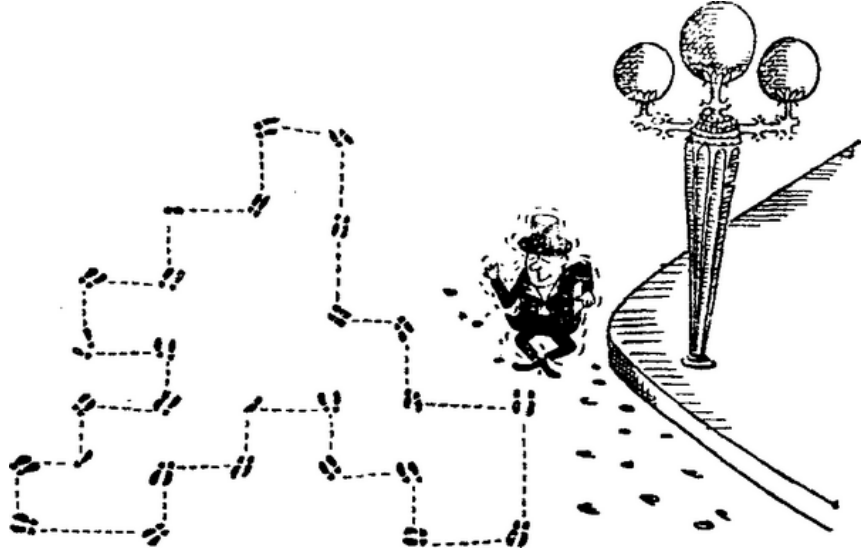

# Машинное обучение и Анализ больших данных

Лекция 4: *Восстановление плотности распределения объектов*

18/02/23

Кабалянц
Петр Степанович

# План

1. Напоминание о методе максимального правдоподобия.
2. Восстановление плотности распределения класса.
3. Ошибки задачи классификации.

# Метод максимального правдоподобия:

Дискретный случай:

$L(x\_1,\ldots,x\_n,\theta)=P(X=x\_1|\ \theta)*\ P(X=x\_2|\ \theta)*\ldots*\ P(X=x\_n|\ \theta)$

Найти θ при котором L максимальна

Непрерывный случай:

$L(x\_1,\ldots,x\_n,\theta)=f_\theta(x\_1)*\ f_\theta(x\_2)\ *\ldots*\ f_\theta(x\_n)$

Найти θ при котором L максимальна,

$f_\theta(x)$ – «предполагаемая» плотность распределения генеральной совокупности

Обычно ищут максимум натурального логарифма функции правдоподобия:

$\ln L(x\_1,\ldots,x\_n,\theta)=\ln P(X=x\_1|\ \theta)+\ \ln P(X=x\_2|\ \theta)+\ldots+\ \ln P(X=x\_n|\ \theta)$

$\ln L(x\_1,\ldots,x\_n,\theta)=\ln f_\theta(x\_1)+\ \ln f_\theta(x\_2)\ +\ldots+\ln f_\theta(x\_n)$

# Пример 1 метод максим. правдоподобия:

10, 4, 6, 8, 9, 1, 3, 8, 7, 5, 7, 4, 5, 9, 6, 7, 8, 7, 9, 6

n=20 – объем выборки,
m=10 – длина серий схемы Бернулли

X~B(10, p): $P(X=k)=C_{10}^{k} p^{k} (1-p)^{10-k}$

p*-?

# Пример 1 метод максим. правдоподобия:

10, 4, 6, 8, 9, 1, 3, 8, 7, 5, 7, 4, 5, 9, 6, 7, 8, 7, 9, 6

n=20 – объем выборки,  m=10 – длина серий схемы Бернулли

X~B(10, p): $P(X=k)=C_{10}^{k} p^{k} (1-p)^{10-k}$

p*-?

$L(x_1,\dots,x_{20},p)=P(X=10|p)* P(X=4|p)*\dots* P(X=6|p)=$

Найти p при котором L максимальна

$= C_{10}^{10} p^{10}(1- p)^{10-10} * C_{10}^{4} p^{4}(1- p)^{10-4} *\dots* C_{10}^{6} p^{6}(1-p)^{10-6} =$

$=(C_{10}^{10}\dots C_{10}^{6}) p^{10+4+\dots+6} (1- p)^{0+6+\dots+4} =(C_{10}^{10}\dots C_{10}^{6}) p^{129}(1-p)^{71}$

$\ln L=\ln(C_{10}^{10}\dots C_{10}^{6}) + 129\ln p + 71\ln(1-p)$

$(\ln L)' = 0+ 129/p – 71/(1-p) =0$

$129(1-p)=71p$  $200p=129$

$p^*=129/200= (x_1+\dots+x_n)/n*m = x_в/m$

# Пример 2 метод максим. правдоподобия:

10, 4, 6, 8, 9, 1, 3, 8, 7, 5, 7, 4, 5, 9, 6, 7, 8, 7, 9, 6

X~R(a, b): P(X<k)=(k-a)/(b-a), если a<k<b

$f_{a,b}(x)=\begin{cases} 0, x \notin [a,b] \\ \frac{1}{b-a}, \ x \in [a,b] \end{cases}$

a*, b*-?

$L(x_1,\ldots, x_{20},a,b)=\begin{cases} 0, \exists i: x_i \notin [a,b] \\ \frac{1}{(b-a)^{20}}, \forall i \ x_i \in [a,b] \end{cases}$

Найти a, b при котором L максимальна.

# Пример 2 метод максим. правдоподобия:

10, 4, 6, 8, 9, 1, 3, 8, 7, 5, 7, 4, 5, 9, 6, 7, 8, 7, 9, 6

X~R(a, b): P(X<k)=(k-a)/(b-a), если a<k<b

a*, b*-?

$L(x_1,\ldots, x_{20},a,b)=\begin{cases} 0, \exists i: x_i \notin (a,b) \\ \frac{1}{(b-a)^{20}}, \forall i\ x_i \in (a,b) \end{cases}$

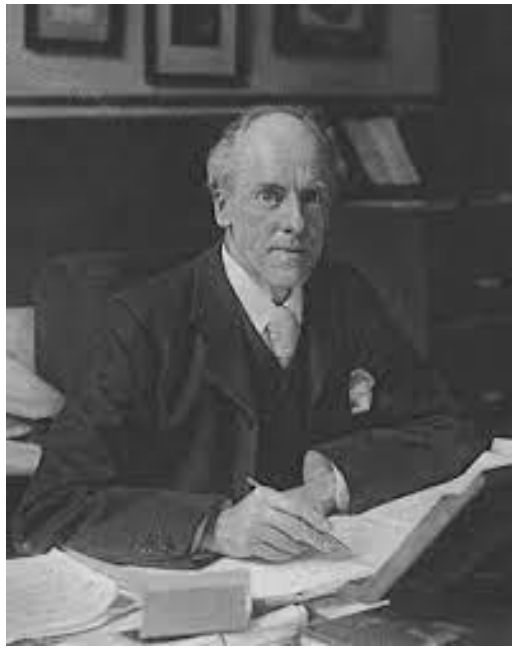
Найти a, b при котором L максимальна.

$a^*=\min x_i=1$

$b^*=\max x_i=10$

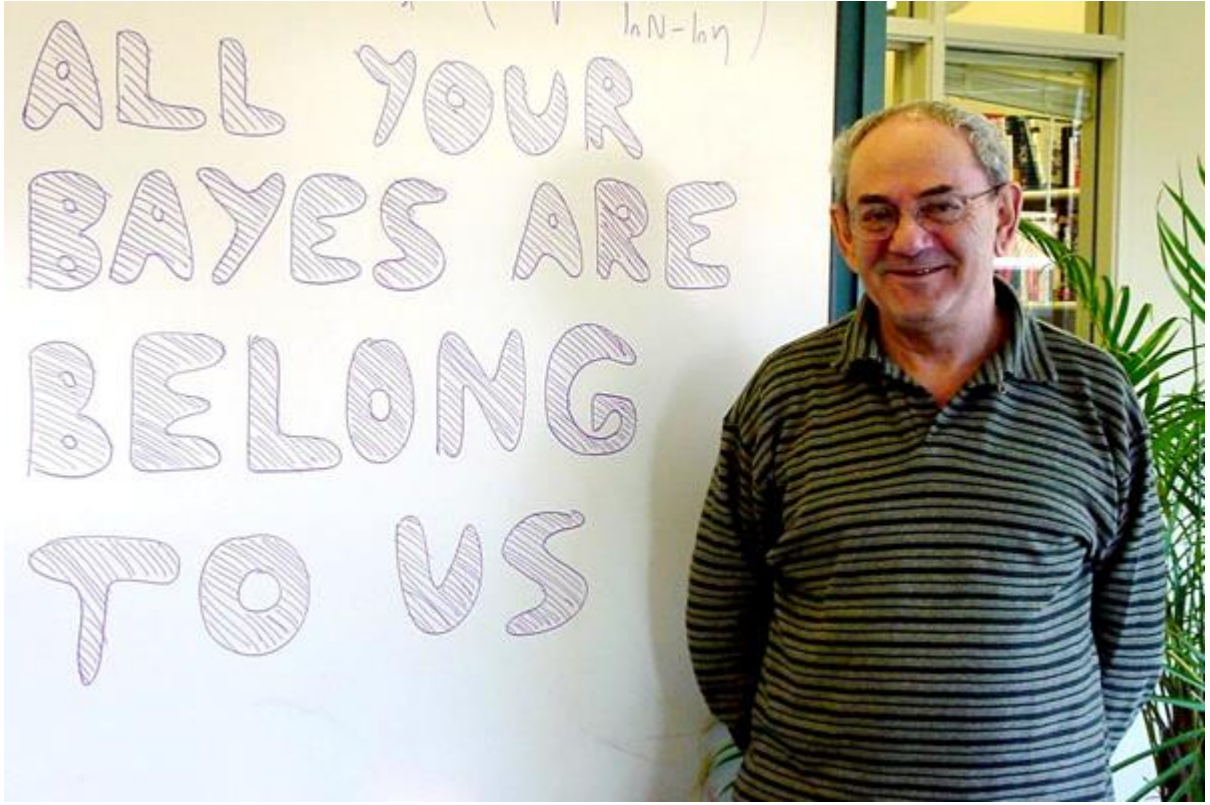
# Neyman–Pearson

Egon Pearson (сын)



K(C)arl Pearson (отец)

Jerzy Neyman (Юрий Нейман, родом из Бендер)




Вапник В. Н. Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979.

Было много критиков. Среди них был Владимир Вапник, математик и отец метода опорных векторов, одной из наиболее широко используемых моделей ИИ. Однажды мартовским днем 1995 года Вапник и Ларри Джакел, которые завербовали его и Лекуна в Bell Labs, заключили пари. Джакел заключил пари, что к 2000 году мы будем иметь представление о том, как работают глубокие искусственные нейронные сети. Вапник не согласился. Он также считал, что к 2005 году "никто в здравом уме не будет использовать нейронные сети, которые по сути похожи на те, что использовались в 1995 году". На карту был поставлен дорогой ужин, поэтому они составили ставку на бумаге и подписали ее - перед свидетелями. ЛеКун был третьим официальным подписантом, Боттту - неофициальным наблюдателем. Вапник выиграет первую половину пари. В 2000 году внутренняя работа нейронных сетей все еще была в значительной степени окутана тайной, и даже сейчас исследователи не могут математически точно определить, что заставляет их работать хорошо. К 2005 году глубокие нейронные сети все еще использовались в банкоматах и банках, и они во многом основывались на работах Лекуна середины 1980-х и начала 90-х.

# Задача восстановления плотности

**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 $\theta$ — параметр, $\varphi$ — 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 $\theta$.

# Случай n-мерной нормальной 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,
$\Sigma \in \mathbb{R}^{n \times n}$ — ковариационная матрица
(симметричная, невырожденная, положительно определённая).

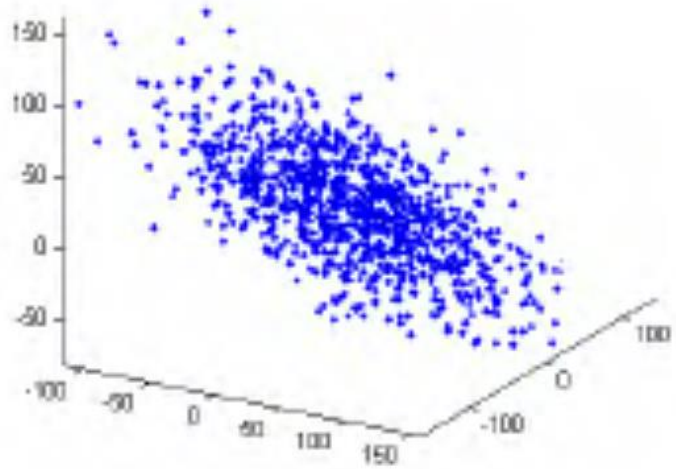
**Выборочные оценки максимального правдоподобия:**

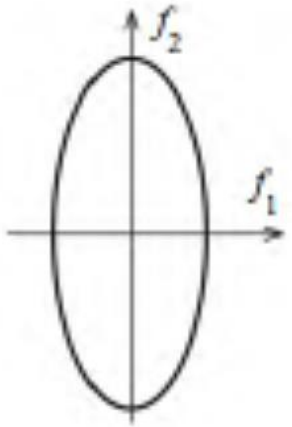
$$\frac{\partial}{\partial \mu} \ln L(\mu, \Sigma; X^\ell) = 0 \quad \Rightarrow \quad \hat{\mu} = \frac{1}{\ell} \sum_{i=1}^{\ell} x_i$$

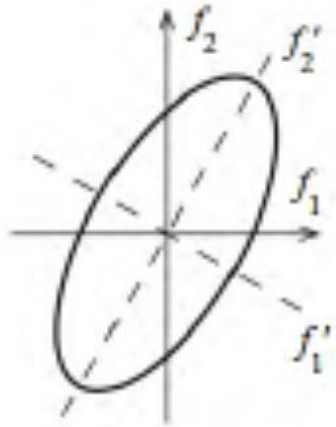
$$\frac{\partial}{\partial \Sigma} \ln L(\mu, \Sigma; X^\ell) = 0 \quad \Rightarrow \quad \hat{\Sigma} = \frac{1}{\ell} \sum_{i=1}^{\ell} (x_i - \hat{\mu})(x_i - \hat{\mu})^\top$$

# Нормальный эллипсоид

*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 собственные векторы,
$S = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ — собственные значения матрицы $\Sigma$

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

$x' = V^\mathsf{T}x$ — *декоррелирующее* ортогональное преобразование

# Проблема мультиколлинеарности

**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 $\tau$, сохраняя собственные векторы (параметр $\tau$ 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 восстановления смесей распределений

**Порождающая модель смеси распределений:**

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

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

**Задача 1:** при фиксированном $k$,
имея простую выборку $X^\ell = \{x_1, \ldots, x_\ell\} \sim p(x)$,
оценить вектор параметров $(w, \theta) = (w_1, \ldots, w_k, \theta_1, \ldots, \theta_k)$.

**Задача 2:** оценить ещё и $k$.

# Функция правдоподобия и EM-алгоритм

**Задача** максимизации логарифма правдоподобия

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

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

**Итерационный алгоритм Expectation–Maximization:**

начальное приближение параметров $(w, \theta)$;

**повторять**

оценка скрытых переменных $G = (g_{ij}),\ \ g_{ij} = P(j|x_i)$:

$G := \text{E-шаг}(w, \theta)$;

максимизация правдоподобия отдельно по компонентам:

$(w, \theta) := \text{M-шаг}(w, \theta, G)$;

**пока** $w, \theta$ и $G$ не стабилизируются;

# Теорема EM-алгоритма

**Теорема (необходимые условия экстремума)**

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

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

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

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

EM-алгоритм — это метод простых итераций для её решения

# Вероятностная интерпретация

**E-шаг** — это формула Байеса:

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

Очевидно, выполнено условие нормировки: $\sum_{j=1}^{k} g_{ij} = 1$.

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

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

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

# Доказательство теоремы

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

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

Приравниваем нулю производные:

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

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

■

# Алгоритм восстановления 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 $j = 1, \ldots, k$

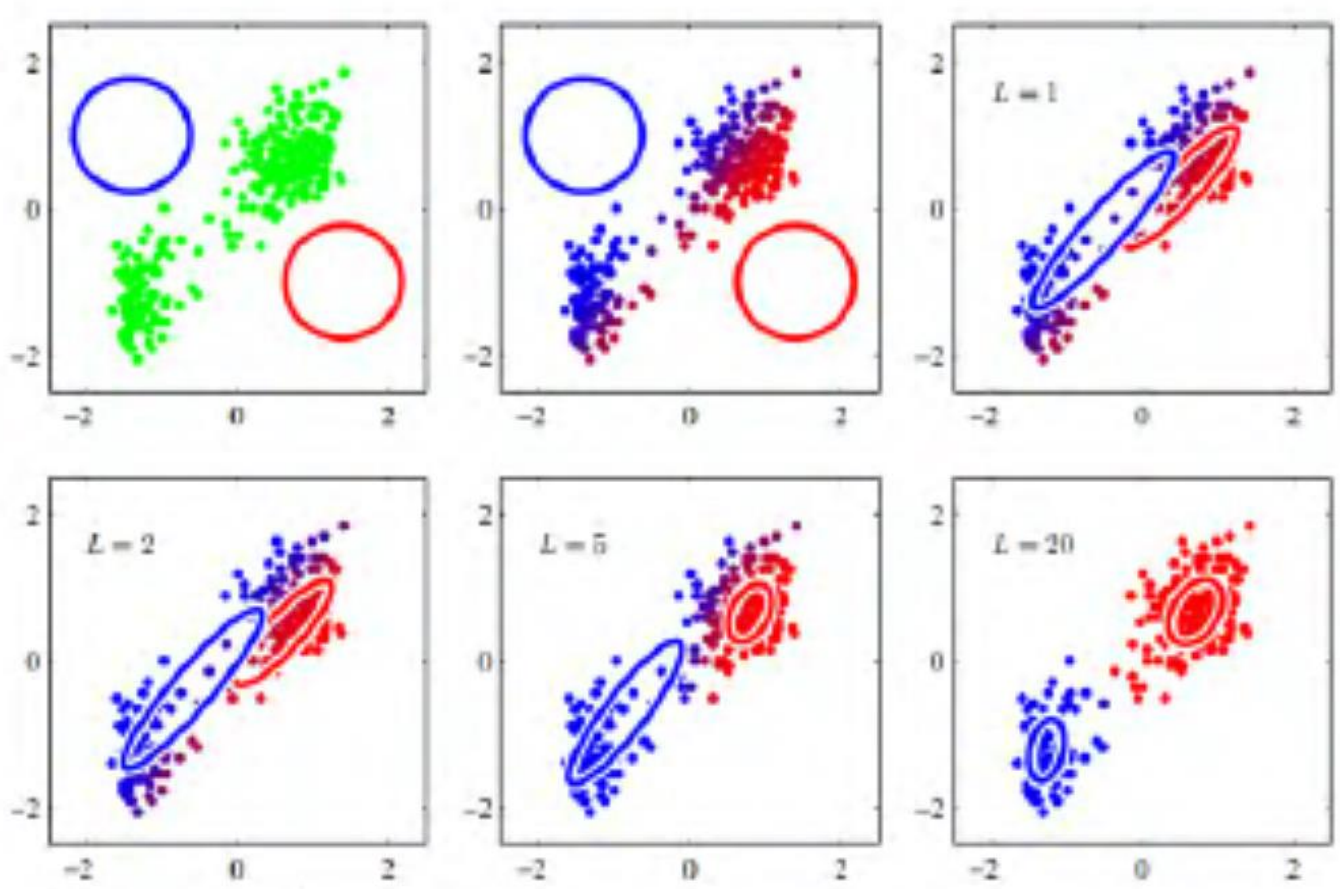
$$\theta_j := \arg\max_\theta \sum_{i=1}^\ell g_{ij} \ln \varphi(x_i, \theta); \qquad w_j := \frac{1}{\ell} \sum_{i=1}^\ell g_{ij};$$

**пока** $w_j, \theta_j$ 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 $L$:



# Модификация алгоритма

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

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

# Непараметрический подход

**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 $[a, b]$:

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

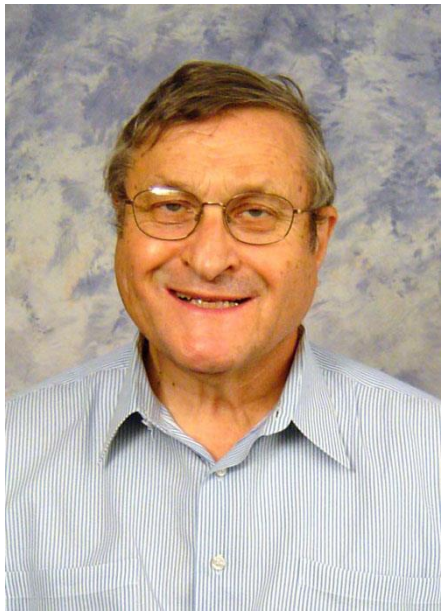
Эмпирическая оценка плотности по окну ширины $h$ (заменяем вероятность на долю объектов выборки):

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

$$p(x) = F'(x) = \lim_{\Delta x \to 0} \frac{\Delta F}{\Delta x} = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h} = \lim_{h \to 0} \frac{P(X < x+h) - P(X < x-h)}{2h} = \lim_{h \to 0} \frac{P(x-h < X < x+h)}{2h}$$

# Метод окна Эмануэля Парзена и Мюррея Розенблатта

Мишель Лоэв (Michel Loève), (1907-1979) ученик Поля Леви, учился в Париже, работал в Беркли, автор известного курса по теории вероятностей



Эмануэль Парзен (1929-2016), Emanuel Parzen, ученик Лоэва, учился в Гарварде и Беркли, работал в Стенфорде, в Техасском университете

Мюррей Розенблатт (1926-2019), университет Калифорнии

# Окно Парзена

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.

# Теорема Парзена

**Теорема (одномерный случай, $x_i \in \mathbb{R}$)**

**Пусть** выполнены следующие условия:

1) $X^\ell$ — простая выборка из распределения $p(x)$;
2) ядро $K(z)$ непрерывно и ограничено: $\int_X K^2(z)\,dz < \infty$;
3) последовательность $h_\ell$: $\lim\limits_{\ell\to\infty} h_\ell = 0$ 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 $x_i \in \mathbb{R}^n$?

# 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 $\rho(x, x')$:

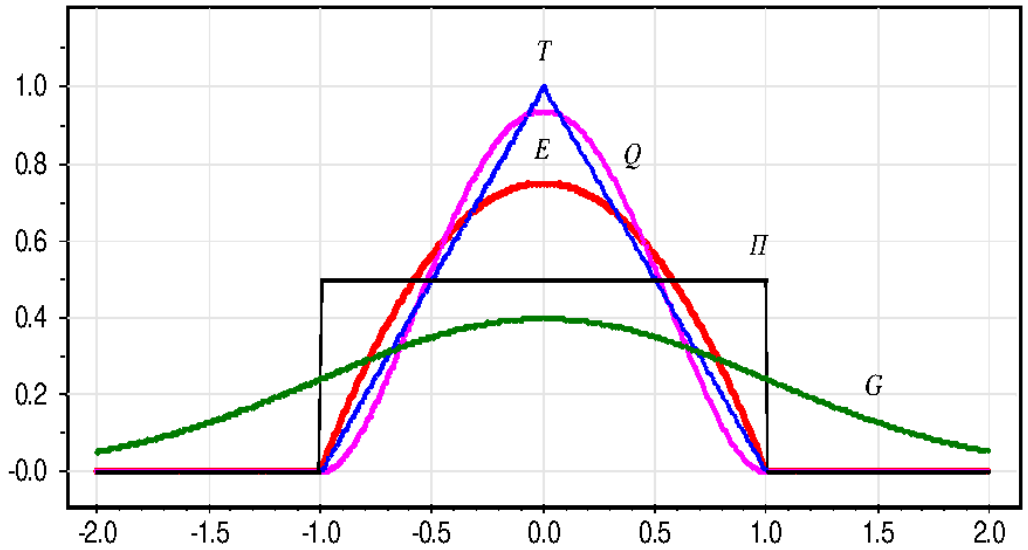
$$\hat{p}_h(x) = \frac{1}{\ell V(h)} \sum_{i=1}^{\ell} K\left( \frac{\rho(x, x_i)}{h} \right)$$

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 не влияет на качество восстановления

Функционал качества восстановления плотности:

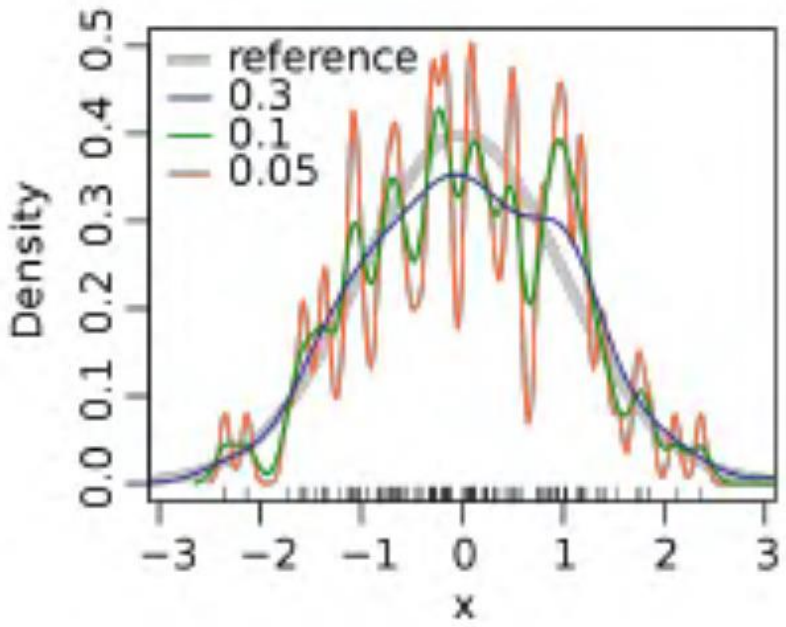
$$J(K) = \int_{-\infty}^{+\infty} \mathsf{E}\big(\hat{p}_h(x) - p(x)\big)^2\, dx.$$

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 | $\infty$ дифференцируема | 0.961 |
| Прямоугольное | $\hat{p}_h$ разрывна | 0.943 |

# Зависимость оценки плотности от ширина окна

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



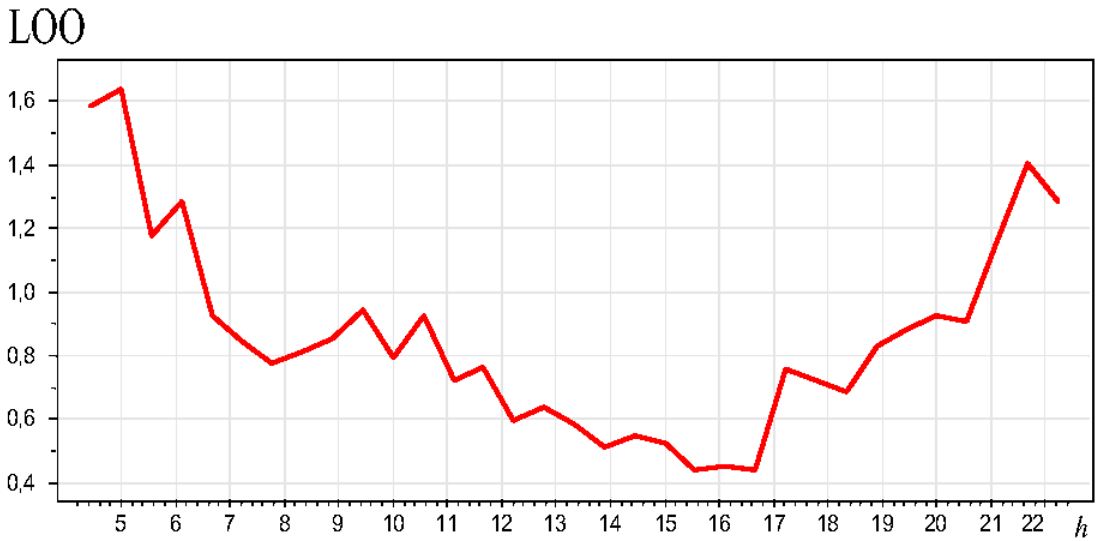

- Качество восстановления плотности существенно зависит от ширины окна $h$, но слабо зависит от вида ядра $K$
- При неоднородности локальных сгущений плотности можно задавать $h_k(x) = \rho\left(x, x^{(k+1)}\right)$, где $k$ — число соседей

# 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 $\mathrm{LOO}(k)$:



# Три подхода

1. **Параметрическое оценивание плотности:**

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

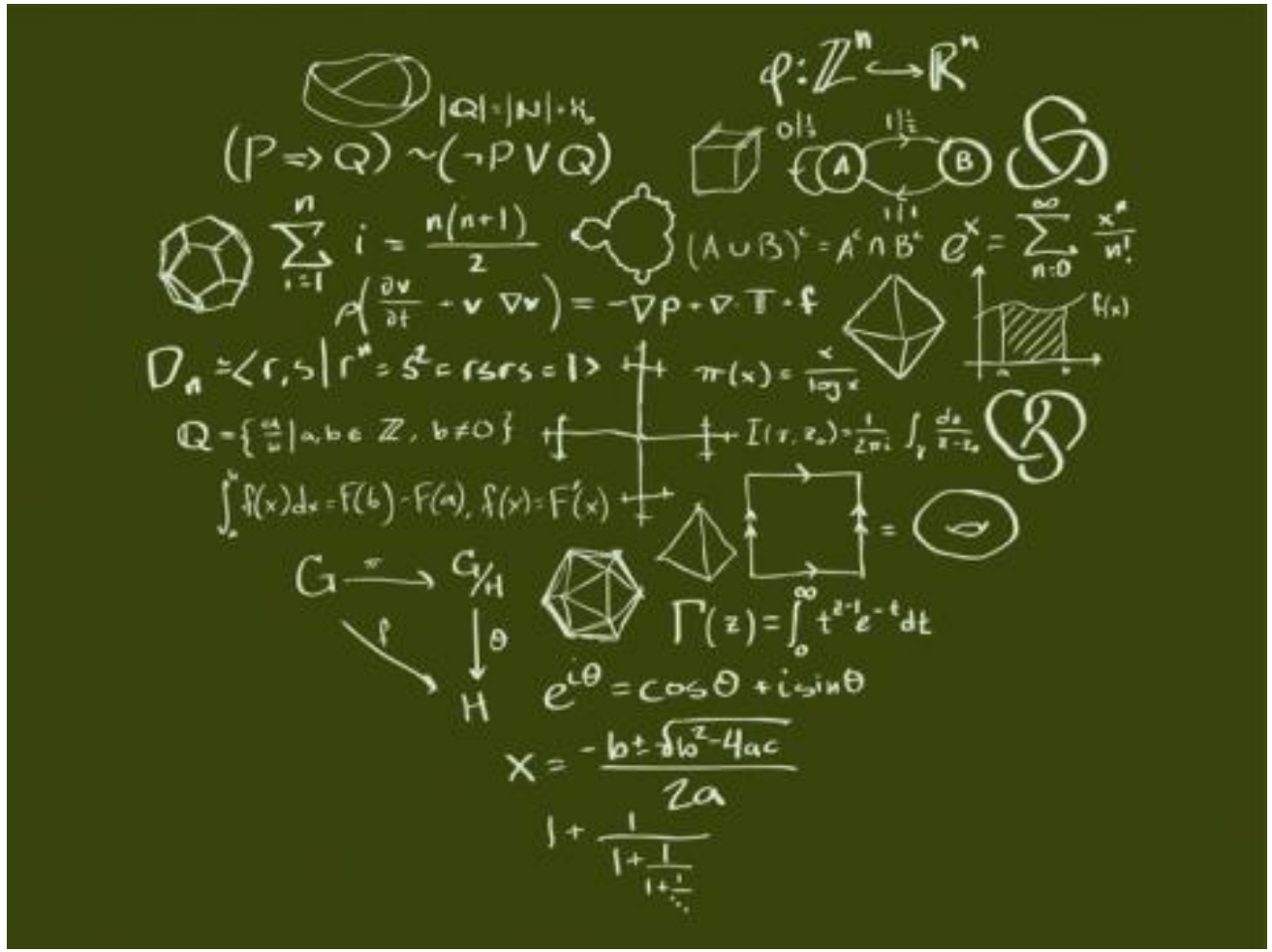
2. **Восстановление смеси распределений:**

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

3. **Непараметрическое оценивание плотности:**

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

# Математика поможет:




Спасибо за терпение!