Методы машинного обучения. Обучение без учителя: оценивание плотности распределения

Воронцов Константин Вячеславович
www.MachineLearning.ru/wiki?title=User:Vokov
вопросы к лектору: k.vorontsov@iai.msu.ru

материалы курса:
github.com/MSU-ML-COURSE/ML-COURSE-24-25
орг.вопросы по курсу: ml.cmc@mail.ru

ВМК МГУ • 4 марта 2025

Содержание

1. **Параметрические методы восстановления плотности**
   - Задача восстановления плотности распределения
   - Восстановление многомерной гауссовской плотности
   - Восстановление дискретного распределения
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

Напоминание. Задача оценивания плотности (density estimation)

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

Найти параметрическую модель плотности распределения:

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

где w — вектор параметров, φ — фиксированная функция

Критерий — максимум (логарифма) правдоподобия выборки, MLE-оценивание параметра w (Maximum Likelihood Estimate):

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

Аналитическое решение: необходимое условие экстремума

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

при условии достаточной гладкости функции $\varphi(x; w)$ по w

Напоминание. Оценка многомерной гауссовской плотности

Пусть объекты описываются n признаками, $x_i \in \mathbb{R}^n$ и выборка порождена n-мерной гауссовской плотностью:

$$x_i \sim p(x) = \mathcal{N}(x;\mu,\Sigma) = \frac{\exp\bigl(-\frac{1}{2}(x-\mu)^{\mathsf T}\Sigma^{-1}(x-\mu)\bigr)}{\sqrt{(2\pi)^n \det\Sigma}}, \quad x \in \mathbb{R}^n$$

$\mu \in \mathbb{R}^n$ — 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)

Выборочная оценка максимального правдоподобия

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

Доказательство: из условий $\frac{\partial}{\partial\mu}\ln L(\mu,\Sigma;X^\ell) = 0, \quad \frac{\partial}{\partial\Sigma}\ln L(\mu,\Sigma;X^\ell) = 0$

Оценка дискретного распределения

Дано: выборка $x_i \in X$, $|X| < \infty$, порождаемая i.i.d. дискретным распределением $(p_x \colon x \in X)$, $\sum_x p_x = 1$, $p_x \geqslant 0$

Найти: параметры распределения $(p_x \colon x \in X)$

Критерий: максимум (логарифма) правдоподобия выборки

$$\ln \prod_{i=1}^{\ell} p_{x_i} = \sum_{x \in X} \underbrace{\textstyle\sum_{i=1}^{\ell} [x_i = x]}_{\ell_x} \ln p_x = \sum_{x \in X} \ell_x \ln p_x \to \max_{(p_x)}$$

Выборочная оценка максимального правдоподобия

$\hat{p}_x = \frac{\ell_x}{\ell}$ — частотные оценки вероятностей $p_x = P(x_i = x)$, оценка минимума кросс-энтропии, эмпирическая гистограмма

Доказательство из условий ККТ: $\frac{\partial}{\partial p_x} \Bigl(\sum\limits_{x \in X} \ell_x \ln p_x + \mu \Bigl(1 - \sum\limits_{x \in X} p_x \Bigr) \Bigr) = 0$

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: $x_i \in X$, $|X| \ll \ell$. Частотная оценка:

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

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

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

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 $f_j\colon X \to \mathbb{R}$:

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

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

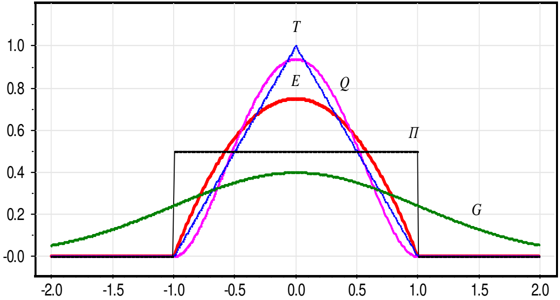
$$\hat{p}_h(x) = \frac{1}{\ell V(h)}\sum_{i=1}^{\ell} K\left(\frac{\rho(x,x_i)}{h}\right)$$

где $V(h) = \int_X K\bigl(\frac{\rho(x,x_i)}{h}\bigr)\,dx$ — 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



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

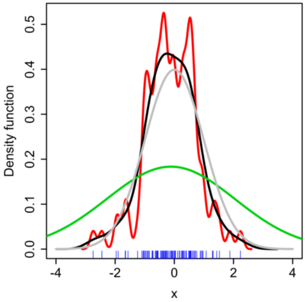
$$J(K) = \int_{-\infty}^{+\infty} \mathsf{E}\bigl(\hat{p}_h(x) - p(x)\bigr)^2\,dx.$$

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

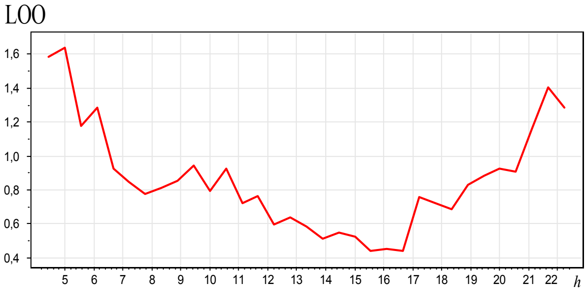
- Качество восстановления плотности существенно зависит от ширины окна h, но слабо зависит от вида ядра K
- При неоднородности локальных сгущений плотности можно задавать $h_k(x) = \rho\bigl(x, x^{(k+1)}\bigr)$, где k — число соседей

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:

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

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}:

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

М-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:

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

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

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

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

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: $X^\ell = \{x_1, \dots, x_\ell\} \subset \mathbb{R}^n$, число компонент смеси 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 $i = 1, \dots, \ell$, $j = 1, \dots, k$

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

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

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

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

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

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

Разделение смеси гауссиан с диагональными матрицами Σ_j

вход: $X^\ell = \{x_1, \dots, x_\ell\} \subset \mathbb{R}^n$, число компонент смеси k;

выход: $\bigl(w_j, \mu_j, \Sigma_j = \mathrm{diag}(\sigma_{j1}^2, \dots, \sigma_{jn}^2)\bigr)_{j=1}^k$;

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

повторять

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

$$g_{ij} := \frac{w_j \prod_{d=1}^n \mathcal{N}\bigl(f_d(x_i); \mu_{jd}, \sigma_{jd}^2\bigr)}{\sum_{s=1}^k w_s \prod_{d=1}^n \mathcal{N}\bigl(f_d(x_i); \mu_{sd}, \sigma_{sd}^2\bigr)};$$

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

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

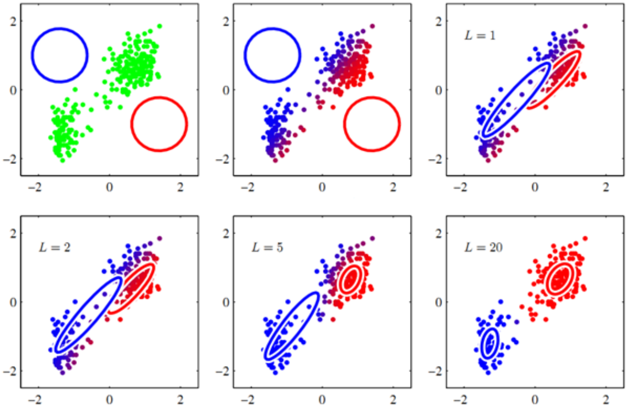
$$\mu_{jd} := \frac{1}{\ell w_j} \sum_{i=1}^\ell g_{ij} f_d(x_i), \quad d = 1, \dots, n;$$

$$\sigma_{jd}^2 := \frac{1}{\ell w_j} \sum_{i=1}^\ell g_{ij} \bigl(f_d(x_i) - \mu_{jd}\bigr)^2, \quad d = 1, \dots, n;$$

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:



Сравнение EM-алгоритма для GMM и метода k-средних

Основные отличия GMM-EM и k-means:

- GMM-EM: мягкая кластеризация: $g_{ij} = \mathsf{P}(j|x_i)$
  k-means: жёсткая кластеризация: $g_{ij} = [j = \arg\max_j \mathsf{P}(j|x_i)]$
- GMM-EM: кластеры эллиптические, настраиваемые
  k-means: кластеры сферические, не настраиваемые

Гибриды (упрощение GMM-EM $\longleftrightarrow$ усложнение k-means):

- GMM-EM с жёсткой кластеризацией на E-шаге
- GMM-EM со сферическими гауссианами

Недостатки k-means (немного компенсируемые GMM-EM):

- чувствительность к выбору начального приближения
- медленная сходимость (пользуйтесь k-means++)

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: $i \sim P(i\,|\,j) = \frac{P(j\,|\,x_i)P(i)}{P(j)} = \frac{g_{ij}}{\ell w_j}$.

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.

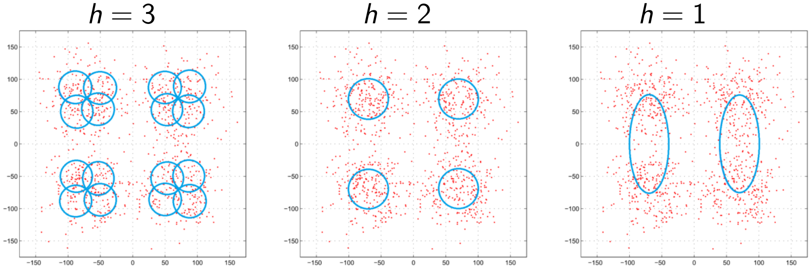
  Регуляризация $L(w,\theta) - \tau \sum\limits_{j=1}^{k} \ln w_j \to \max$:

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

HEM — иерархический EM-алгоритм

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

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



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

Резюме: три подхода к оцениванию плотностей

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