

Методы машинного обучения. Байесовская теория классификации

Воронцов Константин Вячеславович

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

вопросы к лектору: voron@forecsys.ru

материалы курса:

github.com/MSU-ML-COURSE/ML-COURSE-23-24

орг.вопросы по курсу: ml.cmc@mail.ru

1 Байесовская теория классификации

- Задача минимизации вероятности ошибки
- Оптимальный байесовский классификатор
- Задачи эмпирического оценивания

2 Наивный байесовский классификатор

- Гипотеза о независимости признаков
- Линейный наивный байесовский классификатор
- Задачи классификации текстов

3 Обзор байесовских классификаторов

- Метод парзеновского окна
- Нормальный дискриминантный анализ
- Сеть радиальных базисных функций

Вероятностная постановка задачи классификации

X — объекты, Y — классы, $X \times Y$ — 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

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

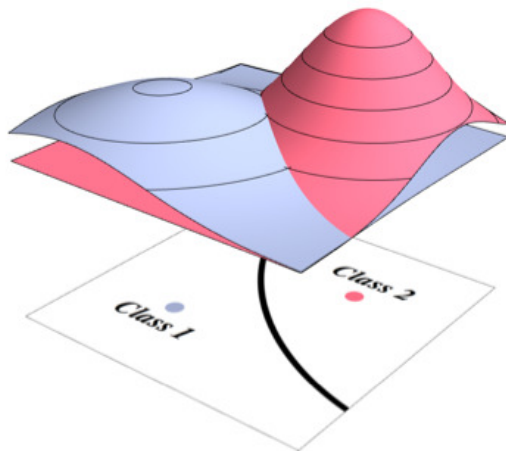
По формуле Байеса: $P(y|x) = \frac{P(y)p(x|y)}{p(x)}$

Байесовский классификатор:

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

Классификация по максимуму функции правдоподобия

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



Два подхода к обучению классификации

1 Дискриминативный (discriminative):

x — неслучайные векторы

$P(y|x, w)$ — модель классификации

Примеры: LR, GLM, SVM, RBF



2 Генеративный (generative):

$x \sim p(x|y)$ — случайные векторы

$p(x|y, \theta)$ — модель генерации данных

Примеры: NB, PW, FLD, RBF



Байесовские модели классификации — генеративные:

- моделируют форму классов не только вдоль границы, но и на всём пространстве, что избыточно для классификации
- требуют больше данных для обучения
- более устойчивы к шумовым выбросам

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 оптимальным байесовским классификатором

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

Замечание 1: после подстановки эмпирических оценок $\hat{P}(y)$ и $\hat{p}(x|y)$ байесовский классификатор уже не оптимален

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

Оптимальный байесовский классификатор (обобщение)

Теорема

Пусть $P(y)$ и $p(x|y)$ известны, $\lambda_{ys} \geq 0$ — потеря от ошибки на объекте класса $y \in Y$, когда $a(x) = s$.

Тогда минимум среднего риска

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

достигается оптимальным байесовским классификатором

$$a(x) = \arg \min_{s \in Y} \sum_{y \in Y} \lambda_{ys} P(y) p(x|y)$$

Предыдущая теорема — частный случай при $\lambda_{yy} = 0$ и $\lambda_{ys} \equiv \lambda_y$

Задачи эмпирического оценивания

Частотная оценка априорной вероятности:

$$\hat{P}(y) = \ell_y / \ell, \quad \ell_y = |X_y|, \quad X_y = \{x_i \in X: y_i = y\}$$

Оценки плотности $\hat{p}(x|y)$ по i.i.d. выборкам X_y , $y \in Y$:

❶ Параметрическая оценка плотности:

$$\hat{p}(x|y) = \varphi(x, \theta_y); \quad \theta_y = \arg \max_{\theta} \sum_{x_i \in X_y} \log \varphi(x_i, \theta)$$

❷ Непараметрическая оценка плотности:

$$\hat{p}(x|y) = \sum_{x_i \in X_y} \frac{1}{\ell V_h} K\left(\frac{\rho(x, x_i)}{h}\right)$$

❸ Восстановление смеси распределений:

$$\hat{p}(x|y) = \sum_{j=1}^k w_{yj} \varphi(x_i, \theta_{yj}); \quad (w_y, \theta_y) = \arg \max_{w, \theta} \sum_{x_i \in X_y} \log \hat{p}(x|y)$$

Наивный байесовский классификатор (Naïve Bayes)

Наивное предположение:

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

Тогда функции правдоподобия классов представимы в виде произведения одномерных плотностей по признакам, $x^j \equiv f_j(x)$:

$$p(x|y) = p_1(x^1|y) \cdots p_n(x^n|y), \quad x = (x^1, \dots, x^n), \quad y \in Y$$

Прологарифмировав под $\arg\max$, получим классификатор

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

Восстановление n одномерных плотностей

— намного более простая задача, чем одной n -мерной

Признаки с плотностями экспоненциального вида

Предположение: одномерные плотности экспоненциальны:

$$p(x^j|y; \theta_{yj}, \varphi_{yj}) = \exp\left(\frac{x^j\theta_{yj} - c(\theta_{yj})}{\varphi_{yj}} + h(x^j, \varphi_{yj})\right)$$

где θ_{yj} , φ_{yj} — параметры, $c(\theta)$, $h(x, \varphi)$ — параметры-функции.

Задача максимизации log-правдоподобия

$$L(\theta, \varphi) = \sum_{j=1}^n \sum_{y \in Y} \left(\sum_{x_i \in X_y} \ln p(x_i^j|y; \theta_{yj}, \varphi_{yj}) \right) \rightarrow \max_{\theta, \varphi}$$

распадается на независимые подзадачи для каждого (y, j) :

$$\sum_{x_i \in X_y} \left(\frac{x_i^j\theta_{yj} - c(\theta_{yj})}{\varphi_{yj}} + h(x_i^j, \varphi_{yj}) \right) \rightarrow \max_{\theta_{yj}, \varphi_{yj}}$$

По θ_{yj} задача решается аналитически, по φ_{yj} — не всегда

Линейный наивный байесовский классификатор

Решение θ_{yj} — функция связи $g(\mu) = [c']^{-1}(\mu)$ от $\mu = \hat{E}(x^j|y)$:

$$\frac{\partial L}{\partial \theta_{yj}} = 0 \Rightarrow c'(\theta_{yj}) = \sum_{x_i \in X_y} \frac{x_i^j}{|X_y|} \equiv \bar{x}_{yj} \Rightarrow \theta_{yj} = [c']^{-1}(\bar{x}_{yj})$$

Решение φ_{yj} не всегда выражается из уравнения $\frac{\partial L}{\partial \varphi_{yj}} = 0$, но для распределений Пуассона, Бернулли, биномиального $\varphi_{yj} = 1$; для гауссовского распределения (и если φ_{yj} не зависит от y):

$$\frac{\partial L}{\partial \varphi_{yj}} = 0 \Rightarrow \varphi_{yj} = \frac{1}{\ell} \sum_{i=1}^{\ell} (x_i^j - \bar{x}_{yij})^2$$

В итоге Naïve Bayes оказывается линейным классификатором:

$$a(x) = \arg \max_{y \in Y} \left(\sum_{j=1}^n x^j \underbrace{\frac{\theta_{yj}}{\varphi_{yj}}}_{w_{yj}} + \underbrace{\ln(\lambda_y P(y)) - \sum_{j=1}^n \frac{c(\theta_{yj})}{\varphi_{yj}}}_{b_y} + \underbrace{h(x^j, \varphi_{yj})}_{\substack{\text{если от } y \\ \text{не зависит}}} \right)$$

Напоминание. Примеры экспоненциальных распределений

μ — параметр матожидания, $\theta = g(\mu)$ — функции связи:

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \exp\left(\frac{x\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2)\right)$$

$$\mu^x (1 - \mu)^{1-x} = \exp\left(x \ln \frac{\mu}{1-\mu} + \ln(1 - \mu)\right)$$

$$C_k^x \left(\frac{\mu}{k}\right)^x \left(1 - \frac{\mu}{k}\right)^{k-x} = \exp\left(x \ln \frac{\mu}{k-\mu} + k \ln(k - \mu) + \ln C_k^x - k \ln k\right)$$

$$\frac{1}{x!} e^{-\mu} \mu^x = \exp\left(x \ln(\mu) - \mu - \ln x!\right)$$

распределение	значения	$c(\theta)$	$c'(\theta)$	$[c']^{-1}(\mu)$	φ	$h(x, \varphi)$
нормальное	\mathbb{R}	$\frac{1}{2}\theta^2$	θ	μ	σ^2	$-\frac{x^2}{2\varphi} - \frac{\ln(2\pi\varphi)}{2}$
Бернулли	$\{0, 1\}$	$\ln(1 + e^\theta)$	$\frac{1}{1+e^{-\theta}}$	$\ln \frac{\mu}{1-\mu}$	1	0
биномиальное	$\{0, \dots, k\}$	$k \ln \frac{1+e^\theta}{k}$	$\frac{k}{1+e^{-\theta}}$	$\ln \frac{\mu}{k-\mu}$	1	$\ln C_k^x - k \ln k$
Пуассона	$\{0, 1, \dots\}$	e^θ	e^θ	$\ln \mu$	1	$-\ln x!$

Задачи классификации (категоризации) текстов

x — текстовый документ (последовательность слов)

$y \in Y$ — класс (тематическая категория или рубрика)

$j \in \{1, \dots, n\}$ — слова, n — число слов в словаре

$f_j(x_i) = x_i^j$ — частота (число вхождений) слова j в документе x ;

$p_j(x^j|y)$ — распределение Пуассона, экспоненциального вида

$\theta_{yj} = \ln \bar{x}_{yj}$ — оценка максимума правдоподобия, $\varphi_{yj} = 1$

Наивный байесовский классификатор — линейный, с весами θ_{yj} :

$$a(x) = \arg \max_{y \in Y} \left(\sum_{j=1}^n \theta_{yj} x^j + \underbrace{\ln(\lambda_y P(y)) - \bar{N}_y}_{b_y} \right),$$

$\bar{N}_y = \sum_{j=1}^n c(\theta_{yj}) = \sum_{j=1}^n \bar{x}_{yj}$ — средняя длина документов в классе y

Замечание: если \bar{x}_{yj} не зависит от y , то слово j не влияет на $a(x)$

Мультиномиальный наивный байесовский классификатор

$x = (j_1, \dots, j_{N_x})$ — текстовый документ, длиной N_x слов

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

$\pi_{yj} = p(j|y)$ — вероятность слова j в текстах класса y

$$\ln p(x|y) = \ln \prod_{t=1}^{N_x} p(j_t|y) = \sum_{j=1}^n \ln(\pi_{yj})^{x^j} = \sum_{j=1}^n x^j \ln \pi_{yj}$$

Частотная оценка (оценка максимума правдоподобия):

$$\pi_{yj} = \frac{\#count(y, j)}{\#count(y)} = \frac{\sum_{i \in X_y} x_i^j}{\sum_{j=1}^n \sum_{i \in X_y} x_i^j} = \frac{\bar{x}_{yj}}{\sum_{j=1}^n \bar{x}_{yj}} = \frac{\bar{x}_{yj}}{\bar{N}_y}$$

Тот же линейный NB, но с другой поправкой на длину текста:

$$a(x) = \arg \max_{y \in Y} \left(\sum_{j=1}^n x^j \ln \bar{x}_{yj} + \ln(\lambda_y P(y)) - N_x \ln \bar{N}_y \right)$$

Выводы про наивный байесовский классификатор

Достоинства:

- очень быстрое обучение за $O(\ell n)$ — вычисление \bar{x}_{yj} , φ_{yj}
- почти нет переобучения, даже на коротких выборках
- единообразная обработка разнотипных признаков
- хорошее начальное приближение для других методов
- оценка полезности и отбор признаков: $\max_y p(y|j)$
- базовый уровень качества при классификации текстов
- при классификации текстов отбор признаков по полезности удаляет стоп-слова, общую и нерелевантную лексику

Ограничения и недостатки:

- гипотеза о независимости признаков
- низкий уровень качества в большинстве приложений

Напоминание. Метод парзеновского окна (Parzen Window, PW)

Непараметрическая оценка плотности Парзена–Розенблатта с функцией расстояния $\rho(x, x')$, для каждого класса $y \in Y$:

$$\hat{p}_h(x|y) = \frac{1}{\ell_y V_h} \sum_{x_i \in X_y} K\left(\frac{\rho(x, x_i)}{h}\right),$$

Метод окна Парзена — это метрический классификатор:

$$a(x) = \arg \max_{y \in Y} \lambda_y \frac{P(y)}{\ell_y} \sum_{x_i \in X_y} K\left(\frac{\rho(x, x_i)}{h}\right).$$

Замечание 1: нормирующий множитель $V_h = \int_X K\left(\frac{\rho(x, x_i)}{h}\right) dx$ сокращается под $\arg\max$, если он не зависит от x_i и y_i .

Замечание 2 (напоминание): имеем проблемы выбора ядра $K(r)$, ширины окна h , функции расстояния $\rho(x, x')$.

Квадратичный дискриминант (Quadratic Discriminant Analysis)

Гипотеза: каждый класс $y \in Y$ имеет n -мерную гауссовскую плотность с центром μ_y и ковариационной матрицей Σ_y :

$$p(x|y) = \mathcal{N}(x; \mu_y, \Sigma_y) = \frac{\exp\left(-\frac{1}{2}(x - \mu_y)^\top \Sigma_y^{-1}(x - \mu_y)\right)}{\sqrt{(2\pi)^n \det \Sigma_y}}$$

Теорема

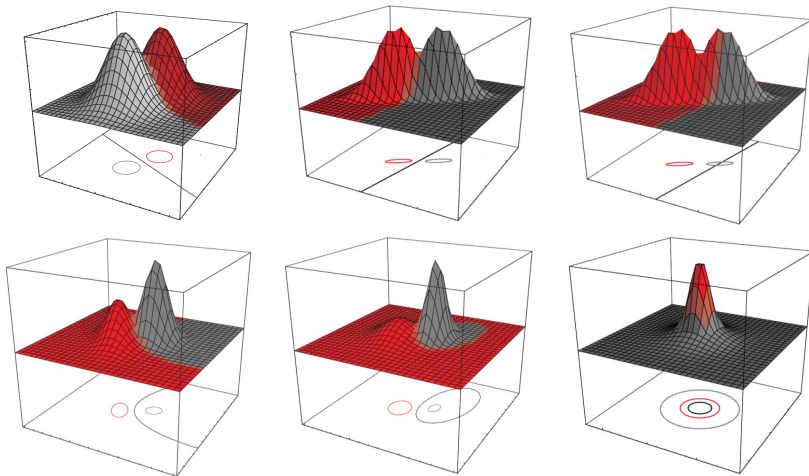
1. Разделяющая поверхность, определяемая уравнением $\lambda_y P(y) p(x|y) = \lambda_s P(s) p(x|s)$, квадратична для всех $y, s \in Y$.
2. Если $\Sigma_y = \Sigma_s$, то поверхность вырождается в линейную.

Квадратичный дискриминант — подстановочный алгоритм:

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

Геометрический смысл квадратичного дискриминанта

Разделяющая поверхность линейна ($\Sigma_y = \Sigma_s$) или квадратична:



Линейный дискриминант Фишера (Fisher Linear Discriminant)

Проблема: для малочисленных классов возможно $\det \hat{\Sigma}_y = 0$.

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

Оценка максимума правдоподобия для Σ :

$$\hat{\Sigma} = \frac{1}{\ell} \sum_{i=1}^{\ell} (x_i - \hat{\mu}_{y_i})(x_i - \hat{\mu}_{y_i})^T, \quad \hat{\mu}_y = \frac{\sum_i [y_i=y] x_i}{\sum_i [y_i=y]}$$

Линейный дискриминант — подстановочный алгоритм:

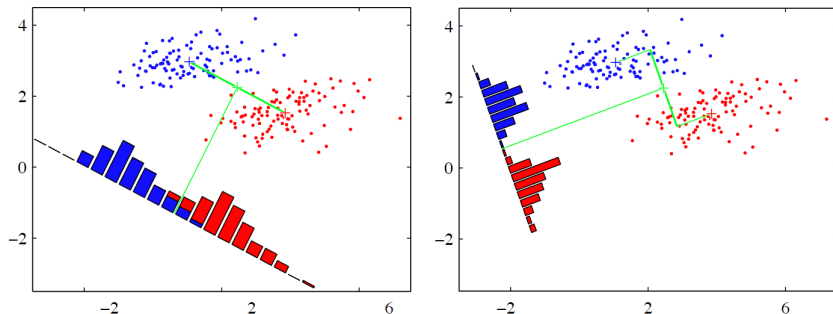
$$\begin{aligned} a(x) &= \arg \max_{y \in Y} \lambda_y \hat{P}(y) \hat{p}(x|y) = \\ &= \arg \max_{y \in Y} \underbrace{\left(\ln(\lambda_y \hat{P}(y)) - \frac{1}{2} \hat{\mu}_y^T \hat{\Sigma}^{-1} \hat{\mu}_y \right)}_{\beta_y} + x^T \underbrace{\hat{\Sigma}^{-1} \hat{\mu}_y}_{\alpha_y}; \end{aligned}$$

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

В случае мультиколлинеарности — обращать матрицу $\hat{\Sigma} + \tau I_n$.

Геометрическая интерпретация линейного дискриминанта

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



Ось проекции перпендикулярна общей касательной эллипсоидов рассеяния

Fisher R. A. The use of multiple measurements in taxonomic problems. 1936.

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, $y \in Y$
- 2 Компоненты $j = 1, \dots, k_y$ имеют n -мерные гауссовские плотности с некоррелированными признаками:
 $\mu_{yj} = (\mu_{yj1}, \dots, \mu_{yjn})$, $\Sigma_{yj} = \text{diag}(\sigma_{yj1}^2, \dots, \sigma_{yjn}^2)$:

$$p(x|y) = \sum_{j=1}^{k_y} w_{yj} p_{yj}(x), \quad p_{yj}(x) = \mathcal{N}(x; \mu_{yj}, \Sigma_{yj})$$
$$\sum_{j=1}^{k_y} w_{yj} = 1, \quad w_{yj} \geq 0$$

Напоминание: ЕМ-алгоритм, оценки средних и дисперсий

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

Е-шаг: для всех $y \in Y$, $j = 1, \dots, k_y$, $d = 1, \dots, n$:

$$g_{yij} = \frac{w_{yj} \mathcal{N}(x_i; \mu_{yj}, \Sigma_{yj})}{p(x_i|y)} \equiv P(j|x_i, y_i = y)$$

М-шаг: для всех $y \in Y$, $j = 1, \dots, k_y$, $d = 1, \dots, n$

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

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

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

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

Байесовский классификатор

Подставим гауссовскую смесь в байесовский классификатор:

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

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

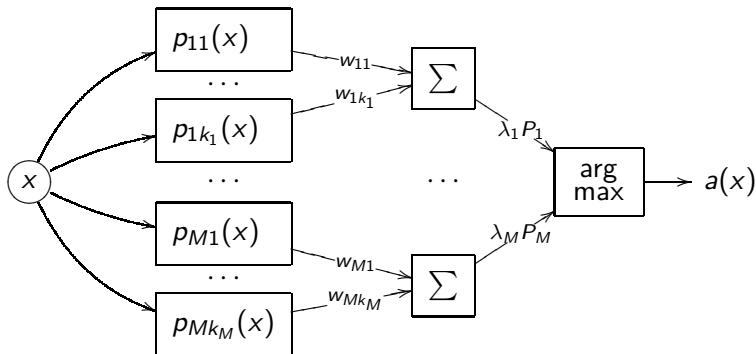
$$\rho_{yj}^2(x, \mu_{yj}) = \sum_{d=1}^n \frac{1}{\sigma_{yjd}^2} (f_d(x) - \mu_{yjd})^2.$$

Интерпретация: это метрический классификатор, в котором
 $p_{yj}(x)$ — близость объекта x к центру j -й компоненты класса y ;
 $\Gamma_y(x)$ — близость объекта x к классу y .

...он же — сеть радиальных базисных функций (RBF)

Трёхслойная сеть RBF (Radial Basis Functions):

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



ЕМ-алгоритм как метод обучения радиальных сетей

Отличия генеративного RBF-ЕМ от дискриминативного RBF-SVM:

- опорные векторы μ_{yj} — это не пограничные объекты выборки, а центры локальных сгущений (эталоны) классов
- автоматически строится *структурное описание* каждого класса в виде совокупности компонент — *кластеров*

Преимущества ЕМ-алгоритма:

- ЕМ-алгоритм легко сделать устойчивым к шуму
- как правило, ЕМ-алгоритм довольно быстро сходится

Недостатки ЕМ-алгоритма:

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

- Основная формула: $a(x) = \arg \max_{y \in Y} \lambda_y P(y)p(x|y)$
- Байесовские модели классификации — генеративные:
 - моделируют форму классов на всём пространстве,
 - требуют большего объёма данных для обучения,
 - менее чувствительны к шумовым выбросам
- Наивный байесовский классификатор
 - основан на предположении о независимости признаков,
 - неплохо работает в задачах категоризации текстов

Три подхода к восстановлению плотности $p(x|y)$ по выборке:

- *Параметрический подход:*
гауссовские классы \Rightarrow нормальный дискриминантный анализ
- *Непараметрический подход:*
задана функция расстояния \Rightarrow метод парзеновского окна
- *Разделение смеси распределений:*
классы описываются смесями гауссиан \Rightarrow сеть RBF