

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>

20 апреля 2017

- 1 Оптимальный байесовский классификатор**
 - Вероятностная постановка задачи классификации
 - Задача восстановления плотности распределения
 - Наивный байесовский классификатор
- 2 Непараметрическое восстановление плотности**
 - Одномерный случай
 - Многомерный случай
 - Метод парзеновского окна
- 3 Параметрическое восстановление плотности**
 - Принцип максимума правдоподобия
 - Нормальный дискриминантный анализ
 - Проблема мультиколлинеарности

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:

классификатор $a: X \rightarrow Y$ 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 ;

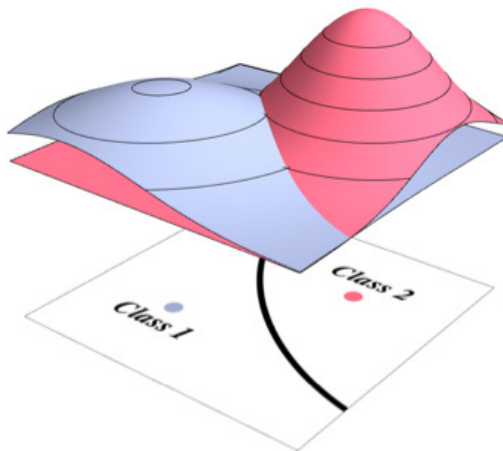
$P(y|x)$ — апостериорная вероятность класса y ;

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 $P(y)$.



Вероятность ошибки и функционал среднего риска

$a: X \rightarrow Y$ разбивает X на непересекающиеся области:

$$A_y = \{x \in X \mid a(x) = y\}, \quad y \in Y.$$

Ошибка: объект x класса y попадает в A_s , $s \neq y$.

Вероятность ошибки: $P(A_s, y) = \int_{A_s} p(x, y) dx$.

Потеря от ошибки: задана $\lambda_{ys} \geq 0$, для всех $(y, s) \in Y \times Y$.

Средний риск — мат.ожидание потери для классификатора a :

$$R(a) = \sum_{y \in Y} \sum_{s \in Y} \lambda_{ys} P(A_s, y),$$

Оптимальный байесовский классификатор

Теорема

Если известны $P(y)$ и $p(x|y)$, то минимум среднего риска $R(a)$ достигается *байесовским классификатором*

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

Теорема

Если к тому же $\lambda_{yy} = 0$ и $\lambda_{ys} \equiv \lambda_y$ для всех $y, s \in Y$, то минимум среднего риска $R(a)$ достигается при

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

Итак, есть две подзадачи, причём вторую мы уже решили!

- 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.

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 $\hat{p}(x)$ по i.i.d. выборке $X_y = \{x_1, \dots, x_m\}$

- 1 Параметрическая МП-оценка $\hat{p}(x) = \varphi(x, \theta)$:

$$\sum_{i=1}^m \log \varphi(x_i, \theta) \rightarrow \max_{\theta}.$$

- 2 Оценка смеси распределений $\hat{p}(x) = \sum_{j=1}^k w_j \varphi(x, \theta_j)$

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

- 3 Непараметрическая оценка плотности:

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

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.

Обоснование оценки Парзена-Розенблатта

Теорема (одномерный случай, $X = \mathbb{R}$)

Пусть выполнены следующие условия:

- 1) X^m — простая выборка из распределения $p(x)$;
- 2) ядро $K(z)$ непрерывно и ограничено: $\int_X K^2(z) dz < \infty$;
- 3) последовательность h_m : $\lim_{m \rightarrow \infty} h_m = 0$ и $\lim_{m \rightarrow \infty} mh_m = \infty$.

Тогда:

- 1) $\hat{p}_{h_m}(x) \rightarrow p(x)$ при $m \rightarrow \infty$ для почти всех $x \in X$;
- 2) скорость сходимости имеет порядок $O(m^{-2/5})$.

А как быть в многомерном случае, когда $X = \mathbb{R}^n$?

Два варианта обобщения на многомерный случай

1. Если объекты описываются n числовыми признаками $f_j: X \rightarrow \mathbb{R}$, $j = 1, \dots, n$.

$$\hat{p}_h(x) = \frac{1}{m} \sum_{i=1}^m \prod_{j=1}^n \frac{1}{h_j} K\left(\frac{f_j(x) - f_j(x_i)}{h_j}\right).$$

2. Если на X задана функция расстояния $\rho(x, x')$:

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

где $V(h) = \int_X K\left(\frac{\rho(x, x_i)}{h}\right) dx$ — нормирующий множитель.

Упражнение: Приведите примеры таких K и ρ , чтобы варианты 1 и 2 оказались эквивалентными.

Метод парзеновского окна (Parzen window)

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

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

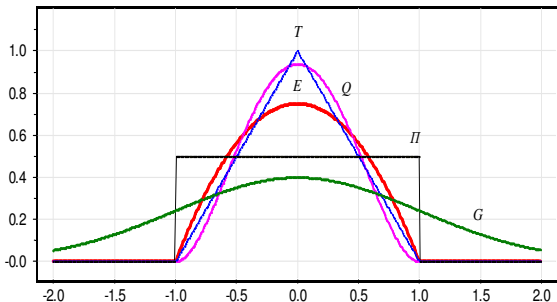
Метод окна Парзена — это метрический классификатор:

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

Замечание 1: множитель $V(h)$ не должен зависеть от x_i (требование однородности пространства $\langle X, \rho \rangle$).

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:

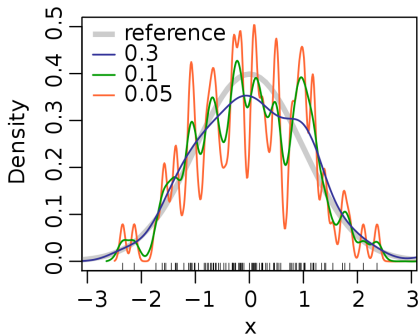
$$J(K) = \int_{-\infty}^{+\infty} E(\hat{p}_h(x) - p(x))^2 dx.$$

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	∞ 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 :

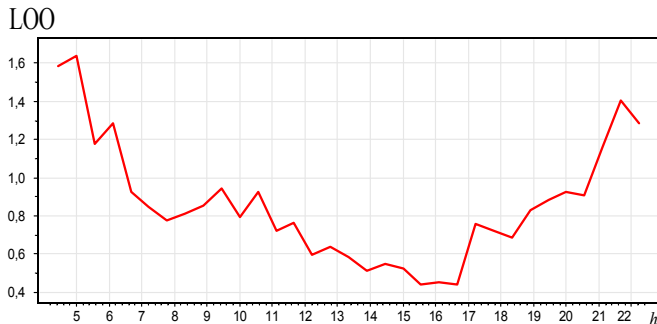


Выбор ширины окна

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^ℓ :

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

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 θ — параметр, φ — фиксированная функция.

Найдём оптимальное θ 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 θ .

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.

Квадратичный дискриминант

Теорема

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

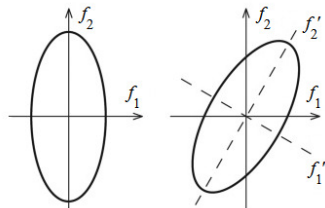
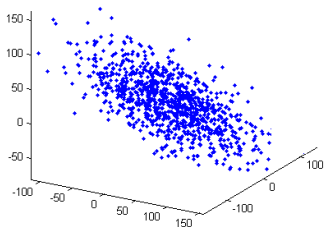
$$\hat{\mu}_y = \frac{1}{\ell_y} \sum_{i: y_i=y} x_i;$$
$$\hat{\Sigma}_y = \frac{1}{\ell_y} \sum_{i: y_i=y} (x_i - \hat{\mu}_y)(x_i - \hat{\mu}_y)^\top.$$

Квадратичный дискриминант — подстановочный алгоритм:

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

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

Линейный дискриминант Фишера

Допущение:

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

Оценка максимума правдоподобия для Σ :

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

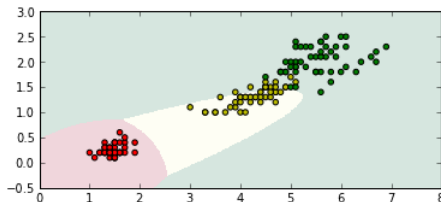
Линейный дискриминант — подстановочный алгоритм:

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

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

Геометрический смысл квадратичного дискриминанта

Пример. Ирисы Фишера: три класса по 50 объектов, разделяющие поверхности — квадратичные

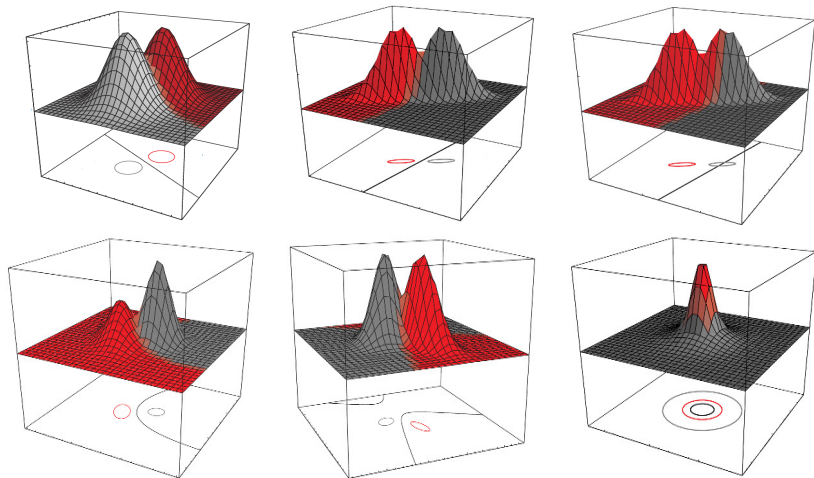


Если $\Sigma_s = \Sigma_t$, то разделяющая поверхность линейная:

$$(x - \mu_{st})^\top \Sigma^{-1} (\mu_s - \mu_t) = \ln(\lambda_t P(t) / \lambda_s P(s)),$$

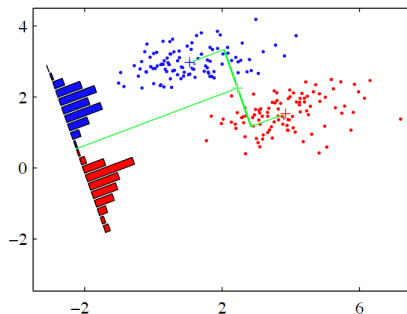
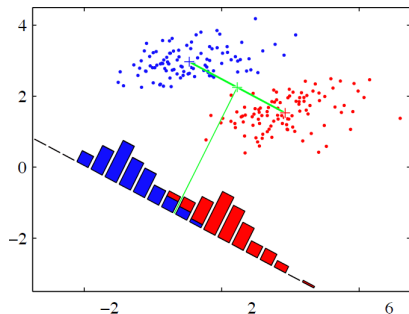
где $\mu_{st} = \frac{1}{2}(\mu_s + \mu_t)$ — середина между центрами классов.

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 τ 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 .

- 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 распределения + принцип максимума правдоподобия.
 - Непараметрический подход наиболее прост и приводит к методу парзеновского окна.
 - Разделение смеси распределений — в следующей лекции