

Восстановление плотности распределения

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)»

МФТИ • 20 ноября 2021

- 1 Параметрические методы восстановления плотности**
 - Задача восстановления плотности распределения
 - Восстановление многомерной гауссовской плотности
 - Проблема мультиколлинеарности
- 2 Непараметрическое восстановление плотности**
 - Восстановление одномерных плотностей
 - Восстановление многомерных плотностей
 - Выбор ядра и ширины окна
- 3 Разделение смеси распределений**
 - Задача разделения смеси распределений
 - EM-алгоритм
 - Обобщения и модификации EM-алгоритма

Восстановление плотности — задача обучения без учителя

Дано: простая (i.i.d.) выборка $X^\ell = \{x_1, \dots, x_\ell\} \sim p(x)$.

Найти параметрическую модель плотности распределения:

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

где θ — параметр, φ — фиксированная функция.

Критерий — максимум (логарифма) правдоподобия выборки:

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

Необходимое условие оптимума:

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

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

Восстановление многомерной гауссовской плотности

Пусть объекты x описываются n признаками $f_j(x) \in \mathbb{R}$
 и выборка порождена n -мерной гауссовской плотностью:

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

$\mu \in \mathbb{R}^n$ — вектор математического ожидания, $\mu = E x$

$\Sigma \in \mathbb{R}^{n \times n}$ — ковариационная матрица, $\Sigma = E(x - \mu)(x - \mu)^\top$
 (симметричная, невырожденная, положительно определённая)

Выборочные оценки максимального правдоподобия:

$$\frac{\partial}{\partial \mu} \ln L(\mu, \Sigma; X^\ell) = 0 \quad \Rightarrow \quad \hat{\mu} = \frac{1}{\ell} \sum_{i=1}^{\ell} x_i$$

$$\frac{\partial}{\partial \Sigma} \ln L(\mu, \Sigma; X^\ell) = 0 \quad \Rightarrow \quad \hat{\Sigma} = \frac{1}{\ell} \sum_{i=1}^{\ell} (x_i - \hat{\mu})(x_i - \hat{\mu})^\top$$

Некоторые приёмы матричного дифференцирования

Производная скалярной функции $f(A)$ по матрице $A = (a_{ij})$:

$$\frac{\partial}{\partial A} f(A) = \left(\frac{\partial}{\partial a_{ij}} f(A) \right)$$

$\text{diag } A$ — диагональ матрицы A , остальные элементы нули

A — квадратная $n \times n$ -матрица

u — вектор размерности n

если A произвольного вида:

$$\frac{\partial}{\partial u} u^T A u = A^T u + A u$$

$$\frac{\partial}{\partial A} \ln |A| = A^{-1T}$$

$$\frac{\partial}{\partial A} u^T A u = u u^T$$

если A симметричная:

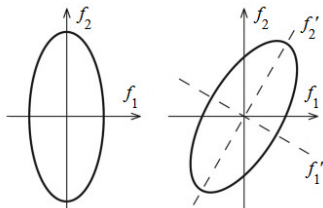
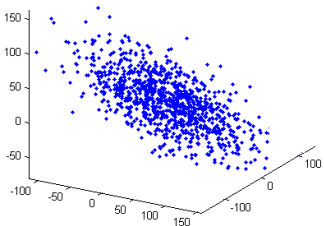
$$\frac{\partial}{\partial u} u^T A u = 2A u$$

$$\frac{\partial}{\partial A} \ln |A| = 2A^{-1} - \text{diag } A^{-1}$$

$$\frac{\partial}{\partial A} u^T A u = 2u u^T - \text{diag } u u^T$$

Геометрический смысл многомерной нормальной плотности

Эллипсоид рассеяния — облако точек эллиптической формы:



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

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

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

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

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

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

Проблема мультиколлинеарности

Проблема: при $\ell < n$ матрица $\hat{\Sigma}$ вырождена, но даже при $\ell \geq n$ она может оказаться плохо обусловленной.

Регуляризация ковариационной матрицы $\hat{\Sigma} + \tau I_n$ увеличивает собственные значения на τ , сохраняя собственные векторы (параметр τ можно подбирать по скользящему контролю)

Диагонализация ковариационной матрицы — оценивание n одномерных плотностей признаков $f_j(x)$, $j = 1, \dots, n$:

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

где $\hat{\mu}_j$ и $\hat{\sigma}_j^2$ — оценки среднего и дисперсии признака j :

$$\hat{\mu}_j = \frac{1}{\ell} \sum_{i=1}^{\ell} f_j(x_i)$$

$$\hat{\sigma}_j^2 = \frac{1}{\ell} \sum_{i=1}^{\ell} (f_j(x_i) - \hat{\mu}_j)^2$$

Задача непараметрического восстановления плотности

Задача: по выборке $X^\ell = (x_i)_{i=1}^\ell$ оценить плотность $\hat{p}(x)$,
без введения параметрической модели плотности

Дискретный случай: $x_i \in D$, $|D| \ll \ell$. Гистограмма частот:

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

Одномерный непрерывный случай: $x_i \in \mathbb{R}$. 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):

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

Локальная непараметрическая оценка Парзена-Розенблатта

Эмпирическая оценка плотности по окну ширины h :

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

Обобщение: оценка Парзена-Розенблатта по окну ширины h :

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

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, $x_i \in \mathbb{R}$)

Пусть выполнены следующие условия:

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

Тогда:

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

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

Два варианта обобщения на многомерный случай

- 1 Если объекты описываются n признаками $f_j: X \rightarrow \mathbb{R}$:

$$\hat{p}_{h_1 \dots h_n}(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \prod_{j=1}^n \frac{1}{h_j} K\left(\frac{f_j(x) - f_j(x_i)}{h_j}\right)$$

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

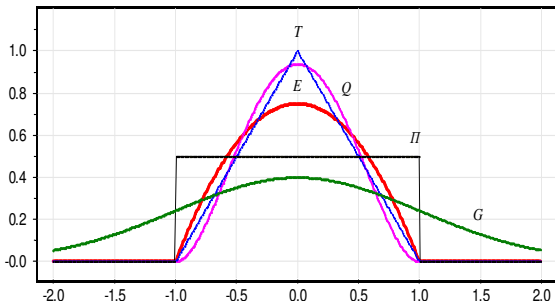
$$\hat{p}_h(x) = \frac{1}{\ell V(h)} \sum_{i=1}^{\ell} K\left(\frac{\rho(x, x_i)}{h}\right)$$

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

Сферическое гауссовское ядро — частный случай обоих:

$$\hat{p}_h(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \prod_{j=1}^n \frac{1}{\sqrt{2\pi}h} \exp\left(-\frac{(f_j(x) - f_j(x_i))^2}{2h^2}\right)$$

Выбор ядра



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

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

$T(r) = (1 - |r|)[|r| \leq 1]$ — 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:

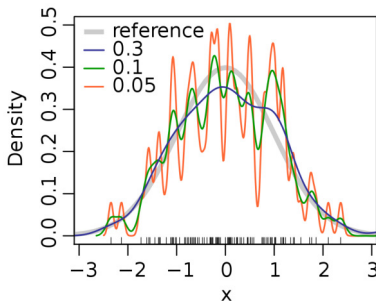
$$J(K) = \int_{-\infty}^{+\infty} E(\hat{p}_h(x) - p(x))^2 dx.$$

Асимптотические значения отношения $J(K^*)/J(K)$ при $l \rightarrow \infty$ не зависят от вида распределения $p(x)$.

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

Зависимость оценки плотности от ширины окна

Оценка $\hat{\rho}_h(x)$ при различных значениях ширины окна h :



- Качество восстановления плотности существенно зависит от ширины окна h , но слабо зависит от вида ядра K
- При неоднородности локальных сгущений плотности можно задавать $h_k(x) = \rho(x, x^{(k+1)})$, где k — число соседей

Выбор ширины окна

Скользящий контроль *Leave One Out* для оценки плотности:

$$\text{LOO}(h) = - \sum_{i=1}^{\ell} \ln \hat{p}_h(x_i; X^{\ell} \setminus x_i) \rightarrow \min_h,$$

Типичный вид зависимости $\text{LOO}(h)$ или $\text{LOO}(k)$:



Ретроспектива: (непара)метрические методы анализа данных

Восстановление плотности. Метод Парзена–Розенблатта:

$$\hat{\rho}_h(x; X^\ell) = \frac{1}{\ell V(h)} \sum_{i=1}^{\ell} K\left(\frac{\rho(x, x_i)}{h}\right)$$

Классификация. Метод парзеновского окна:

$$a_h(x; X^\ell, Y^\ell) = \arg \max_{y \in Y} \sum_{i=1}^{\ell} [y_i = y] K\left(\frac{\rho(x, x_i)}{h}\right)$$

Регрессия. Метод ядерного сглаживания Надарая–Ватсона:

$$a_h(x; X^\ell, Y^\ell) = \frac{\sum_{i=1}^{\ell} y_i K\left(\frac{\rho(x, x_i)}{h}\right)}{\sum_{i=1}^{\ell} K\left(\frac{\rho(x, x_i)}{h}\right)}$$

Задача разделения смеси распределений

Порождающая модель смеси k распределений:

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

$\varphi(x, \theta_j) = p(x|j)$ — функция правдоподобия j -й компоненты;
 $w_j = P(j)$ — априорная вероятность j -й компоненты.

Задача максимизации логарифма правдоподобия:

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

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

EM-алгоритм для разделения смеси распределений

Теорема (необходимые условия экстремума)

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

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

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

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

EM-алгоритм — метод простых итераций для решения системы

Вероятностная интерпретация шагов EM-алгоритма

E-шаг — это формула Байеса:

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

Нормировка условных вероятностей: $\sum_{j=1}^k g_{ij} = 1$

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

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

вес компоненты определяется как средний вес её объектов:

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

Доказательство. Условия Каруша–Куна–Таккера

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

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

Приравниваем нулю производные:

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

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

EM-алгоритм

вход: $X^\ell = \{x_1, \dots, x_\ell\}$, k ;

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

инициализировать $(\theta_j)_{j=1}^k$, $w_j := \frac{1}{k}$;

повторять

Е-шаг (expectation): для всех $i = 1, \dots, \ell$, $j = 1, \dots, k$

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

М-шаг (maximization): для всех $j = 1, \dots, k$

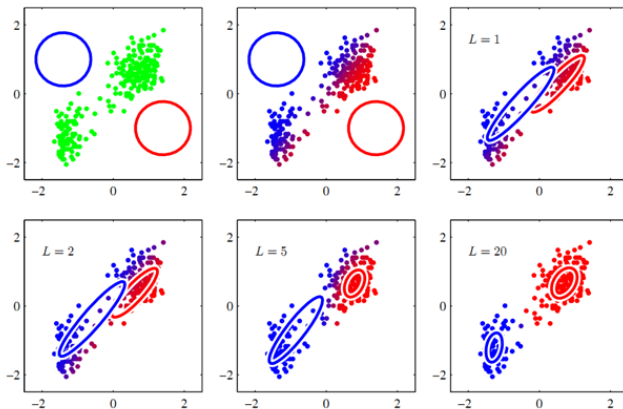
$$\theta_j := \arg \max_{\theta} \sum_{i=1}^{\ell} g_{ij} \ln \varphi(x_i, \theta); \quad w_j := \frac{1}{\ell} \sum_{i=1}^{\ell} g_{ij};$$

пока w_j, θ_j и/или g_{ij} не сошлись;

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 :



GEM — обобщённый EM-алгоритм

Идея:

Не обязательно добиваться высокой точности на M-шаге.
Достаточно лишь сместиться в направлении максимума,
сделав одну или несколько итераций, и вернуться на E-шаг.

Преимущества:

- сохраняется свойство слабой локальной сходимости (в смысле увеличения правдоподобия на каждом шаге)
- повышается скорость сходимости при сопоставимом качестве решения

SEM — стохастический EM-алгоритм

Идея: на M-шаге вместо максимизации

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

максимизируется обычное, невзвешенное, правдоподобие

$$\theta_j := \arg \max_{\theta} \sum_{x_i \in X_j} \ln \varphi(x_i, \theta),$$

выборки X_j строятся путём сэмплирования объектов из X^{ℓ}
 ℓ раз с возвращениями: $i \sim P(i|j) = \frac{P(j|x_i)P(i)}{P(j)} = \frac{g_{ij}}{\ell w_j}$.

Преимущества:

ускорение сходимости, предотвращение зацикливаний.

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

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

HEM — иерархический EM-алгоритм

Идея:

«Плохо описанные» компоненты с низким правдоподобием расщепляются на две или более *дочерних* компонент.

Преимущество:

автоматически выявляется иерархическая кластерная структура выборки, полезная для понимания исходных данных (data understanding).

Резюме: три подхода к оцениванию плотностей

- 1 **Параметрическое оценивание плотности**
модель плотности + максимизация правдоподобия:

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

- 2 **Непараметрическое оценивание плотности**
наиболее прост, приводит к методу парзеновского окна:

$$\hat{\rho}(x) = \sum_{i=1}^{\ell} \frac{1}{\ell V(h)} K\left(\frac{\rho(x, x_i)}{h}\right)$$

- 3 **Разделение смеси распределений**
в случае смеси гауссиан приводит к сети RBF:

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