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

ШАД Яндекс • 13 апреля 2015

Содержание

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:

$X^\ell = (x_i, y_i)_{i=1}^\ell$ — простая выборка;

Найти:

классификатор $a\colon X \to Y$ с минимальной вероятностью ошибки.

Временное допущение: пусть известна совместная плотность

$$p(x,y) = p(x)\,\mathsf{P}(y|x) = \mathsf{P}(y)p(x|y).$$

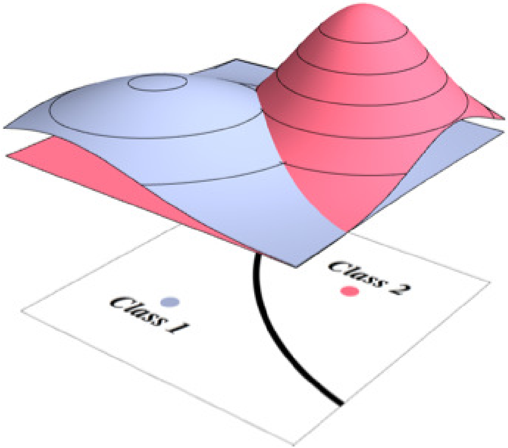
$\mathsf{P}(y)$ — *априорная вероятность* класса y;
$p(x|y)$ — *функция правдоподобия* класса y;
$\mathsf{P}(y|x)$ — *апостериорная вероятность* класса y;

Принцип максимума апостериорной вероятности:

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

Классификация по максимуму функции правдоподобия

Частный случай: $a(x) = \arg\max\limits_{y \in Y} p(x|y)$ при $P(y) = \text{const}$.



Вероятность ошибки и функционал среднего риска

$a\colon X \to Y$ 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: $\mathsf{P}(A_s, y) = \int_{A_s} p(x,y)dx$.

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

Средний риск — мат.ожидание потери для классификатора a:

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

Оптимальный байесовский классификатор

Теорема

Если известны $\mathsf{P}(y)$ и $p(x|y)$, то минимальный средний риск $R(a)$ имеет *байесовский классификатор*

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

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

Ехидное замечание: Когда вместо $P(y)$ и $p(x|y)$ подставляются их эмпирические оценки, байесовский классификатор перестаёт быть оптимальным.

Задачи эмпирического оценивания

- **Оценивание априорных вероятностей** частотами

$$\hat{P}(y) = \frac{\ell_y}{\ell}, \quad \ell_y = |X_y|, \quad X_y = \{x_i \in X \colon y_i = y\}, \quad y \in Y.$$

- **Оценивание функций правдоподобия:**
  Дано:
  $X^m = \{x_1, \dots, x_m\}$ — простая выборка (X_y без ответов y_i).
  Найти:
  эмпирическую оценку плотности $\hat{p}(x)$,
  аппроксимирующую истинную плотность $p(x)$ на всём X:

$$\hat{p}(x) \to p(x) \text{ при } m \to \infty.$$

Анонс: три подхода к оцениванию плотностей

1. **Параметрическое оценивание плотности:**

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

2. **Восстановление смеси распределений:**

$$\hat{p}(x) = \sum_{j=1}^{k} w_j \varphi(x, \theta_j), \quad k \ll m.$$

3. **Непараметрическое оценивание плотности:**

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

Наивный байесовский классификатор

Допущение (наивное):

Признаки $f_j\colon X \to D_j$ — 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:

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

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 $[a,b]$:

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

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

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

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:

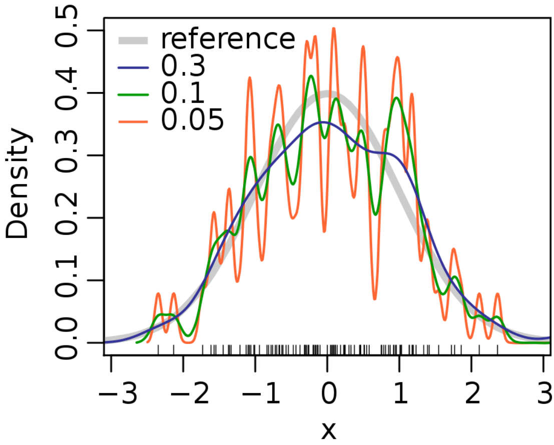
$$\hat{p}_h(x) = \frac{1}{mh} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h} \right),$$

где $K(r)$ — *ядро*, удовлетворяющее требованиям:
— чётная функция;
— нормированная функция: $\int K(r)\,dr = 1$;
— (как правило) невозрастающая, неотрицательная функция.

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

Пример. Ядерные оценки плотности при разных h

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



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: $\lim\limits_{m\to\infty} h_m = 0$ и $\lim\limits_{m\to\infty} mh_m = \infty$.

Тогда:

1) $\hat{p}_{h_m}(x) \to p(x)$ при $m \to \infty$ 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 $f_j\colon X \to \mathbb{R}$, $j = 1, \dots, n$.
$$\hat{p}_h(x) = \frac{1}{m} \sum_{i=1}^{m} \prod_{j=1}^{n} \frac{1}{h_j} K\left(\frac{f_j(x) - f_j(x_i)}{h_j} \right).$$

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 $y \in Y$:

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

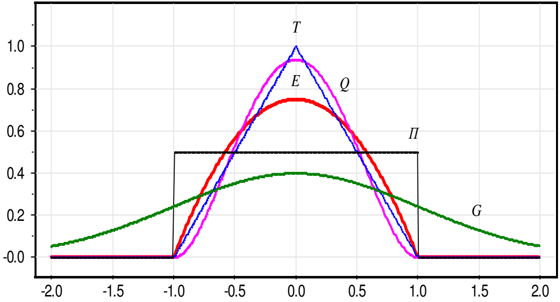
Метод окна Парзена — это метрический классификатор:

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

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



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

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

$T(r) = \bigl(1 - |r|\bigr)\bigl[|r| \leqslant 1\bigr]$ — треугольное;

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

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

Выбор ядра почти не влияет на качество восстановления

Функционал качества восстановления плотности:

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

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 $m \to \infty$, причём это отношение не зависит от $p(x)$.

Окна переменной ширины

Проблема:

при наличии локальных сгущений любая h не оптимальна.

Идея:

задавать не ширину окна h, а число соседей k.

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

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

$$\rho(x, x^{(1)}) \leqslant \cdots \leqslant \rho(x, x^{(\ell)})$$

Замечание 1:

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

Замечание 2:

Оптимизация k по LOO аналогична оптимизации h.

Принцип максимума правдоподобия

Пусть известна параметрическая модель плотности

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

где θ — параметр, φ — фиксированная функция.
Задача — найти оптимальное θ по простой выборке X^m.
Принцип максимума правдоподобия:

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

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:

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

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
   $\bigl\{x \in X \bigm| \lambda_y P(y) p(x|y) = \lambda_s P(s) p(x|s)\bigr\}$
   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 максимума правдоподобия, $y \in Y$:

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

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

Квадратичный дискриминант — подстановочный алгоритм:

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

Линейный дискриминант Фишера

Допущение:

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

Оценка максимума правдоподобия для Σ:

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

Линейный дискриминант — подстановочный алгоритм:

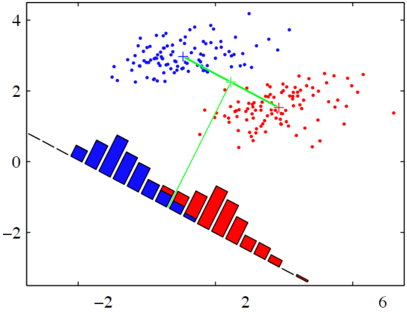
$$a(x) = \arg\max_{y \in Y} \lambda_y \hat{P}(y) \hat{p}(x|y) =$$

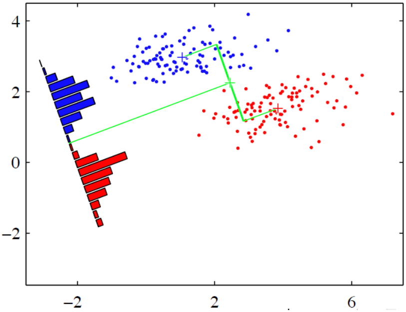
$$= \arg\max_{y \in Y} \Bigl(\underbrace{\ln(\lambda_y \hat{P}(y)) - \tfrac{1}{2} \hat{\mu}_y^{\mathsf{T}} \hat{\Sigma}^{-1} \hat{\mu}_y}_{\beta_y} + x^{\mathsf{T}} \underbrace{\hat{\Sigma}^{-1} \hat{\mu}_y}_{\alpha_y} \Bigr);$$

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

Геометрический смысл линейного дискриминанта

В одномерной проекции на направляющий вектор разделяющей гиперплоскости классы разделяются наиболее чётко, то есть вероятность ошибки минимальна.





Проблема мулитиколлинеарности

Недостатки квадратичного дискриминанта:

- Необходимость обращать матрицы $\hat{\Sigma}_y$
- Матрица $\hat{\Sigma}_y$ может быть плохо обусловлена
- При $\ell_y < n$ матрица $\hat{\Sigma}_y$ вырождена

Линейный дискриминант:

- требует обращения только одной матрицы, более устойчив
- хуже описывает классы различной формы

Далее — **меры по улучшению алгоритма**:

- регуляризация ковариационной матрицы
- диагонализация ковариационной матрицы
- преобразование или отбор признаков

Методы устранения мультиколлинеарности

- 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:

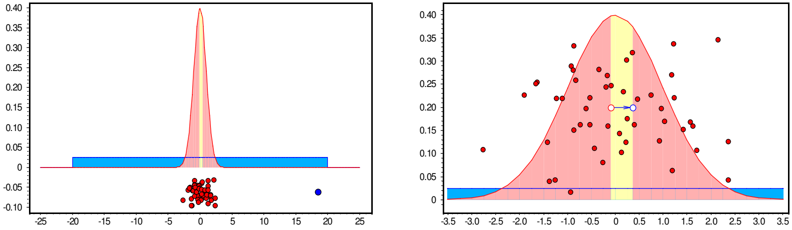
$$a(x) = \arg\max_{y \in Y} \biggl(\ln \lambda_y \hat{P}(y) + \sum_{j=1}^{n} \ln \hat{p}_j(\xi_j|y) \biggr), \quad x \equiv (\xi_1, \ldots, \xi_n);$$

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

где $\hat{\mu}_{yj}$ и $\hat{\sigma}_{yj}$ — оценки среднего и дисперсии j-го признака, вычисленные по подвыборке $X_y \subset X^\ell$ класса y.

Проблема выбросов (outliers)

Эмпирическое среднее является оценкой матожидания, неустойчивой к редким большим выбросам.

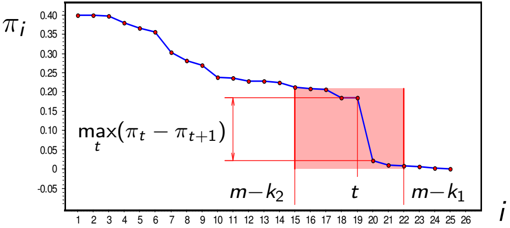


Пример. Одномерная нормальная плотность $\mathcal{N}(0, 1)$, загрязнённая равномерным на $[-20, +20]$ распределением, $\ell = 50$, смещение эмпирического среднего 0.359.

Цензурирование выборки (отсев выбросов)

Идея: задача решается дважды; после первого раза объекты с наибольшими ошибками исключаются из обучения.

Алгоритм (для задачи восстановления плотности)
1) оценить параметр $\hat\theta$ по всей выборке X^m;
2) вычислить правдоподобия $\pi_i = \varphi(x_i;\hat\theta)$ для всех $x_i \in X^m$;
3) отсортировать выборку по убыванию: $\pi_1 \geqslant \ldots \geqslant \pi_m$;
4) удалить из X^m от k_1 до k_2 объектов, попавших в конец ряда;
5) оценить параметр $\hat\theta$ по укороченной выборке X^m;



Резюме в конце лекции

- Эту формулу полезно помнить:
  $a(x) = \arg\max\limits_{y\in Y} \lambda_y P(y) p(x|y)$.
- 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