Методы машинного обучения. Восстановление плотности распределения

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

ВМК МГУ • 31 октября 2023

Содержание

Восстановление плотности — задача обучения без учителя

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

Найти параметрическую модель плотности распределения:

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

где θ — параметр, φ — фиксированная функция.

Критерий — максимум (логарифма) правдоподобия выборки:

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

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:

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

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

Выборочные оценки максимального правдоподобия:

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

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

Некоторые приёмы матричного дифференцирования

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

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

$\operatorname{diag} A$ — диагональ матрицы A, остальные элементы нули
A — квадратная $n \times n$-матрица
u — вектор размерности n

если A произвольного вида:

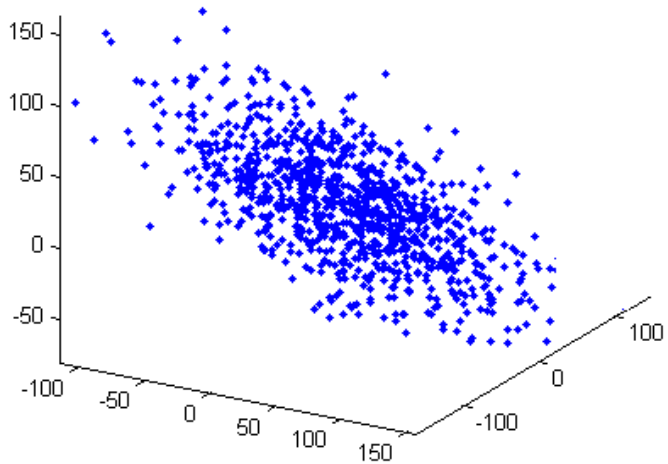
$$\frac{\partial}{\partial u} u^{\mathsf{T}} A u = A^{\mathsf{T}} u + A u$$
$$\frac{\partial}{\partial A} \ln |A| = A^{-1\mathsf{T}}$$
$$\frac{\partial}{\partial A} u^{\mathsf{T}} A u = u u^{\mathsf{T}}$$

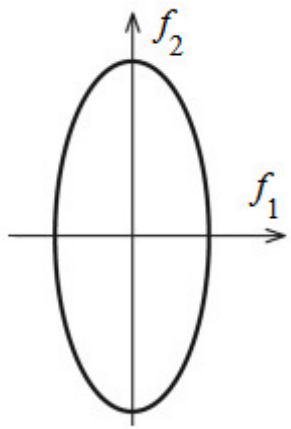
если A симметричная, $A^{\mathsf{T}} = A$:

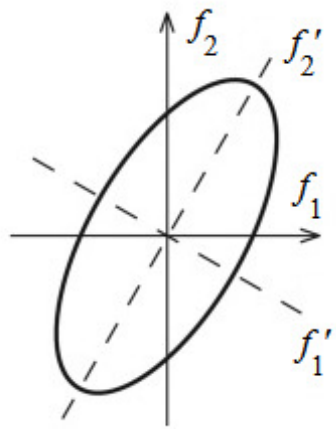
$$\frac{\partial}{\partial u} u^{\mathsf{T}} A u = 2 A u$$
$$\frac{\partial}{\partial A} \ln |A| = 2 A^{-1} - \operatorname{diag} A^{-1}$$
$$\frac{\partial}{\partial A} u^{\mathsf{T}} A u = 2 u u^{\mathsf{T}} - \operatorname{diag} u u^{\mathsf{T}}$$

Геометрический смысл многомерной нормальной плотности

Эллипсоид рассеяния — облако точек эллиптической формы:







При $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ оси эллипсоида параллельны ортам.
В общем случае: $\Sigma = VSV^\mathsf{T}$ — спектральное разложение,
$V = (v_1, \ldots, v_n)$ — ортогональные собств. векторы, $V^\mathsf{T}V = I_n$
$S = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ — собственные значения матрицы Σ

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

$x' = V^\mathsf{T}x$ — ортогональное преобразование поворот/отражение

Проблема мулитиколлинеарности

Проблема: при $\ell < n$ матрица $\hat{\Sigma}$ вырождена, но даже при $\ell \geqslant n$ 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 $f_j(x)$, $j = 1, \ldots, n$:

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

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 $[a, b]$:

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

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

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

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

Обоснование оценки Парзена-Розенблатта

Другое название — Kernel Density Estimate (KDE)

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: $\lim\limits_{\ell\to\infty} h_\ell = 0$ и $\lim\limits_{\ell\to\infty} \ell h_\ell = \infty$.

Тогда:

1) $\hat{p}_{h_\ell}(x) \to p(x)$ при $\ell \to \infty$ 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 $f_j : X \to \mathbb{R}$:

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

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

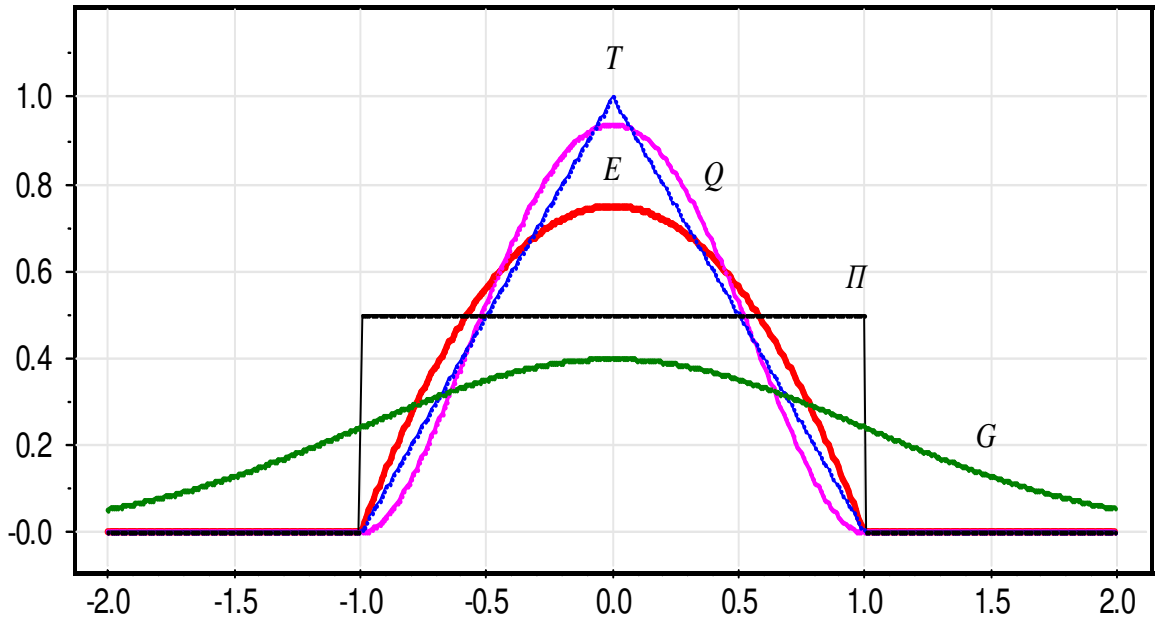
$$\hat{p}_h(x) = \frac{1}{\ell V(h)} \sum_{i=1}^{\ell} K\left(\frac{\rho(x, x_i)}{h} \right)$$

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



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

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

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

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

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

Выбор ядра почти не влияет на качество восстановления

Функционал качества восстановления плотности:

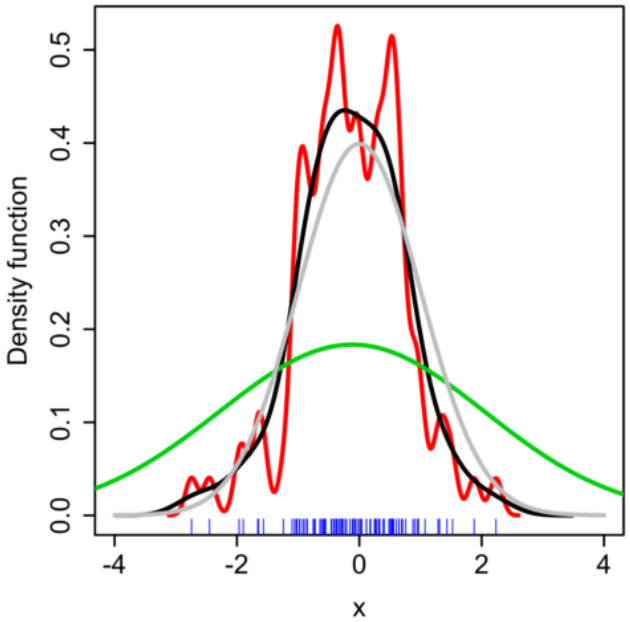
$$J(K) = \int_{-\infty}^{+\infty} \mathsf{E}\big(\hat{p}_h(x) - p(x)\big)^2\, dx.$$

Асимптотические значения отношения $J(K^*)/J(K)$ при $\ell \to \infty$ 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 $\hat{p}_h(x)$ при различных значениях ширины окна h:



истинная плотность
(стандартная гауссовская)

$h = 0.05$ — переобучение
$h = 0.337$ — оптимальная
$h = 2.0$ — недообучение

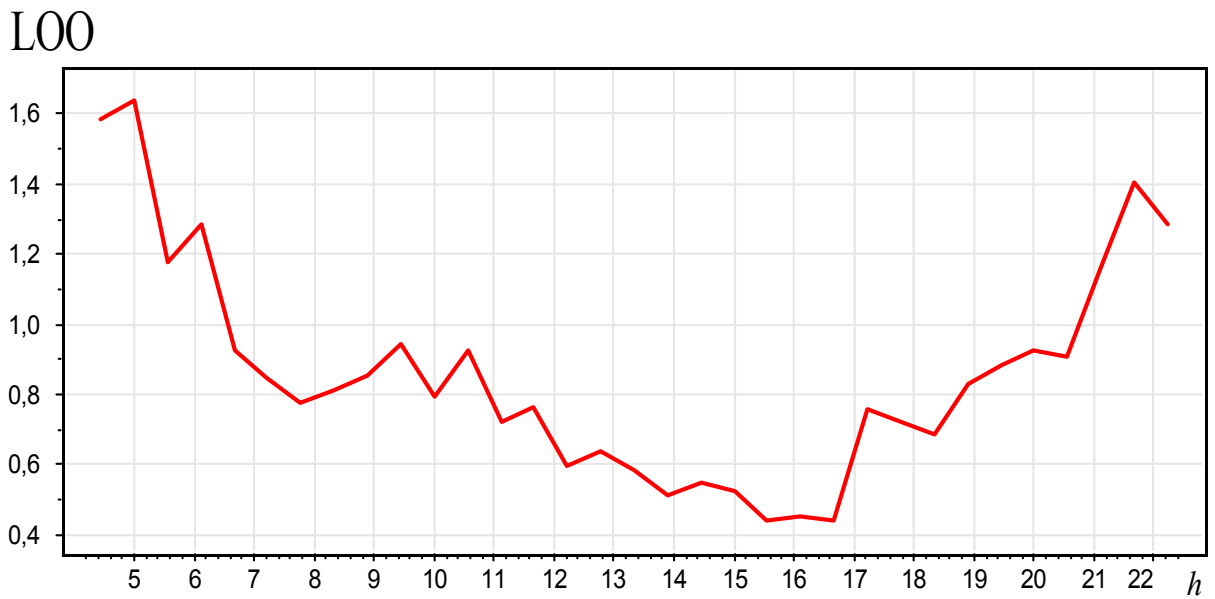
- Качество восстановления плотности существенно зависит от ширины окна h, но слабо зависит от вида ядра K
- При неоднородности локальных сгущений плотности можно задавать $h_k(x) = \rho\left(x, x^{(k+1)}\right)$, где k — число соседей

Выбор ширины окна

Скользящий контроль *Leave One Out* для оценки плотности:

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

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



Ретроспектива: (непара)метрические методы анализа данных

Восстановление плотности. Метод Парзена–Розенблатта:

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

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:

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

описывает двухуровневый процесс порождения данных:

1. $j \sim P(j) \equiv w_j$ — дискретное *априорное распределение*
2. $x \sim p(x|j) \equiv \varphi(x, \theta_j)$ — плотность j-й компоненты

Максимизация (логарифма) правдоподобия
приводит к задаче математического программирования:

$$\begin{cases} L(w, \theta) = \ln \prod_{i=1}^{\ell} p(x_i) = \sum_{i=1}^{\ell} \ln \sum_{j=1}^{k} w_j \varphi(x_i, \theta_j) \to \max\limits_{w,\theta} \\ \sum_{j=1}^{k} w_j = 1, \quad w_j \geqslant 0 \end{cases}$$

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}:

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

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

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

EM-алгоритм — метод простых итераций для решения системы

Вероятностная интерпретация шагов EM-алгоритма

E-шаг — это формула Байеса:

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

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 $L(w,\theta) \to \max$:

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

Приравниваем нулю производные:

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

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

■

EM-алгоритм для разделения смеси распределений

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

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

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

повторять

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

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

M-шаг (maximization): для всех $j = 1, \ldots, k$

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

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

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

Разделение смеси гауссиан (Gaussian Mixture Model, GMM)

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

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

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

повторять

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

$$g_{ij} := \frac{w_j \mathcal{N}(x_i; \mu_j, \Sigma_j)}{\sum_{s=1}^k w_s \mathcal{N}(x_i; \mu_s, \Sigma_s)};$$

M-шаг (maximization): для всех $j = 1, \ldots, k$

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

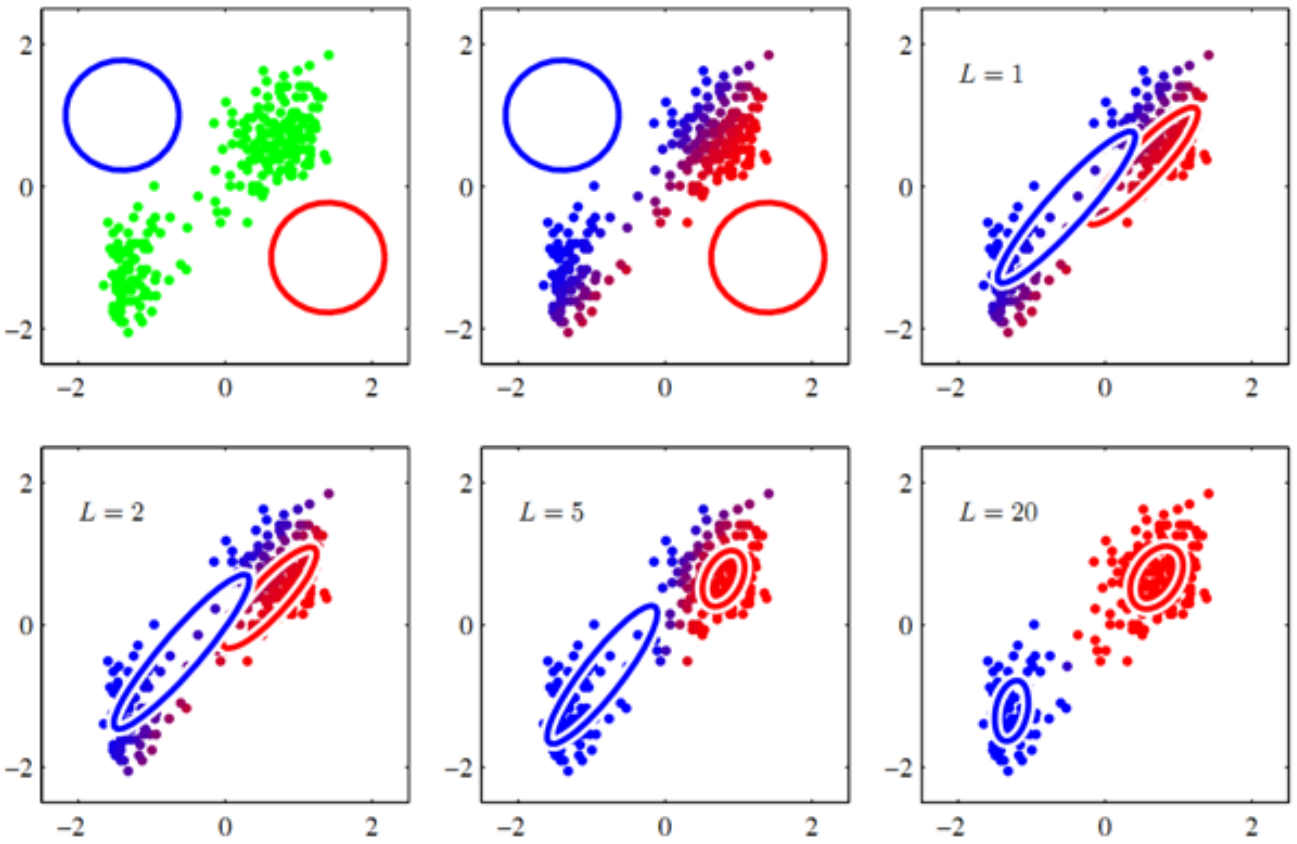
$$\mu_j := \frac{1}{\ell w_j} \sum_{i=1}^\ell g_{ij} x_i;$$

$$\Sigma_j := \frac{1}{\ell w_j} \sum_{i=1}^\ell g_{ij} (x_i - \mu_j)(x_i - \mu_j)^\mathsf{T};$$

пока (w_j, μ_j, Σ_j) и/или g_{ij} не сошлись;

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 нужно добиваться точного решения задачи M-шага

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

достаточно сместиться в направлении максимума,
сделав одну или несколько итераций, затем выполнить 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: $i \sim P(i \mid j) = \frac{P(j \mid x_i)P(i)}{P(j)} = \frac{g_{ij}}{\ell w_j}$.

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

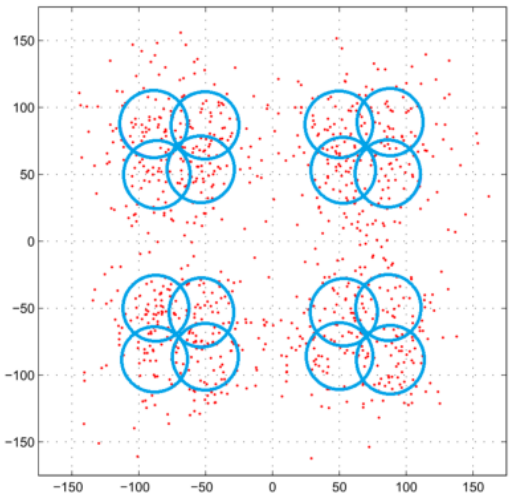
$$w_j \propto \left(\frac{1}{\ell} \sum_{i=1}^{\ell} g_{ij} - \tau \right)_+$$

HEM — иерархический EM-алгоритм

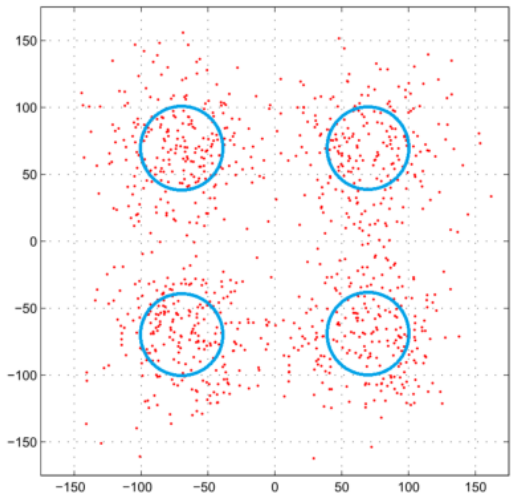
Связь $w_s^{h+1} = p(s|j) w_j^h$ между соседними уровнями h и $h+1$:

$$p^h(x) = \sum_{j=1}^{k_h} w_j^h \varphi(x, \theta_j^h) \qquad p^{h+1}(x) = \sum_{s=1}^{k_{h+1}} w_s^{h+1} \varphi(x, \theta_s^{h+1})$$

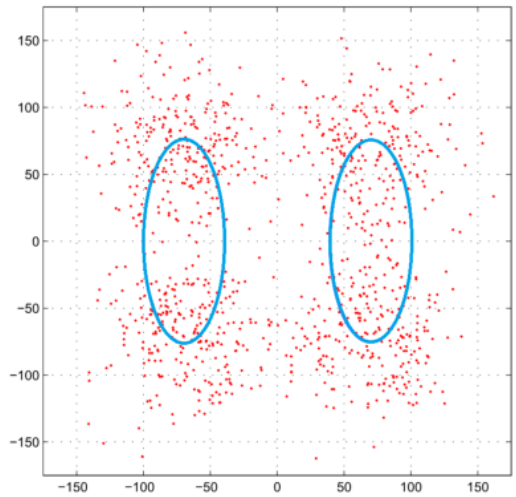
$h = 3$



$h = 2$



$h = 1$



N.Vasconcelos, A.Lippman. Learning Mixture Hierarchies. NIPS 1998.

Резюме: три подхода к оцениванию плотностей

1. **Параметрическое оценивание плотности**
   модель плотности + максимизация правдоподобия:
   $$\hat{p}(x) = \varphi(x, \theta)$$
2. **Непараметрическое оценивание плотности**
   наиболее прост, парзеновская оценка плотности:
   $$\hat{p}(x) = \sum_{i=1}^{\ell} \frac{1}{\ell V(h)} K\left(\frac{\rho(x, x_i)}{h}\right)$$
3. **Разделение смеси распределений**
   максимизация правдоподобия итерациями EM-алгоритма:
   $$\hat{p}(x) = \sum_{j=1}^{k} w_j \varphi(x, \theta_j), \quad k \ll \ell$$