

# Байесовская теория классификации и методы восстановления плотности

Воронцов Константин Вячеславович

vokov@forecsys.ru

<http://www.MachineLearning.ru/wiki?title=User:Vokov>

Этот курс доступен на странице вики-ресурса

<http://www.MachineLearning.ru/wiki>

«Машинное обучение (курс лекций, К.В.Воронцов)»

Видеолекции: <http://shad.yandex.ru/lectures>

26 марта 2018

- 1 Оптимальный байесовский классификатор**
  - Вероятностная постановка задачи классификации
  - Задача восстановления плотности распределения
  - Наивный байесовский классификатор
- 2 Восстановление плотности вероятности**
  - Непараметрическое восстановление плотности
  - Параметрическое восстановление плотности
  - Проблема мультиколлинеарности
- 3 Разделение смеси распределений**
  - EM-алгоритм
  - Разделение гауссовских смесей
  - Сеть радиальных базисных функций

## Постановка задачи

$X$  — объекты,  $Y$  — ответы,  $X \times Y$  — в.п. с плотностью  $p(x, y)$ ;

**Дано:**  $X^\ell = (x_i, y_i)_{i=1}^\ell \sim p(x, y)$  — простая выборка (i.i.d.);

**Найти:**  $a: X \rightarrow Y$  с минимальной вероятностью ошибки.

**Временное допущение:** пусть известна совместная плотность

$$p(x, y) = p(x)P(y|x) = P(y)p(x|y).$$

$P(y)$  — априорная вероятность класса  $y$ ;

$p(x|y)$  — функция правдоподобия класса  $y$ ;

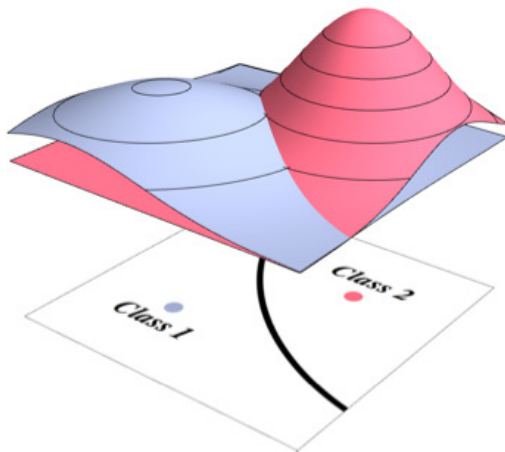
$P(y|x)$  — апостериорная вероятность класса  $y$ ;

**Принцип максимума апостериорной вероятности:**

$$a(x) = \arg \max_{y \in Y} P(y|x) = \arg \max_{y \in Y} P(y)p(x|y).$$

## Классификация по максимуму функции правдоподобия

Частный случай:  $a(x) = \arg \max_{y \in Y} p(x|y)$  при равных  $P(y)$ .



## Оптимальный байесовский классификатор

### Теорема

Пусть  $P(y)$  и  $p(x|y)$  известны,  $\lambda_y \geq 0$  — потеря от ошибки на объекте класса  $y \in Y$ . Тогда минимум среднего риска

$$R(a) = \sum_{y \in Y} \lambda_y \int [a(x) \neq y] p(x, y) dx$$

достигается *байесовским классификатором*

$$a(x) = \arg \max_{y \in Y} \lambda_y P(y) p(x|y).$$

## Две подзадачи, причём вторая уже решена!

- 1 Восстановление плотности распределения по выборке  
**Дано:**  $X^\ell = (x_i, y_i)_{i=1}^\ell$  — обучающая выборка.  
**Найти:** эмпирические оценки  $\hat{P}(y)$  и  $\hat{p}(x|y)$ ,  $y \in Y$
- 2 Построение классификатора  
**Дано:** вероятности  $P(y)$  и плотности  $p(x|y)$ ,  $y \in Y$ .  
**Найти:** классификатор  $a: X \times Y$ , минимизирующий  $R(a)$ .

**Замечание 1:** после замены  $P(y)$  и  $p(x|y)$  их эмпирическими оценками байесовский классификатор уже не оптимален.

**Замечание 2:** задача оценивания плотности распределения — более сложная, чем задача классификации.

## Наивный байесовский классификатор

**Допущение (действительно наивное):**

Признаки  $f_j: X \rightarrow D_j$  — независимые случайные величины с плотностями распределения,  $p_j(\xi_j|y)$ ,  $y \in Y$ ,  $j = 1, \dots, n$ .

Тогда функции правдоподобия классов представимы в виде произведения одномерных плотностей по признакам:

$$p(x|y) = p_1(\xi_1|y) \cdots p_n(\xi_n|y), \quad x = (\xi_1, \dots, \xi_n), \quad y \in Y.$$

Прологарифмируем (для удобства). Получим классификатор

$$a(x) = \arg \max_{y \in Y} \left( \ln \lambda_y \hat{P}(y) + \sum_{j=1}^n \ln \hat{p}_j(\xi_j|y) \right).$$

Восстановление  $n$  одномерных плотностей

— намного более простая задача, чем одной  $n$ -мерной.

## Начнём с определения плотности вероятности

**Дискретный случай:**  $|X| \ll m$ . Гистограмма значений  $x_i$ :

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

**Одномерный непрерывный случай:**  $X = \mathbb{R}$ . По определению плотности, если  $P[a, b]$  — вероятностная мера отрезка  $[a, b]$ :

$$p(x) = \lim_{h \rightarrow 0} \frac{1}{2h} P[x - h, x + h],$$

Эмпирическая оценка плотности по окну ширины  $h$   
(заменяем вероятность на долю объектов выборки):

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



## Локальная непараметрическая оценка Парзена-Розенблатта

Эмпирическая оценка плотности по окну ширины  $h$ :

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

Обобщение: оценка Парзена-Розенблатта по окну ширины  $h$ :

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

где  $K(r)$  — ядро, удовлетворяющее требованиям:

- чётная функция;
- нормированная функция:  $\int K(r) dr = 1$ ;
- невозрастающая при  $r > 0$ , неотрицательная функция.

В частности, при  $K(r) = \frac{1}{2} [ |r| < 1 ]$  имеем эмпирическую оценку.

## Метод парзеновского окна (Parzen window)

Многомерное обобщение:  $\rho(x, x')$  — метрика на  $X$ .

Парзеновская оценка плотности для каждого класса  $y \in Y$ :

$$\hat{p}_h(x|y) = \frac{1}{\ell_y V(h)} \sum_{i: y_i=y} K\left(\frac{\rho(x, x_i)}{h}\right),$$

Метод окна Парзена — это метрический классификатор:

$$a(x; X^\ell, h) = \arg \max_{y \in Y} \lambda_y \frac{P(y)}{\ell_y} \sum_{i: y_i=y} K\left(\frac{\rho(x, x_i)}{h}\right).$$

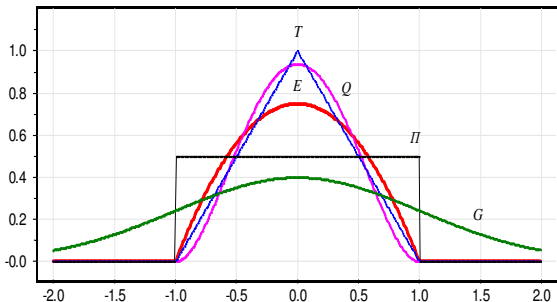
**Замечание 1:** нормирующий множитель

$V(h) = \int_X K\left(\frac{\rho(x, x_i)}{h}\right) dx$  не должен зависеть от  $x_i$  и  $y_i$ .

**Замечание 2:** имеем проблемы выбора

ядра  $K(r)$ , ширины окна  $h$ , функции расстояния  $\rho(x, x')$ .

## Выбор ядра



$E(r) = \frac{3}{4}(1 - r^2)[|r| \leq 1]$  — оптимальное (Епанечникова);

$Q(r) = \frac{15}{16}(1 - r^2)^2[|r| \leq 1]$  — квартическое;

$T(r) = (1 - |r|)[|r| \leq 1]$  — треугольное;

$G(r) = (2\pi)^{-1/2} \exp(-\frac{1}{2}r^2)$  — гауссовское;

$\Pi(r) = \frac{1}{2}[|r| \leq 1]$  — прямоугольное.

## Выбор ядра почти не влияет на качество восстановления

Функционал качества восстановления плотности:

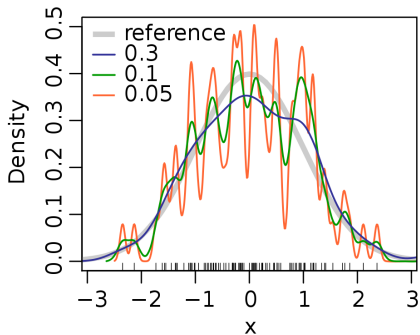
$$J(K) = \int_{-\infty}^{+\infty} E(\hat{p}_h(x) - p(x))^2 dx.$$

ядро $K(r)$	степень гладкости	$J(K^*)/J(K)$
Епанечникова $K^*(r)$	$\hat{p}'_h$ разрывна	1.000
Квартическое	$\hat{p}''_h$ разрывна	0.995
Треугольное	$\hat{p}'_h$ разрывна	0.989
Гауссовское	$\infty$ дифференцируема	0.961
Прямоугольное	$\hat{p}_h$ разрывна	0.943

**Замечание:** в таблице представлены асимптотические значения отношения  $J(K^*)/J(K)$  при  $m \rightarrow \infty$ , причём это отношение не зависит от  $p(x)$ .

## Пример. Зависимость оценки плотности от ширины окна

Оценка  $\hat{p}_h(x)$  существенно зависит от ширины окна  $h$ :

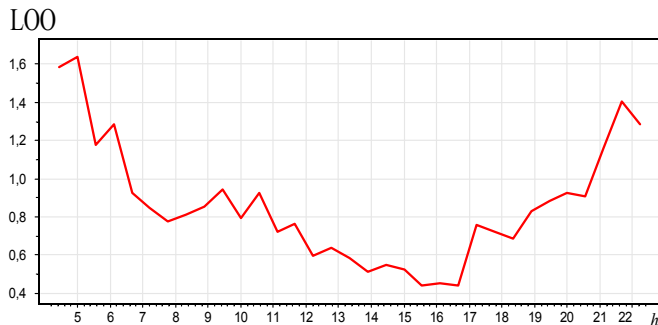


## Выбор ширины окна

Скольльзящий контроль *Leave One Out*:

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

Типичный вид зависимости  $LOO(h)$ :



## Окна переменной ширины

### Проблема:

при наличии локальных сгущений любая  $h$  не оптимальна.

### Идея:

задавать не ширину окна  $h$ , а число соседей  $k$ .

$$h_k(x) = \rho(x, x^{(k+1)}),$$

где  $x^{(i)}$  —  $i$ -й сосед объекта  $x$  при ранжировании выборки  $X^\ell$ :

$$\rho(x, x^{(1)}) \leq \dots \leq \rho(x, x^{(\ell)}).$$

**Замечание 1:** нормировка  $V(h_k)$  не должна зависеть от  $y$ , поэтому выборка ранжируется целиком, а не по классам  $X_y$ .

**Замечание 2:** оптимизация  $LOO(k)$  аналогична  $LOO(h)$ .

## Принцип максимума правдоподобия

Задана параметрическая модель плотности

$$p(x) = \varphi(x; \theta),$$

где  $\theta$  — параметр,  $\varphi$  — фиксированная функция.

Найдём оптимальное  $\theta$  по i.i.d. выборке  $X^m = \{x_1, \dots, x_m\}$ .

**Принцип максимума правдоподобия:**

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

**Необходимое условие оптимума:**

$$\frac{\partial}{\partial \theta} L(\theta; X^m) = \sum_{i=1}^m \frac{\partial}{\partial \theta} \ln \varphi(x_i; \theta) = 0,$$

где функция  $\varphi(x; \theta)$  достаточно гладкая по параметру  $\theta$ .



## Многомерное нормальное распределение

Пусть  $X = \mathbb{R}^n$  — объекты описываются  $n$  числовыми признаками.

**Гипотеза:** классы имеют  $n$ -мерные гауссовские плотности:

$$p(x|y) = \mathcal{N}(x; \mu_y, \Sigma_y) = \frac{e^{-\frac{1}{2}(x-\mu_y)^T \Sigma_y^{-1}(x-\mu_y)}}{\sqrt{(2\pi)^n \det \Sigma_y}}, \quad y \in Y,$$

где  $\mu_y \in \mathbb{R}^n$  — вектор матожидания (центр) класса  $y \in Y$ ,  
 $\Sigma_y \in \mathbb{R}^{n \times n}$  — ковариационная матрица класса  $y \in Y$   
(симметричная, невырожденная, положительно определённая).

### Теорема

1. Разделяющая поверхность

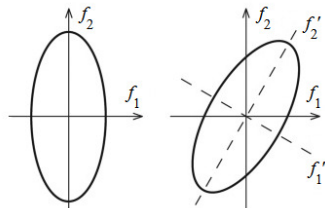
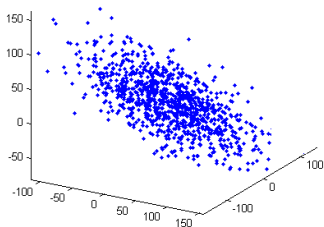
$$\{x \in X \mid \lambda_y P(y)p(x|y) = \lambda_s P(s)p(x|s)\}$$

квадратична для всех  $y, s \in Y$ ,  $y \neq s$ .

2. Если  $\Sigma_y = \Sigma_s$ , то она вырождается в линейную.

## Геометрический смысл предположения о нормальности классов

Каждый класс — облако точек эллиптической формы:



Если  $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ : оси эллипсоида параллельны осям

В общем случае:  $\Sigma = VSV^T$  — спектральное разложение,

$V = (v_1, \dots, v_n)$  — ортогональные собственные векторы  $\Sigma$ ,

$S = \text{diag}(\lambda_1, \dots, \lambda_n)$  — собственные значения

$$(x - \mu)^T \Sigma^{-1} (x - \mu) = (x - \mu)^T V S^{-1} V^T (x - \mu) = (x' - \mu')^T S^{-1} (x' - \mu').$$

$x' = V^T x$  — декоррелирующее ортогональное преобразование

## Квадратичный дискриминант

### Теорема

Оценки максимального правдоподобия для  $n$ -мерных гауссовских плотностей классов  $y \in Y$ :

$$\hat{\mu}_y = \frac{1}{\ell_y} \sum_{i: y_i=y} x_i;$$

$$\hat{\Sigma}_y = \frac{1}{\ell_y} \sum_{i: y_i=y} (x_i - \hat{\mu}_y)(x_i - \hat{\mu}_y)^\top.$$

Квадратичный дискриминант — подстановочный алгоритм:

$$a(x) = \arg \max_{y \in Y} \left( \ln \lambda_y P(y) - \frac{1}{2} (x - \hat{\mu}_y)^\top \hat{\Sigma}_y^{-1} (x - \hat{\mu}_y) - \frac{1}{2} \ln \det \hat{\Sigma}_y \right).$$

## Линейный дискриминант Фишера

Допущение:

ковариационные матрицы классов равны:  $\Sigma_y = \Sigma$ ,  $y \in Y$ .

Оценка максимума правдоподобия для  $\Sigma$ :

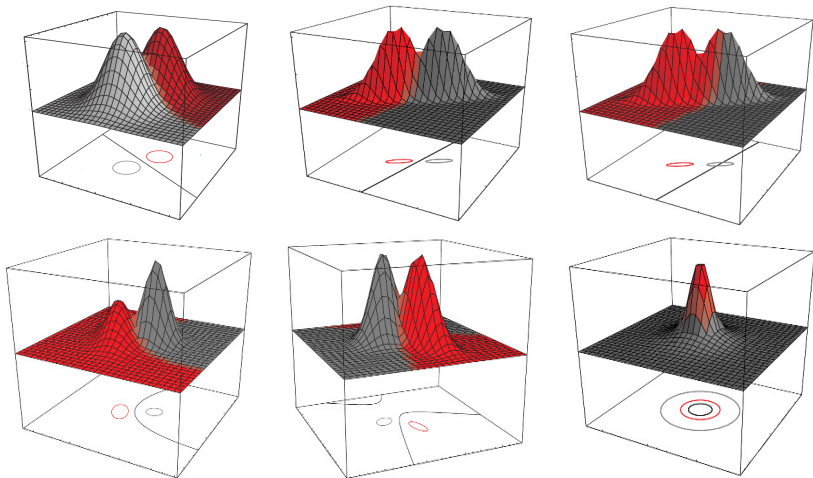
$$\hat{\Sigma} = \frac{1}{\ell} \sum_{i=1}^{\ell} (x_i - \hat{\mu}_{y_i})(x_i - \hat{\mu}_{y_i})^T$$

Линейный дискриминант — подстановочный алгоритм:

$$\begin{aligned} a(x) &= \arg \max_{y \in Y} \lambda_y \hat{P}(y) \hat{p}(x|y) = \\ &= \arg \max_{y \in Y} \underbrace{(\ln(\lambda_y \hat{P}(y)) - \frac{1}{2} \hat{\mu}_y^T \hat{\Sigma}^{-1} \hat{\mu}_y)}_{\beta_y} + x^T \underbrace{\hat{\Sigma}^{-1} \hat{\mu}_y}_{\alpha_y}; \end{aligned}$$

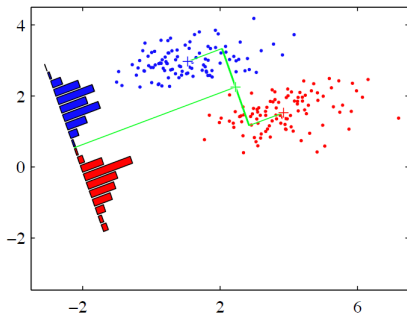
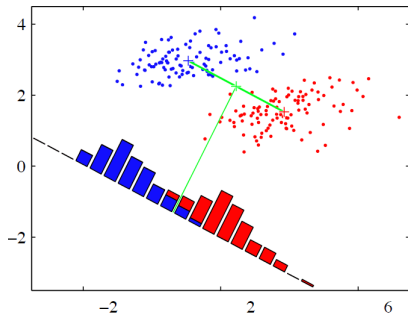
$$a(x) = \arg \max_{y \in Y} (x^T \alpha_y + \beta_y).$$

## Геометрический смысл квадратичного дискриминанта



## Геометрический смысл линейного дискриминанта

В одномерной проекции на направляющий вектор разделяющей гиперплоскости классы разделяются наилучшим образом, то есть с минимальной вероятностью ошибки.



## Проблема мультиколлинеарности

Матрица  $\hat{\Sigma}_y$  вырождена при  $\ell_y < n$   
 и может быть плохо обусловлена при  $\ell_y \geq n$

- Регуляризация ковариационной матрицы:
  - 1) обращение  $\hat{\Sigma} + \tau I_n$  вместо  $\hat{\Sigma}$
  - 2) выбор параметра  $\tau$  по скользящему контролю
- Диагонализация ковариационной матрицы,  
*нормальный наивный байесовский классификатор:*

$$a(x) = \arg \max_{y \in Y} \left( \ln \lambda_y \hat{P}(y) + \sum_{j=1}^n \ln \hat{p}_j(\xi_j | y) \right), \quad x \equiv (\xi_1, \dots, \xi_n);$$

$$\hat{p}_j(\xi | y) = \frac{1}{\sqrt{2\pi\hat{\sigma}_{yj}}} \exp\left(-\frac{(\xi - \hat{\mu}_{yj})^2}{2\hat{\sigma}_{yj}^2}\right), \quad y \in Y, \quad j = 1, \dots, n;$$

$\hat{\mu}_{yj}$  и  $\hat{\sigma}_{yj}$  — оценки среднего и дисперсии признака  $j$  в  $X_y$ .

## Задача восстановления смеси распределений

Порождающая модель смеси распределений:

$$p(x) = \sum_{j=1}^k w_j \varphi(x, \theta_j), \quad \sum_{j=1}^k w_j = 1, \quad w_j \geq 0,$$

$k$  — число компонент смеси;

$\varphi(x, \theta_j) = p(x|j)$  — функция правдоподобия  $j$ -й компоненты;

$w_j = P(j)$  — априорная вероятность  $j$ -й компоненты.

**Задача 1:** при фиксированном  $k$ ,

имея простую выборку  $X^m = \{x_1, \dots, x^m\} \sim p(x)$ ,

оценить вектор параметров  $(w, \theta) = (w_1, \dots, w_k, \theta_1, \dots, \theta_k)$ .

**Задача 2:** оценить ещё и  $k$ .



## Максимизация правдоподобия и EM-алгоритм

Задача максимизации логарифма правдоподобия

$$L(w, \theta) = \ln \prod_{i=1}^m p(x_i) = \sum_{i=1}^m \ln \sum_{j=1}^k w_j \varphi(x_i, \theta_j) \rightarrow \max_{w, \theta}.$$

при ограничениях  $\sum_{j=1}^k w_j = 1$ ;  $w_j \geq 0$ .

**Итерационный алгоритм Expectation–Maximization:**

- 1: начальное приближение параметров  $(w, \theta)$ ;
- 2: **повторять**
- 3: оценка скрытых переменных  $G = (g_{ij})$ ,  $g_{ij} = P(j|x_i)$ :  
 $G := E\text{-шаг}(w, \theta)$ ;
- 4: максимизация правдоподобия, отдельно по компонентам:  
 $(w, \theta) := M\text{-шаг}(w, \theta, G)$ ;
- 5: **пока**  $w, \theta$  и  $G$  не стабилизируются.

## EM-алгоритм как способ решения системы уравнений

### Теорема (необходимые условия экстремума)

Точка  $(w_j, \theta_j)_{j=1}^k$  локального экстремума  $L(w, \theta)$  удовлетворяет системе уравнений относительно  $w_j, \theta_j$  и  $g_{ij}$ :

$$\text{E-шаг: } g_{ij} = \frac{w_j \varphi(x_i, \theta_j)}{\sum_{s=1}^k w_s \varphi(x_i, \theta_s)}, \quad i = 1, \dots, m, \quad j = 1, \dots, k;$$

$$\text{M-шаг: } \theta_j = \arg \max_{\theta} \sum_{i=1}^m g_{ij} \ln \varphi(x_i, \theta), \quad j = 1, \dots, k;$$

$$w_j = \frac{1}{m} \sum_{i=1}^m g_{ij}, \quad j = 1, \dots, k.$$

EM-алгоритм — это метод простых итераций для её решения

## Вероятностная интерпретация

**E-шаг** — это формула Байеса:

$$g_{ij} = P(j|x_i) = \frac{P(j)p(x_i|j)}{p(x_i)} = \frac{w_j\varphi(x_i, \theta_j)}{p(x_i)} = \frac{w_j\varphi(x_i, \theta_j)}{\sum_{s=1}^k w_s\varphi(x_i, \theta_s)}.$$

Очевидно, выполнено условие нормировки:  $\sum_{j=1}^k g_{ij} = 1$ .

**M-шаг** — это максимизация взвешенного правдоподобия, с весами объектов  $g_{ij}$  для  $j$ -й компоненты смеси:

$$\theta_j = \arg \max_{\theta} \sum_{i=1}^m g_{ij} \ln \varphi(x_i, \theta),$$

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

## Доказательство. Условия Каруша–Куна–Таккера

Лагранжиан оптимизационной задачи « $L(w, \theta) \rightarrow \max$ »:

$$\mathcal{L}(w, \theta) = \sum_{i=1}^m \ln \left( \underbrace{\sum_{j=1}^k w_j \varphi(x_i, \theta_j)}_{p(x_i)} \right) - \lambda \left( \sum_{j=1}^k w_j - 1 \right).$$

Приравниваем нулю производные:

$$\frac{\partial \mathcal{L}}{\partial w_j} = 0 \quad \Rightarrow \quad \lambda = m; \quad w_j = \frac{1}{m} \sum_{i=1}^m \underbrace{\frac{w_j \varphi(x_i, \theta_j)}{p(x_i)}}_{g_{ij}} = \frac{1}{m} \sum_{i=1}^m g_{ij},$$

$$\frac{\partial \mathcal{L}}{\partial \theta_j} = \sum_{i=1}^m \underbrace{\frac{w_j \varphi(x_i, \theta_j)}{p(x_i)}}_{g_{ij}} \frac{\partial}{\partial \theta_j} \ln \varphi(x_i, \theta_j) = \frac{\partial}{\partial \theta_j} \sum_{i=1}^m g_{ij} \ln \varphi(x_i, \theta_j) = 0.$$

## EM-алгоритм

**Вход:**  $X^m = \{x_1, \dots, x_m\}$ ,  $k$ ,  $\delta$ , начальные  $(w_j, \theta_j)_{j=1}^k$ ;

**Выход:**  $(w_j, \theta_j)_{j=1}^k$  — параметры смеси распределений

1: **повторять**

2: E-шаг (expectation):

для всех  $i = 1, \dots, m$ ,  $j = 1, \dots, k$

$$g_{ij}^0 := g_{ij}; \quad g_{ij} := \frac{w_j \varphi(x_i, \theta_j)}{\sum_{s=1}^k w_s \varphi(x_i, \theta_s)};$$

3: M-шаг (maximization):

для всех  $j = 1, \dots, k$

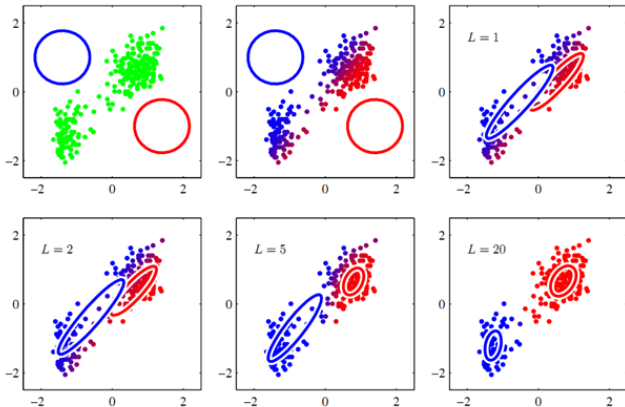
$$\theta_j := \arg \max_{\theta} \sum_{i=1}^m g_{ij} \ln \varphi(x_i, \theta); \quad w_j := \frac{1}{m} \sum_{i=1}^m g_{ij};$$

4: **пока**  $\max_{i,j} |g_{ij} - g_{ij}^0| > \delta$ ;

5: **вернуть**  $(w_j, \theta_j)_{j=1}^k$ ;

## Пример

Две гауссовские компоненты  $k = 2$  в пространстве  $X = \mathbb{R}^2$ .  
Расположение компонент в зависимости от номера итерации  $L$ :



## EM-алгоритм с добавлением и удалением компонент

Проблемы базового варианта EM-алгоритма:

- Как выбирать начальное приближение?
- Как определять число компонент?
- Как ускорить сходимость?

Добавление и удаление компонент в EM-алгоритме:

- Если слишком много объектов  $x_i$  имеют слишком низкие правдоподобия  $p(x_i)$ , то создаём новую  $k+1$ -ю компоненту, по этим объектам строим её начальное приближение.
- Если у  $j$ -й компоненты слишком низкий  $w_j$ , удаляем её.

Регуляризация  $L(w, \theta) - \tau \sum_{j=1}^k \ln w_j \rightarrow \max:$

$$w_j \propto \left( \frac{1}{m} \sum_{i=1}^m g_{ij} - \tau \right)_+$$

## Гауссовская смесь с диагональными матрицами ковариации

Гауссовская смесь GMM — Gaussian Mixture Model

**Допущения:**

1. Функции правдоподобия классов  $p(x|y)$  представимы в виде смесей  $k_y$  компонент,  $y \in Y = \{1, \dots, M\}$ .
2. Компоненты имеют  $n$ -мерные гауссовские плотности с некоррелированными признаками:

$$\mu_{yj} = (\mu_{yj1}, \dots, \mu_{yjn}), \quad \Sigma_{yj} = \text{diag}(\sigma_{yj1}^2, \dots, \sigma_{yjn}^2), \quad j = 1, \dots, k_y:$$

$$p(x|y) = \sum_{j=1}^{k_y} w_{yj} p_{yj}(x), \quad p_{yj}(x) = \mathcal{N}(x; \mu_{yj}, \Sigma_{yj}),$$

$$\sum_{j=1}^{k_y} w_{yj} = 1, \quad w_{yj} \geq 0;$$



## Эмпирические оценки средних и дисперсий

Числовые признаки:  $f_d: X \rightarrow \mathbb{R}$ ,  $d = 1, \dots, n$ .

**Решение задачи M-шага:**

для всех классов  $y \in Y$  и всех компонент  $j = 1, \dots, k_y$ ,

$$w_{yj} = \frac{1}{\ell_y} \sum_{i: y_i=y} g_{yij}$$

для всех размерностей (признаков)  $d = 1, \dots, n$

$$\hat{\mu}_{yjd} = \frac{1}{\ell_y w_{yj}} \sum_{i: y_i=y} g_{yij} f_d(x_i);$$

$$\hat{\sigma}_{yjd}^2 = \frac{1}{\ell_y w_{yj}} \sum_{i: y_i=y} g_{yij} (f_d(x_i) - \hat{\mu}_{yjd})^2;$$

**Замечание:** компоненты «наивны», но смесь не «наивна».

## Байесовский классификатор

Подставим гауссовскую смесь в байесовский классификатор:

$$a(x) = \arg \max_{y \in Y} \underbrace{\lambda_y P_y}_{\Gamma_y(x)} \sum_{j=1}^{k_y} \underbrace{w_{yj} \mathcal{N}_{yj} \exp\left(-\frac{1}{2} \rho_{yj}^2(x, \mu_{yj})\right)}_{\rho_{yj}(x)}$$

$\mathcal{N}_{yj} = (2\pi)^{-\frac{n}{2}} (\sigma_{yj1} \cdots \sigma_{yjn})^{-1}$  — нормировочные множители;  
 $\rho_{yj}(x, \mu_{yj})$  — взвешенная евклидова метрика в  $X = \mathbb{R}^n$ :

$$\rho_{yj}^2(x, \mu_{yj}) = \sum_{d=1}^n \frac{1}{\sigma_{yjd}^2} (f_d(x) - \mu_{yjd})^2.$$

**Интерпретация** — как у метрического классификатора:

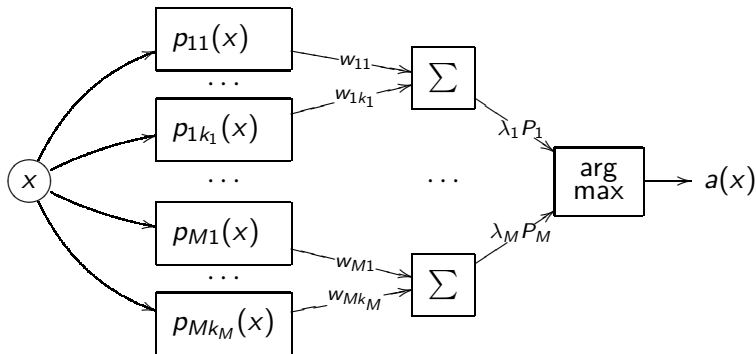
$\rho_{yj}(x)$  — близость объекта  $x$  к  $j$ -й компоненте класса  $y$ ;

$\Gamma_y(x)$  — близость объекта  $x$  к классу  $y$ .

## Байесовский классификатор — сеть RBF

Radial Basis Functions (RBF) — трёхуровневая суперпозиция:

$$a(x) = \arg \max_{y \in Y} \lambda_y P_y \sum_{j=1}^{k_y} w_{yj} p_{yj}(x)$$



## Преимущества RBF-EM

EM — один из лучших алгоритмов обучения радиальных сетей.

### Преимущества EM-алгоритма по сравнению с SVM:

- 1 EM-алгоритм легко сделать устойчивым к шуму
- 2 EM-алгоритм довольно быстро сходится
- 3 автоматически строится *структурное описание* каждого класса в виде совокупности компонент — *кластеров*

### Недостатки EM-алгоритма:

- 1 EM-алгоритм чувствителен к начальному приближению
- 2 Определение числа компонент — трудная задача (простые эвристики могут плохо работать)

- Эту формулу полезно помнить:  
$$a(x) = \arg \max_{y \in Y} \lambda_y P(y) p(x|y).$$
- Наивный байесовский классификатор:  
предположение о независимости признаков.  
Как ни странно, иногда это работает.
- Три подхода к восстановлению плотности  $p(x|y)$  по выборке:
  - *Параметрический подход:*  
модель плотности + максимизация правдоподобия;
  - *Непараметрический подход:*  
наиболее прост и приводит к методу парзеновского окна;
  - *Разделение смеси распределений:*  
в случае смеси гауссиан приводит к методу RBF.