# Методы восстановления плотности распределения

Евгений Борисов

# Восстановление плотности

**Вероятностное пространство** математическая модель случайного эксперимента (опыта)

$$(\Omega, A, P)$$

**Ω** - элементарные исходы эксперимента, множество объектов ω.

**A** - случайные события, набор подмножеств **Ω**.

$A \ni \Omega$ - достоверное событие

$A \ni \varnothing$ - невозможное событие

**P** - функция вероятности $P: A \longrightarrow [0,1]$

$0 \leqslant P(a) \leqslant 1$ вероятность события **a** из **A**

$P(\varnothing)=0$ ; $P(\Omega)=1$

**Случайная величина** в пространстве **(Ω,A,P)** это числовая функция

$$X: A \longrightarrow \mathbb{R}$$

**типы случайных величин:**

<u>дискретные (discrete)</u> - принимающая конечное или счетное число значений
(*<u>Пример</u>: частота слов в тексте, количество детей в семье*)

<u>непрерывные (continuous)</u> - принимают значение в определённом интервале
(*<u>Пример</u>: рост людей*)

# Восстановление плотности

Вероятностное пространство **(Ω,A,P)**
**Ω** - элементарные исходы эксперимента
**A** - случайные события, набор подмножеств **Ω**
**P** - функция вероятности **P: A ⟶ [0,1]**
**X** - случайная величина **X: A ⟶** $\mathbb{R}$

случайная величина **X** задаётся распределением вероятностей **F** своих значений

**F(x) = P( X ⩽ x )**

Рассмотрим интервалы **(x,x+Δx)**, где **Δx** - бесконечно малые приращения **x** для **F(x)**

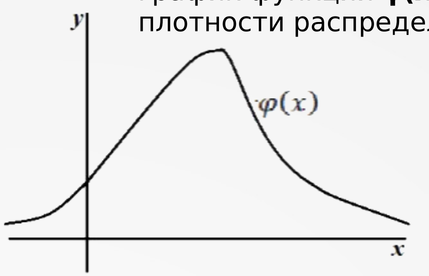
$$\varphi(x)=\lim_{\Delta x\to 0}\frac{F(x+\Delta x)-F(x)}{\Delta x}$$

**Плотностью распределения** (вероятности) **φ(x)** непрерывной случайной величины **X** назовём первую производную функции распределения **F(x)**

# Восстановление плотности

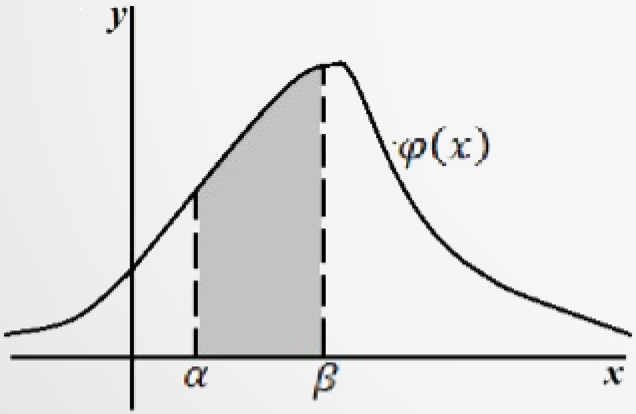
<u>Вероятностное пространство</u> **(Ω,A,P)**
**Ω** - элементарные исходы эксперимента
**A** - случайные события, набор подмножеств **Ω**
**P** - функция вероятности **P: A ⟶ [0,1]**
**X** - случайная величина **X: A ⟶** ℝ
**F(x)=P(X⩽x)** - распределение вероятностей **P**
**φ(x)=F`(x)** - плотность распределения **F**

график функции **φ(x)**
плотности распределения **X**



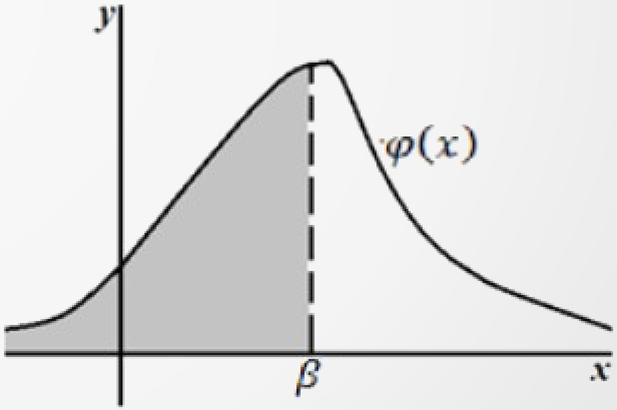

$$P(\Omega)=\int_{-\infty}^{+\infty}\varphi(x)\,dx=1$$

**площадь** криволинейной трапеции, ограниченной графиком **φ(x)** и прямыми **x=a, x=b, y=0** это вероятность **P(a⩽X⩽b)** попадания **X** в интервал **[a,b]**




$$P(\alpha\leqslant X\leqslant\beta)=\int_{\alpha}^{\beta}\varphi(x)\,dx$$

**площадь** бесконечной криволинейной трапеции, ограниченной графиком **φ(x)**, прямой **x=b, y=0** это функция распределения **F(b)=P(X⩽b)**




$$F(\beta)=P(X\leqslant\beta)=\int_{-\infty}^{\beta}\varphi(x)\,dx$$

https://spravochnick.ru/matematika/plotnost_raspredeleniya_veroyatnostey_nepreryvnoy_sluchaynoy_velichiny/

# Байесовский классификатор

**принцип максимума апостериорной вероятности**

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

**формула Байеса**

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

**байесовский классификатор**

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

$\lambda_y$ - потеря для объектов y

$P(y)$ - доля примеров класса y (априорная вероятность)

$p(x|y)$ - плотность  класса y

# Восстановление плотности

**подходы к оценке плотности распределения**

- непараметрический $$\hat{p}(x)=\frac{1}{mV(h)}\sum_{j=1}^{m}K\left(\frac{\rho(x,x_j)}{h}\right)$$
- параметрический $$\hat{p}(x)=\varphi(x,\theta)$$
- **смеси распределений** $$\hat{p}(x)=\sum_{j=1}^{k}w_j\varphi_j(x,\theta_j)$$

# Восстановление плотности

**смеси распределений**

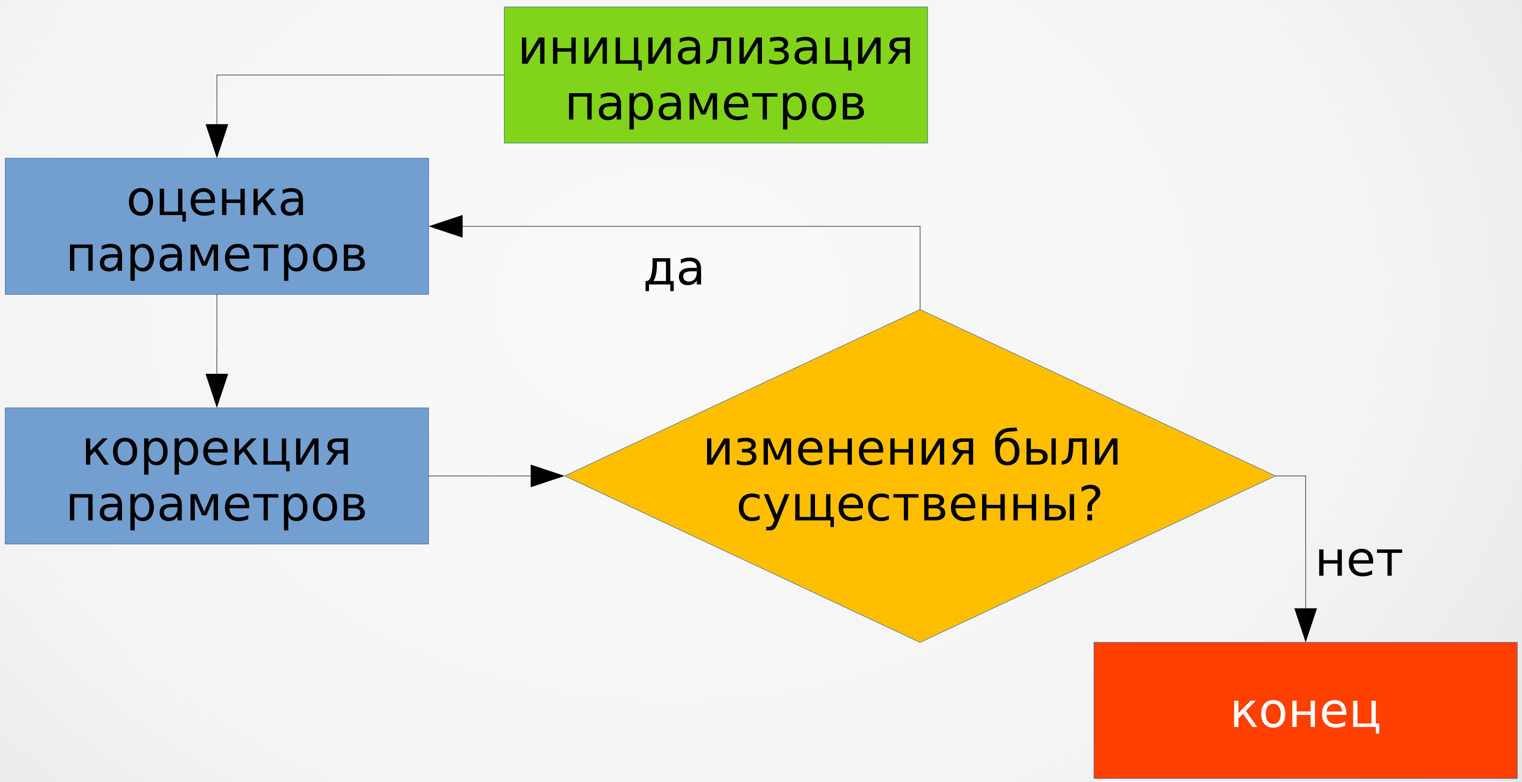
$$\hat{p}(x)=\sum_{j=1}^{k} w_j \varphi_j(x,\theta_j);$$

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

# Восстановление плотности

**EM (expectation–maximization algorithm)**:

базовый вариант алгоритма



# Восстановление плотности

**EM (expectation–maximization algorithm)**

оценка

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

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

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

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

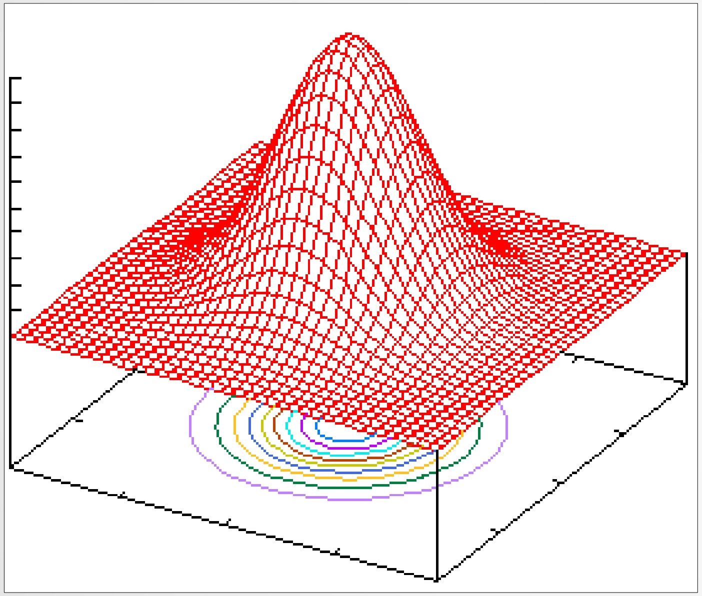
i=1...m
m - количество примеров X
s - количество компонент смеси

условие остановки: параметры не изменились

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

# Восстановление плотности

**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

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

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

задача:

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

# 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:

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

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

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:

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

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

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

вес компоненты

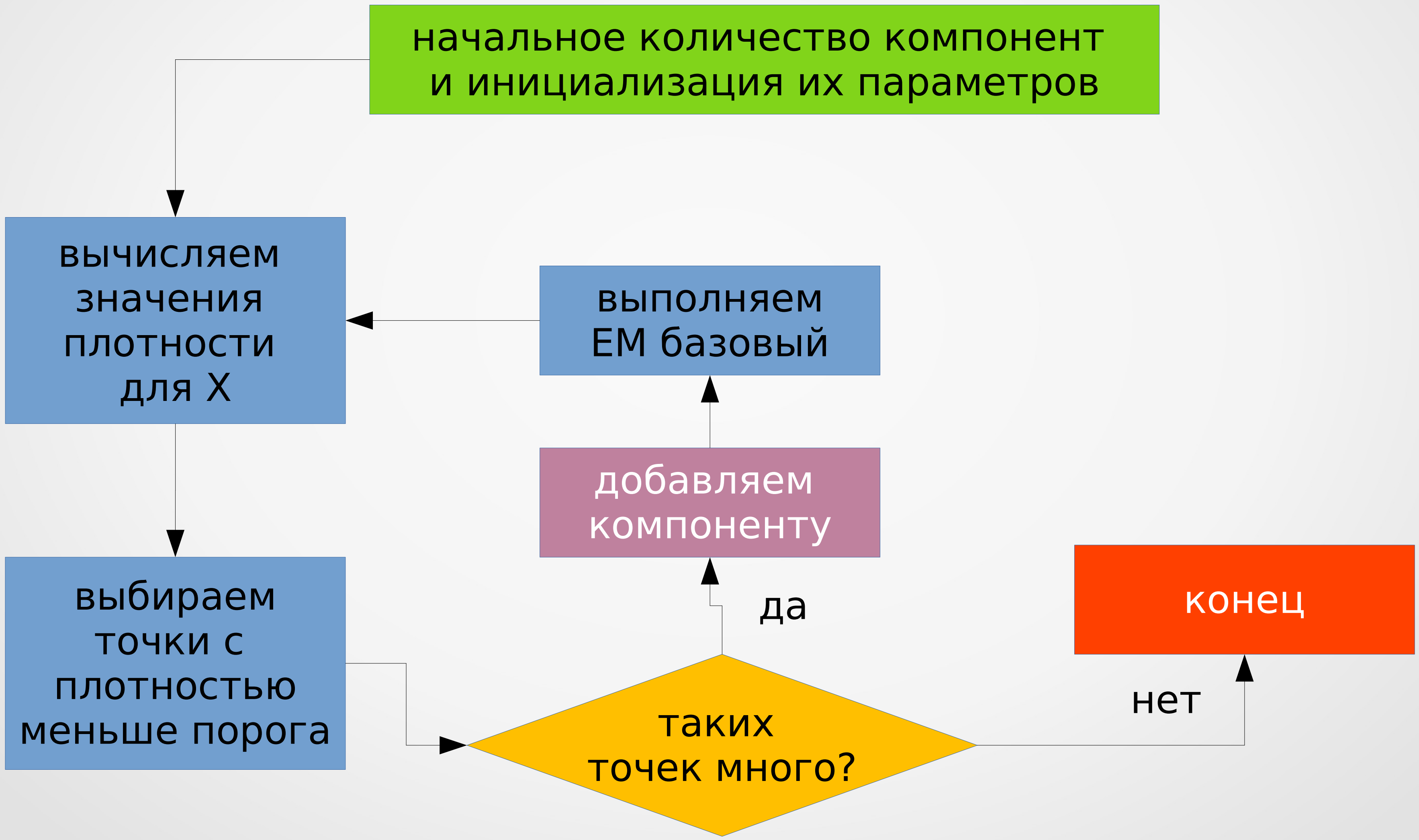
мат.ожидание компоненты

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

матрица ковариаций компоненты

# Восстановление плотности

**ЕМ 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

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

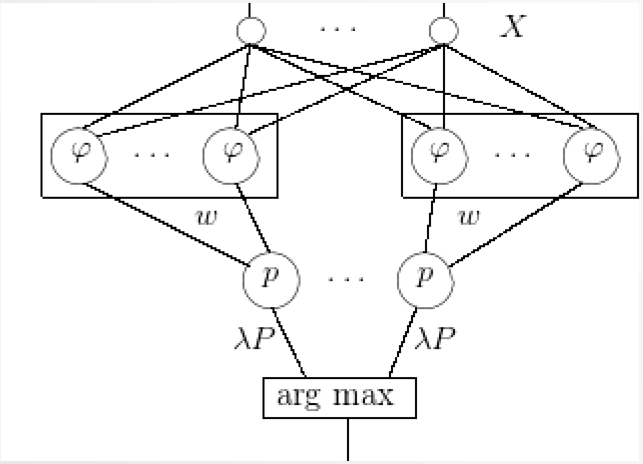
после определения новых параметров смеси запускаем EM

# Восстановление плотности

**RBF -** сеть радиальных базисных функций

Байесовский классификатор

плотности классов - смеси нормальных распределений



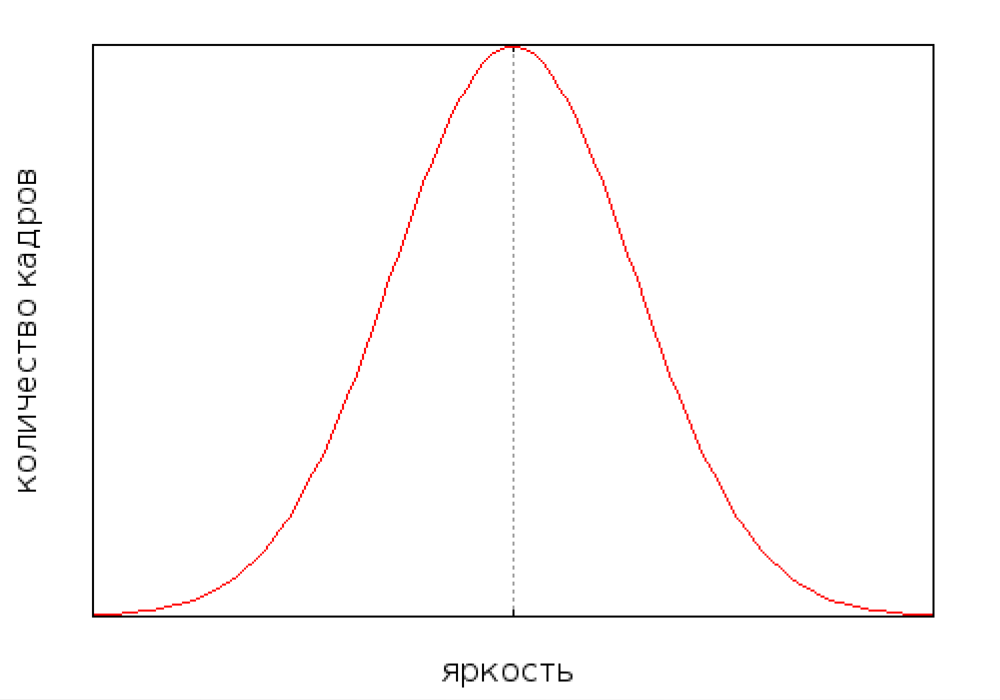

$$a(x) = \underset{y \in Y}{argmax}\, \lambda_y P(y)\, p(x|y)$$

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

http://www.machinelearning.ru/wiki/index.php?title=Сеть_радиальных_базисных_функций

# Восстановление плотности

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




# Литература

Борисов Е.С. Методы машинного обучения. 2024
https://github.com/mechanoid5/ml_lectorium_2024_I

Константин Воронцов   Машинное обучение. ШАД Яндекс
https://www.youtube.com/playlist?list=PLJOzdkh8T5kp99tGTEFjH_b9zqEQiiBtC

Константин Воронцов   Машинное_обучение. курс_лекций.
http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение_(курс_лекций,_К.В.Воронцов)

SciKit-Learn : Naive Bayes
https://scikit-learn.org/stable/modules/naive_bayes.html

Евгений Борисов   Детектор объектов для неподвижных камер.
http://mechanoid.su/cv-backgr.html