$\Sigma_{k+1} = cov(X_{low})$$

матрица ковариаций компоненты

$$\mu_{k+1} = w_{k+1} \frac{1}{|X_{low}|} \sum X_{low}$$

мат.ожидание компоненты

коррекция весов старых компонент смеси

$$w_i := w_i(1 - w_{k+1})$$

после определения новых параметров смеси запускаем EM

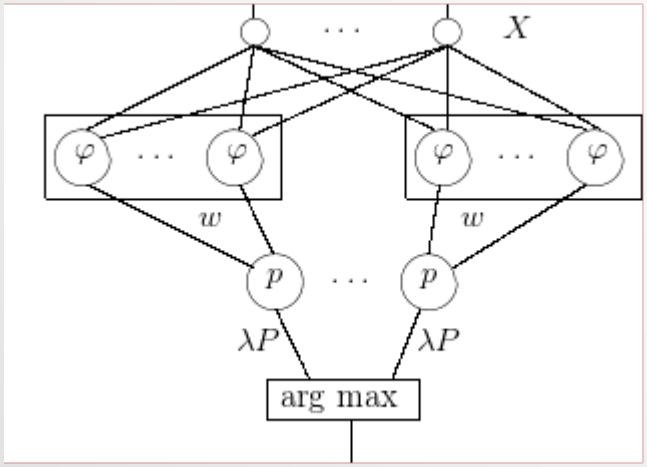
Восстановление плотности

RBF - сеть радиальных базисных функций

Байесовский классификатор

плотности классов - смеси нормальных распределений

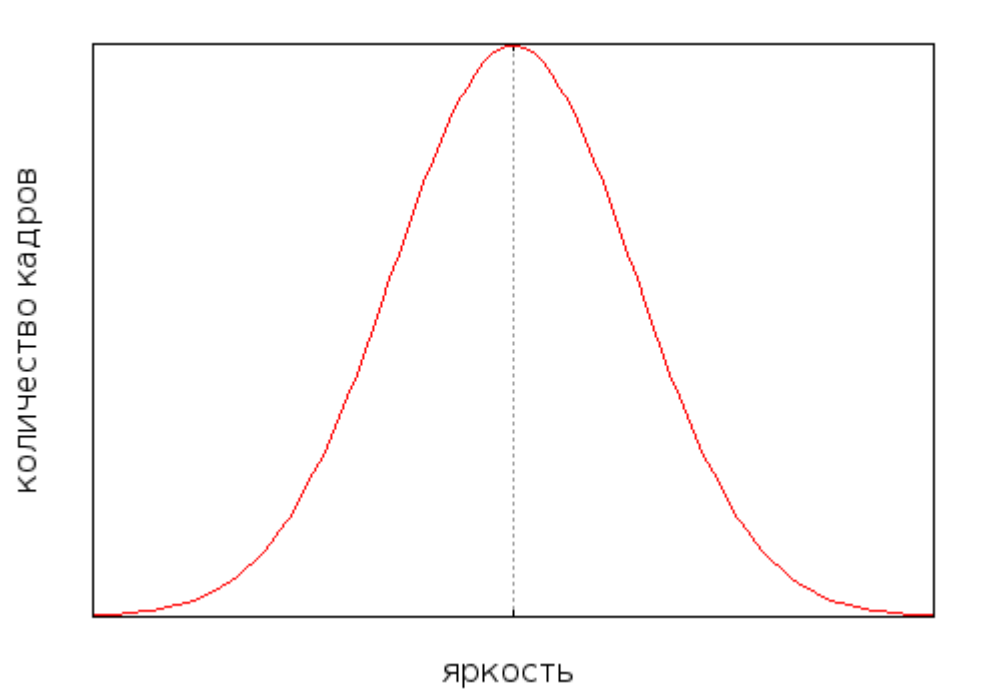
$$a(x) = \underset{y \in Y}{argmax}\ \lambda_y P(y)\, p(x|y)$$



$$p(x|y) = \sum_k w_k^y \varphi_k^y(x;\theta_k^y) = \sum_k w_k^y N(x;\Sigma_k^y,\mu_k^y)$$

Восстановление плотности

Пример: детектор новых объектов для неподвижных камер



Восстановление плотности: 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*
- Борисов Е.С. Классификатор на основе RBF. *http://mechanoid.kiev.ua/ml-rbf.html*
- *Борисов Е.С. Детектор объектов для неподвижных камер. http://mechanoid.kiev.ua/cv-backgr.html*

Восстановление плотности

Вопросы ?

Восстановление плотности: практика

OpenCV MOG

источники данных для экспериментов

sklearn.datasets

UCI Repository

kaggle