2.1. Байесовская классификация

Y – зависимая категориальная переменная, $D_Y = \{\omega_1, ...\omega_K\}$,
ω_l - l-*й* класс $(D_Y = \{1, ..., \omega, ..., K\})$; $D_X = R^n$

Необходимо разработать классификатор, оптимальный по некоторому критерию:

$$x = X(a) \in D_X: \quad x \xrightarrow{f} \omega.$$

Критерий качества, например, - риск ошибочной классификации

f - называется решающей функцией или классификатором.

Пусть дано распределение $P(x, y)$. По правилу условной вероятности

$$P(x, y) = P(y)P(x \mid y) = P(x)P(y \mid x),$$

где

$P(y) = P(Y(a) = \omega) = P(\omega)$ - это априорная вероятность класса ω,

$P(x \mid y) = P_{\omega}(x)$ - это условное распределение переменной для класса ω (функция правдоподобия класса)

$P(y \mid x) = P(Y(a) = \omega \mid X(a) = x) = P_{x}(\omega)$ - это вероятность класса ω для точки x (апостериорная вероятность класса),

$$\sum_{\omega=1}^{K} P(\omega) = 1, \ \sum_{\omega=1}^{K} P_{x}(\omega) = 1$$

Байесовская решающая функция

$$f_B : x \to \omega^*: \ \omega^* = \arg\max_{\omega} P_x(\omega)$$

- максимизирует апостериорную вероятность класса.

По формуле Байеса: $P_x(\omega) = \dfrac{P(\omega)P_\omega(x)}{P(x)}$,

где $P(x) = \sum_{\omega=1}^{K} P(\omega)P_\omega(x)$ (по правилу полной вероятности).

Тогда, $\omega^* = \arg\max_{\omega} P(\omega)P_\omega(x)$.

Например, для K=2: f_B соответствует разбиению

$$D_{X,f_B}^{(1)} = \{x \mid P(1)P_1(x) \ge P(2)P_2(x)\}, \quad D_{X,f_B}^{(2)} = \overline{D_{X,f_B}^{(1)}}$$

Теорема. Для Байесовской решающей функции вероятность ошибки является минимальной

Доказательство. (K=2)

Пусть f произвольная решающая функция. Тогда

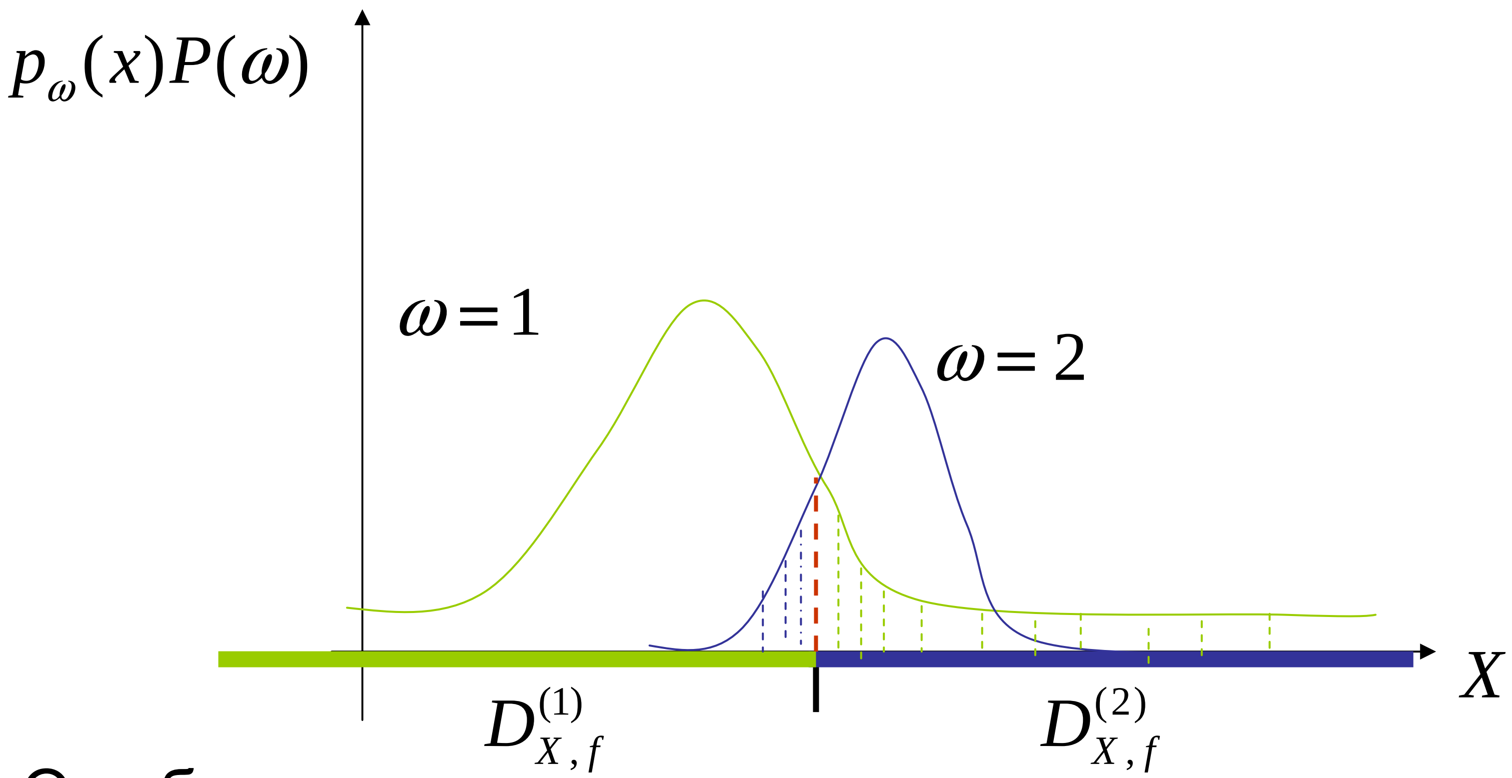
$$P_f^{err} = P(1) \int\limits_{x \in D_{X,f}^{(2)}} p_1(x)\,dx + P(2) \int\limits_{x \in D_{X,f}^{(1)}} p_2(x)\,dx =$$

$$= \int\limits_{D_X} h(x, f(x))\,dx,$$

где

$$h(x, f(x)) = \begin{cases} P(1)\,p_1(x) \quad for \quad x \in D_{X,f}^{(2)}, \\ P(2)\,p_2(x) \quad for \quad x \in D_{X,f}^{(1)}. \end{cases}$$

.



Ошибка минимальна, если

$$P(1)\,p_1(x) < P(2)\,p_2(x) \text{ for } x \in D_{X,f}^{(2)}$$

$$P(1)\,p_1(x) > P(2)\,p_2(x) \text{ for } x \in D_{X,f}^{(1)}.$$

$\Rightarrow$ x относится к классу с максимальным значением $P(\omega)\,p_\omega(x) = P_x(\omega)\cdot P(x)$ $\Rightarrow$ f это Байесовская решающая функция.

Оптимальная разделяющая функция

Пусть K=2, f_B это Байесовская решающая функция:

$$D^{(1)}_{X,f_B} = \{x \mid P(1)P_1(x) \geq P(2)P_2(x)\}, \quad D^{(2)}_{X,f_B} = \overline{D^{(1)}_{X,f_B}}.$$

$$L = \{x \mid P(1)P_1(x) = P(2)P_2(x)\}.$$

$$P(1)P_1(x) = P(2)P_2(x)$$

$$\frac{P(1)P_1(x)}{P(2)P_2(x)} = 1,$$

Тогда

$$l_0(x) = \log\frac{P_1(x)}{P_2(x)} + \log\frac{P(1)}{P(2)} = 0.$$

Функция, определяемая уравнением $l_0(x) = 0$ называется оптимальной разделяющей функцией.

Оценивание распределения по выборке

На практике распределение вероятностей неизвестно и должно оцениваться по обучающей выборке.

Оценка априорной вероятности класса:

$$\hat{P}(\omega) = N(\omega) / N,$$

где $N(\omega)$ это число объектов класса ω в таблице данных,

$$N = \sum_{\omega=1}^{k} N(\omega).$$

По закону больших чисел: $\hat{P}(\omega) \xrightarrow[p]{} P(\omega)$ при $N(\omega) \to \infty$

Как оценить априорное распределение переменной $P_{\omega}(x)$?

Параметрический подход: предполагается общая форма распределения $P_\omega(x)$ как функция с неизвестными параметрами

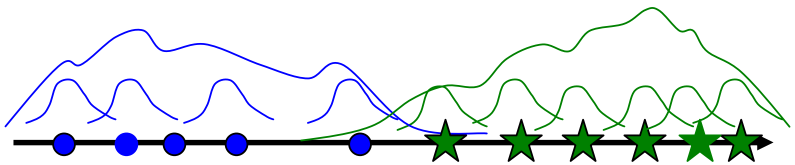
$$P_\omega(x) = \varphi(x,\theta);$$

или

$$P_\omega(x) = \sum_{l=1}^{L} w_l \, \varphi_l(x,\theta_l)$$

(модель смеси распределений).

Непараметрический: $\hat{p}_\omega(x) = \frac{1}{Nh^n} \sum_{i=1}^{N} \phi\left(\frac{x - x^{(i)}}{h}\right)$



Оценивание параметров

Пусть класс ω фиксирован. Предположим, что

$$P(x) = \varphi(x, \theta),$$

где $\theta \in \mathbf{\Theta}$, $\theta = (\theta_1, \ldots, \theta_m)$.

Для нахождения оценки $\hat{\theta}$ используется, например, метод максимального правдоподобия.
Логарифмическая функция правдоподобия:

$$L(\theta) = \log \prod_i P(x^{(i)}) = \sum_i \log(\varphi(x^{(i)}, \theta)); \ \ (\log \leftrightarrow \ln).$$

Оценка МП $\theta^*: L(\theta^*) = \max_\theta L(\theta)$.

Необходимое условие экстремума:

$$\frac{\partial}{\partial \theta_l} L(\theta_1, \ldots, \theta_m) = \sum_i \frac{\partial}{\partial \theta_l} \log(\varphi(x^{(i)}, \theta_1, \ldots, \theta_m)) = 0, \ \ l = 1, \ldots, m$$

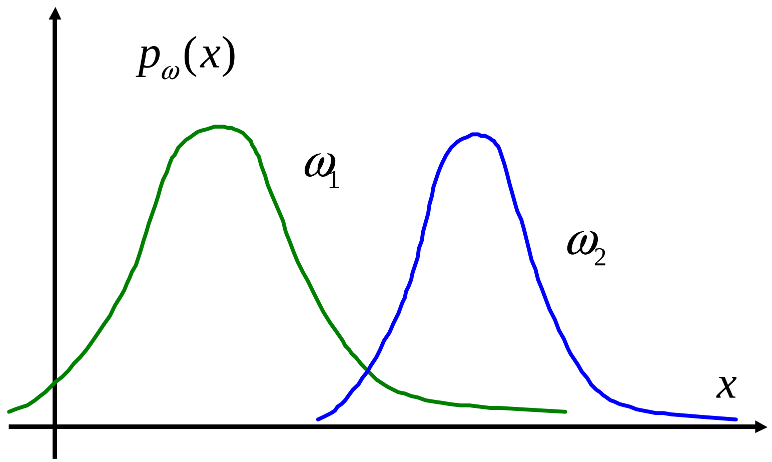
(система m уравнений; m - число неизвестных параметров)
$\rightarrow$ наилучший вектор $\theta^* = (\theta_1^*, \ldots, \theta_m^*)$.

Пример: 2 класса, 1 переменная, нормальное распределение

Предположим нормальное распределение для каждого класса:

$\omega_1: \; p_1(x) \sim N(\mu_1, \sigma_1); \; \omega_2: \; p_2(x) \sim N(\mu_2, \sigma_2);$

априорные вероятности $P(1)$, $P(2)$ $\quad (P(1)+P(2)=1)$.



$$p_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}, \; k = 1, 2.$$

Разделяющая функция (дискриминант):

$$l_0(x) = \ln \frac{p_1(x)}{p_2(x)} + \ln \frac{P(1)}{P(2)} = 0;$$

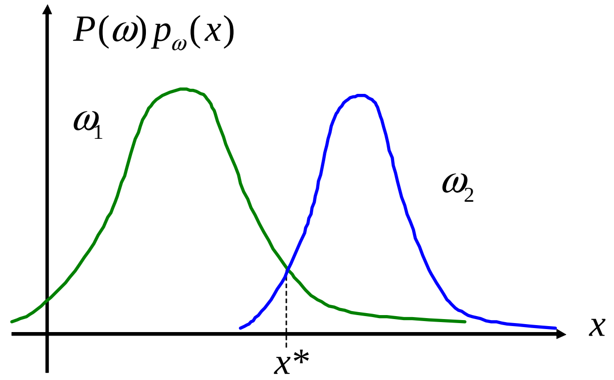
$$\ln \left(\frac{e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}}{\sqrt{2\pi}\sigma_1} \Bigg/ \frac{e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}}{\sqrt{2\pi}\sigma_2} \right) + \ln \frac{P(1)}{P(2)} = 0;$$

$$\ln \frac{\sigma_2}{\sigma_1} + \left(-\frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2} \right) + \ln \frac{P(1)}{P(2)} = 0;$$

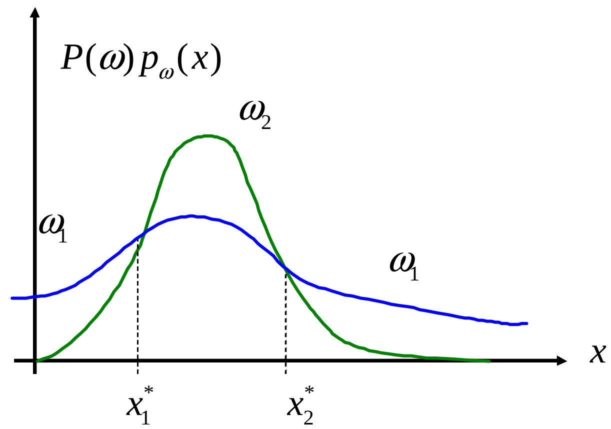
$$\frac{(x-\mu_2)^2}{\sigma_2^2} - \frac{(x-\mu_1)^2}{\sigma_1^2} = 2\ln \left(\frac{P(2)}{P(1)} \frac{\sigma_1}{\sigma_2} \right)$$

- квадратное уравнение

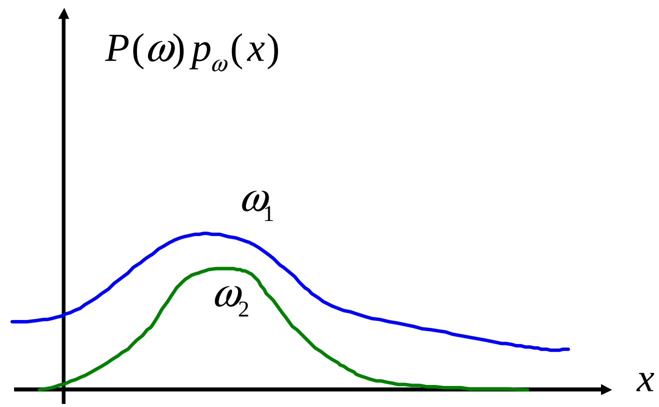
а) один корень x^*:



б) два корня x_1^*, x_2^*:



в) нет корней



Подставим вместо параметров из оценки в $l_0(x)$:

$$\hat{P}(1) = \frac{N(1)}{N}, \hat{P}(2) = \frac{N(2)}{N},$$

<u>Теорема</u>. Для нормального распределения оценки максимального правдоподобия равны:

$$\hat{\mu}_1 = \frac{1}{N(1)} \sum_{i:Y(i)=1} x^{(i)},\ \hat{\mu}_2 = \frac{1}{N(2)} \sum_{i:Y(i)=2} x^{(i)},$$

$$\hat{\sigma}_1^2 = \frac{1}{N(1)} \sum_{i:Y(i)=1} (x^{(i)} - \hat{\mu}_1)^2,\ \hat{\sigma}_2^2 = \frac{1}{N(2)} \sum_{i:Y(i)=2} (x^{(i)} - \hat{\mu}_2)^2$$

Пример: 2 класса, 2 переменные X_1, X_2, нормальное распределение

Нормальное распределение

$$p_1(x_1, x_2) = \frac{1}{2\pi |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right),$$

где $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$ - ковариационная матрица,

$\sigma_{ij} = \mathbf{E}\left[(X_i - \mu_i)(X_j - \mu_j) \right]$, $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ центроид первого класса,

$$p_2(x_1,x_2)=\frac{1}{2\pi\,|\Lambda|^{1/2}}\exp\left(-\frac{1}{2}(x-\nu)^T\Lambda^{-1}(x-\nu)\right),$$

где $\Lambda=\begin{pmatrix}\lambda_{11} & \lambda_{12}\\ \lambda_{12} & \lambda_{22}\end{pmatrix}$ это ковариационная матрица,

$\lambda_{ij}=\mathbf{E}\left[(X_i-\nu_i)(X_j-\nu_j)\right]$, $\nu=\begin{pmatrix}\nu_1\\ \nu_2\end{pmatrix}$ это центроид

второго класса.

Для нахождения оптимальной разделяющей функции решим уравнение

$$p_1(x_1,x_2)P(1)=p_2(x_1,x_2)P(2)\,.$$

Случай 1: независимые переменные и одинаковые дисперсии для каждого класса:

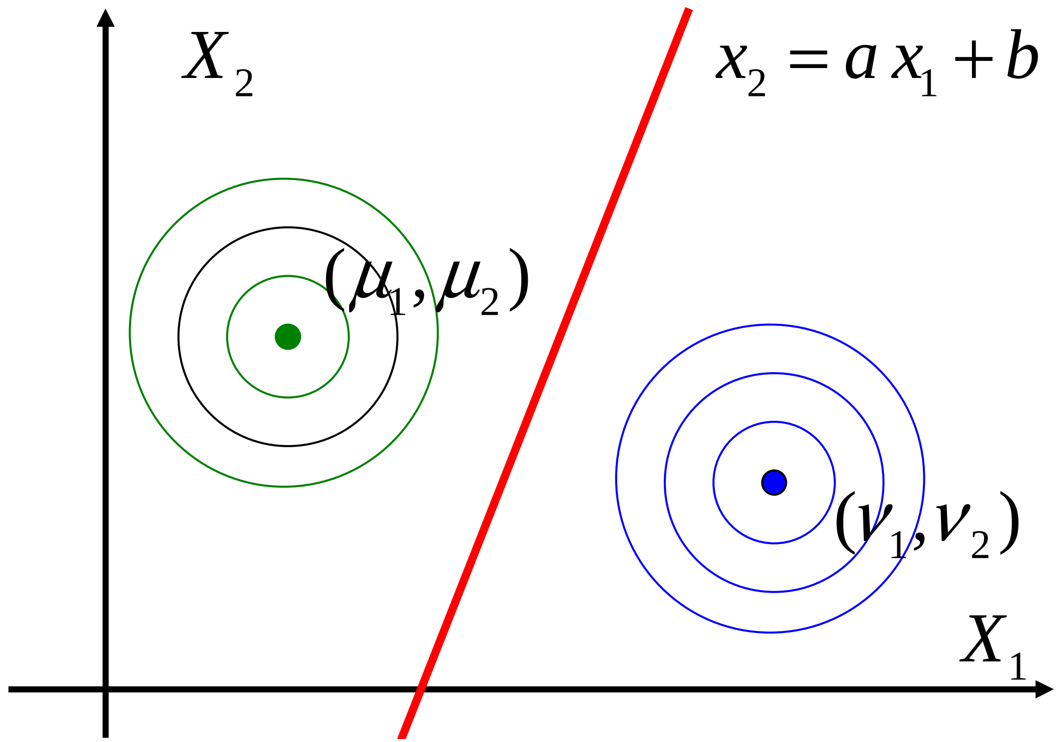
$$\Sigma = \Lambda = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}, \text{ then}$$

$$P(1)\exp\left\{-\frac{1}{2\sigma^2}\left[(x_1-\mu_1)^2+(x_2-\mu_2)^2\right]\right\}=$$

$$=P(2)\exp\left\{-\frac{1}{2\sigma^2}\left[(x_1-\nu_1)^2+(x_2-\nu_2)^2\right]\right\} \Rightarrow$$

$$x_2=\frac{\mu_1-\nu_1}{\nu_2-\mu_2}x_1+\left(\frac{\nu_1{}^2+\nu_2{}^2-\mu_1{}^2-\mu_2{}^2}{2(\nu_2-\mu_2)}-\frac{\sigma^2}{\nu_2-\mu_2}\ln\frac{P(2)}{P(1)}\right)$$

или $x_2 = a\,x_1 + b$.

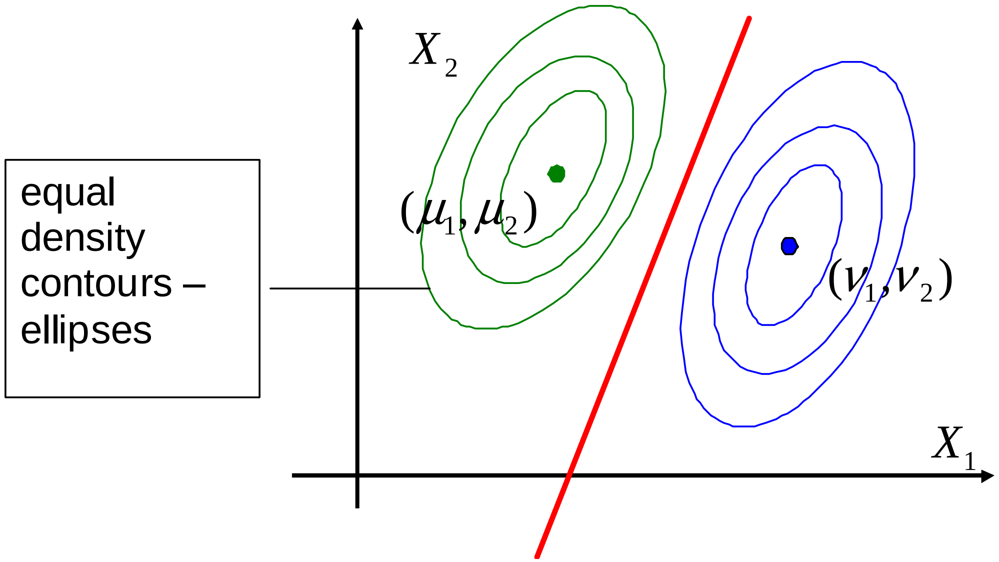


Поскольку параметры неизвестны, они оцениваются по выборке.

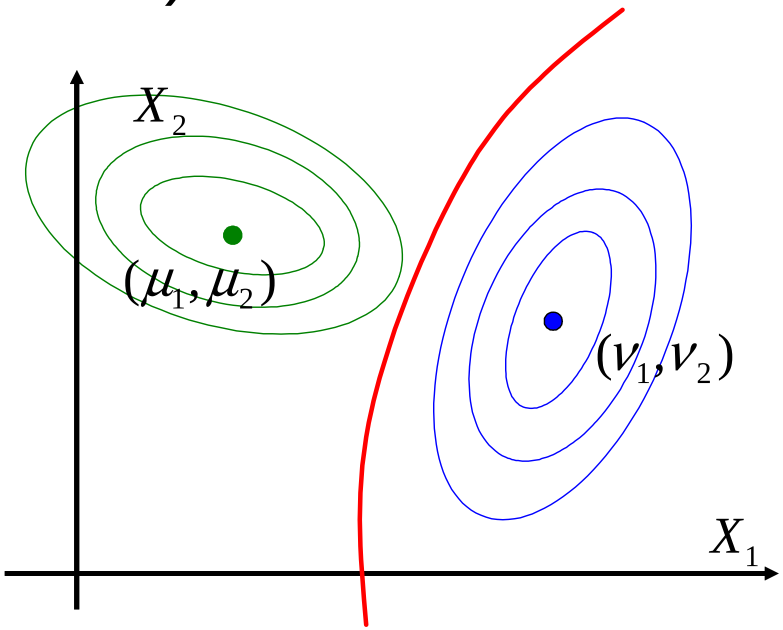
Случай 2. Пусть Σ, Λ произвольные, $\Sigma = \Lambda$. Тогда получаем линейную разделяющую функцию (Линейный дискриминант Фишера):

$$w_1 x_1 + w_2 x_2 = w_1 x_{01} + w_2 x_{02},$$

where $\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \Sigma^{-1}(\mu - \nu)$, $\begin{pmatrix} x_{01} \\ x_{02} \end{pmatrix} = \frac{1}{2}(\mu + \nu) - \frac{\ln P(1) - \ln P(2)}{(\mu - \nu)^T \Sigma^{-1}(\mu - \nu)}$



Случай 3. Для произвольных ковариационных матриц разделяющая функция - это кривая второго порядка (парабола, эллипс, окружность, прямая линия, гипербола)



$$l_0(x) = \frac{1}{2}\left[Q_2(x) - Q_1(x)\right] + \ln\frac{|\Lambda|^{1/2}}{|\Sigma|^{1/2}} + \ln\frac{P(1)}{P(2)} = 0,$$

где Q_1, Q_2 квадратичные формы, $Q_1 = (x-\mu)^T \Sigma^{-1}(x-\mu)$, $Q_2 = (x-\nu)^T \Lambda^{-1}(x-\nu)$ - квадратичный дискриминант.

Таким образом, предположение о нормальном распределении приводит к

линейной разделяющей функции (для одинаковых ковариационных матриц),

или

квадратичной разделяющей функции
(для произвольных ковариационных матриц).

Двойственный подход: определим семейство линейных или квадратичных разделяющих функций и найдем оптимальную функцию в этом семействе.

Оба этих подхода эквивалентны.

Байесовская классификация в случае дискретных переменных

1. Наивный Байесовский классификатор

Рассмотрим случай произвольных категориальных переменных $X_1,\dots,X_j,\dots,X_n$. Пусть $D_j=\{u_{1,j},\dots,u_{l_j,j}\}$, $l_j \ge 2$. Предположим, что все переменные независимы.

Тогда $x_j \in \{u_{1,j},\dots,u_{l_j,j}\}$:

$$P_\omega(x)=P_\omega(x_1)\cdot\ldots\cdot P_\omega(x_j)\cdot\ldots\cdot P_\omega(x_n),\ \omega=1,\dots,K.$$

Байесовская решающая функция:

$$x\to\omega^*:\ \omega^*=\arg\max_\omega P(\omega)P_\omega(x_1)\cdot\ldots\cdot P_\omega(x_n).$$

Оценки вероятностей:

$$\widehat{P}_\omega(x_j)=\frac{N_\omega(X_j=x_j)}{N(\omega)},\quad \widehat{P}(\omega)=\frac{N(\omega)}{N},$$

где $N_\omega(X_j=x_j)$ - число объектов класса ω, для которых X_j принимает значение x_i.

2. Бинарный классификатор, $K = 2$

$$X = \{X_1, \ldots, X_j, \ldots, X_n\},\ D_j = \{0,1\}$$

Обозначим:

$$P_1(x_j = 1) = p_j;$$

Тогда $P_1(x_j = 0) = 1 - p_j$.

$$q_j = P_2(x_j = 1);$$

Тогда $P_2(x_j = 0) = 1 - q_j$.

В результате получим

$$P_1(x_j) = p_j^{x_j}(1 - p_j)^{1 - x_j};$$

$$P_2(x_j) = q_j^{x_j}(1 - q_j)^{1 - x_j}.$$

$$l_0(x) = \log\frac{P_1(x)}{P_2(x)} + \log\frac{P(1)}{P(2)} =$$

$$= \sum_{j=1}^{n} \log[p_j^{x_j}(1-p_j)^{1-x_j}] - \sum_{j=1}^{n} \log[q_j^{x_j}(1-q_j)^{1-x_j}] + \log\frac{P(1)}{P(2)} =$$

$$= \sum_{j=1}^{n} \left[x_j \log\frac{p_j(1-q_j)}{q_j(1-p_j)} \right] + \sum_{j=1}^{n} \log\frac{1-p_j}{1-q_j} + \log\frac{P(1)}{P(2)} =$$

$$= \sum_{j=1}^{n} b_j x_j + b_0 = 0$$

тогда $b_j = \log\frac{p_j(1-q_j)}{q_j(1-p_j)}$, $b_0 = \sum_{j=1}^{n} \log\frac{1-p_j}{1-q_j} + \log\frac{P(1)}{P(2)}$.

$l_0(x) = 0$ - уравнение гиперплоскости в D_X.

Оптимальное разбиение $\beta_0 = \{D_X^{(1)}, D_X^{(2)}\}$, где $D_X^{(1)} = \{x \mid l_0(x) \geq 0\}$, $D_X^{(2)} = \overline{D_X^{(1)}}$.

Разделяющая функция (дискриминант):

$$\hat{l}_0(x) = \sum_{j=1}^{n} \hat{b}_j x_j + \hat{b}_0 = 0,$$

где $\hat{b}_j = \log \dfrac{\hat{p}_j(1-\hat{q}_j)}{\hat{q}_j(1-\hat{p}_j)}$, $\hat{b}_0 = \sum\limits_{j=1}^{n} \log \dfrac{1-\hat{p}_j}{1-\hat{q}_j} + \log \dfrac{\hat{P}(1)}{\hat{P}(2)}$.

оценки $\hat{P}(1) = \dfrac{N(1)}{N}, \hat{P}(2) = \dfrac{N(2)}{N}$,

$$\hat{p}_j = \frac{N_1(x_j = 1)}{N(1)}, \; \hat{q}_j = \frac{N_2(x_j = 1)}{N(2)},$$

или Байсовские оценки *:

$$\hat{p}_j = \frac{N_1(x_j = 1) + 1}{N(1) + 2}, \ \hat{q}_j = \frac{N_2(x_j = 1) + 1}{N(2) + 2},$$

где $N(1)$, $N(2)$ - число объектов первого и второго класса ($N = N(1) + N(2)$), $N_1(x_j = 1), N_2(x_j = 1)$ -число объектов первого и второго класса, для которых x_j принимает значение 1.

Borovkov A. A. On the Problem of Pattern Recognition // Theory Probab. Appl. 1971. 16(1), 141–144.

Непараметрический подход

Идея: чтобы классифицировать x, необходимо оценит апостериорную вероятность только в этой точке

Пусть класс ω фиксирован. Необходимо найти функцию плотности $p(x)$, используя

$$x^{(1)},\ldots,x^{(N)},$$

где $x^{(i)} = \left(x_1^{(i)},\ldots,x_n^{(i)}\right) \in \mathbf{R}^n$.

Предположим, что $X^{(1)},\ldots,X^{(N)}$ независимы и одинаково распределены;

$$X^{(i)} \sim p(x).$$

Пусть V небольшая область около x;

$|V|$ - это объем;

$k =$ число точек в V.

Оценить плотность можно так

$$\boxed{p_N(x) = \frac{k}{N} / |V_N|}$$

Желаемое свойство - состоятельность, т.е.

$p_N(x) \to p(x)$ когда $N \to \infty$.

Есть два способа:

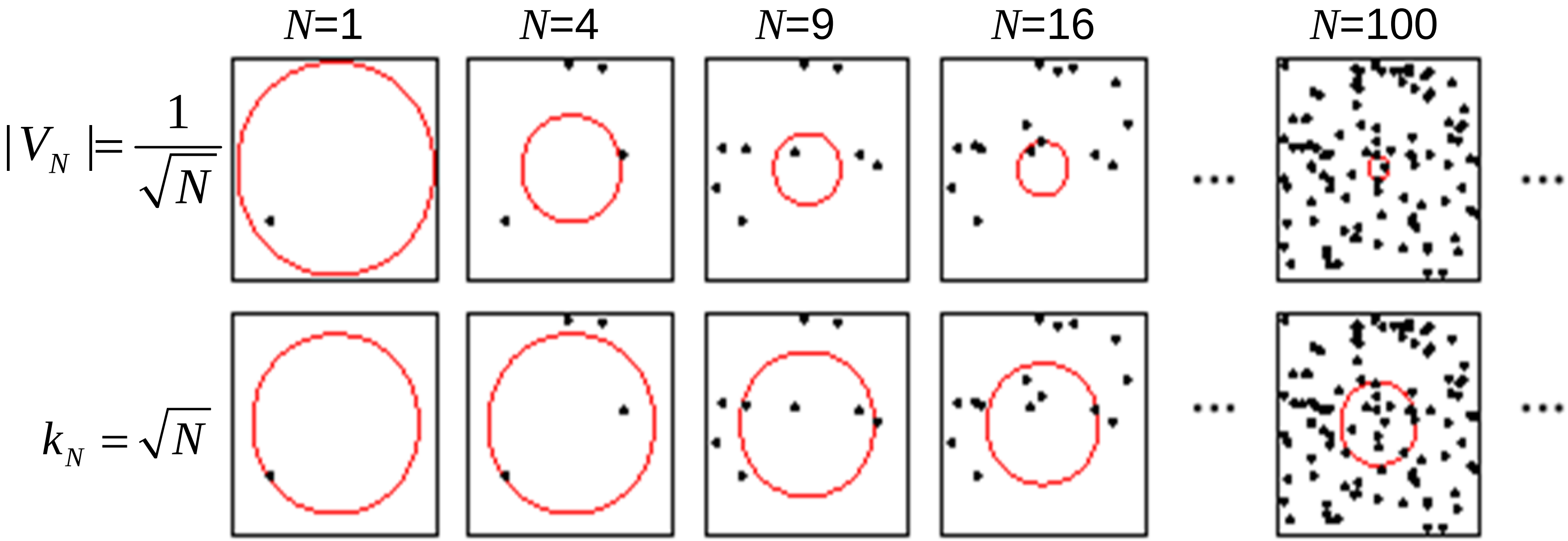
Окно Парзена

V_N - серия регулярных областей (гиперсфера, гиперкуб),

объем $|V_N| = \dfrac{V_0}{\sqrt{N}}$, где V_0 это параметр.

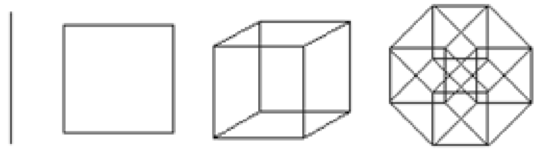
2) k_N ближайших соседей (nearest neighbors): $k_N = \sqrt{N}$;

Размер V_N растет пока V не будет включать k_N ближайших к x точек.



N=1
N=4
N=9
N=16
N=100
$|V_N| = \frac{1}{\sqrt{N}}$
$k_N = \sqrt{N}$

Для фиксированного N, рассмотрим окно Парзена:



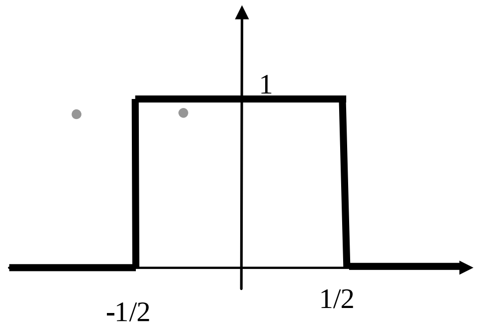
V - это n-мерный гиперкуб с центром в точке x.

Обозначим $u = \dfrac{x - x'}{h}$, где x, x' произвольные точки,

h - это длина ребра гиперкуба (размер окна), $|V| = h^n$.

Пусть $\phi(u)$: $\phi(u) \geq 0$,

$$\int_{x \in D} \phi(u) du = 1, \quad \phi(u) = \begin{cases} 1, & if \ |u_j| \leq \frac{1}{2}, \ j = 1, \ldots, n \\ 0, & othervise \end{cases}.$$



Число точек, попавших в V:

$$k = \sum_i \phi\left(\frac{x - x^{(i)}}{h}\right).$$

Оценка функции плотности: $p_N(x) = \dfrac{1}{Nh^n} \sum_i \phi\left(\dfrac{x - x^{(i)}}{h}\right)$

Обобщение: оценка Розенблатта-Парзена:

$$\boxed{p_N(x) = \frac{1}{Nh^n} \sum_i \phi\left(\frac{\| x^i - x \|}{h}\right)},$$

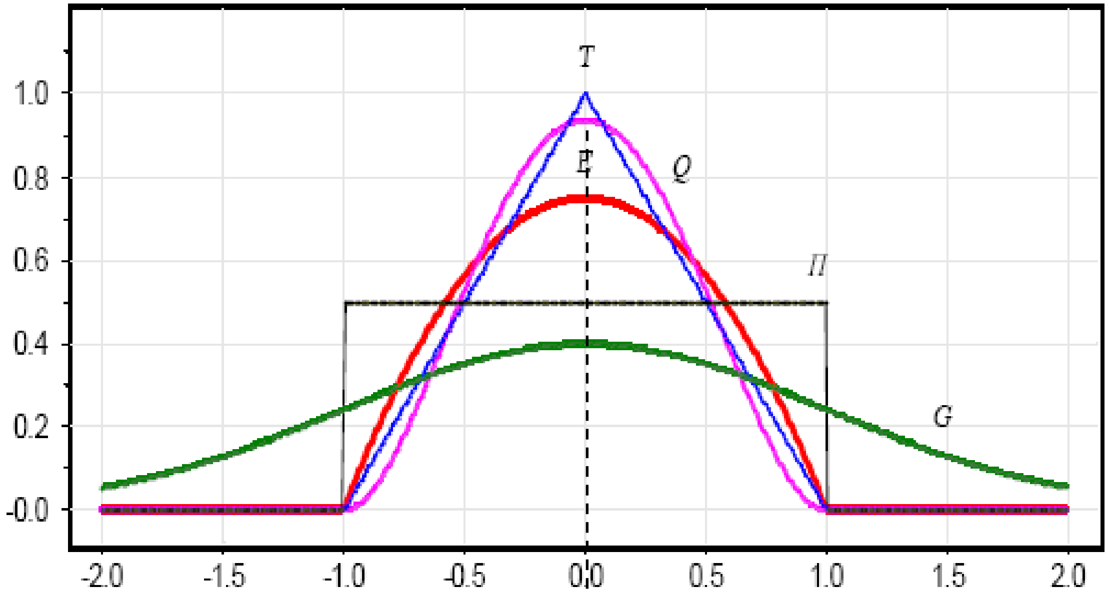
где $\phi(\cdot)$ - ядерная функция:

- $\phi(r) \geq 0$
- $\phi(\cdot)$ четная функция;
- $\int \phi(r) dr = 1$;
- $r > 0$.

В общем случае ядро это функция $\phi(x, x')$.

$$V_N = \frac{|V_1|}{\sqrt{N}} = \frac{h_1^n}{\sqrt{N}},$$ где h_1 это параметр

Основные типы ядер:



$E(r) = \frac{3}{4}(1 - r^2)\left[|r| \leqslant 1\right]$ — Epanechnikov kernel;

$Q(r) = \frac{15}{16}(1 - r^2)^2\left[|r| \leqslant 1\right]$ – quartic;

$T(r) = \left(1 - |r|\right)\left[|r| \leqslant 1\right]$ — triangular;

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

$\Pi(r) = \frac{1}{2}\left[|r| \leqslant 1\right]$ — rectangular;

$$[\mathbf{I}] = \begin{cases} 1, & \mathbf{I} = \mathit{true} \\ 0, & \mathbf{I} = \mathit{false} \end{cases}$$

Многомерный случай

Ядро Епанечникова:

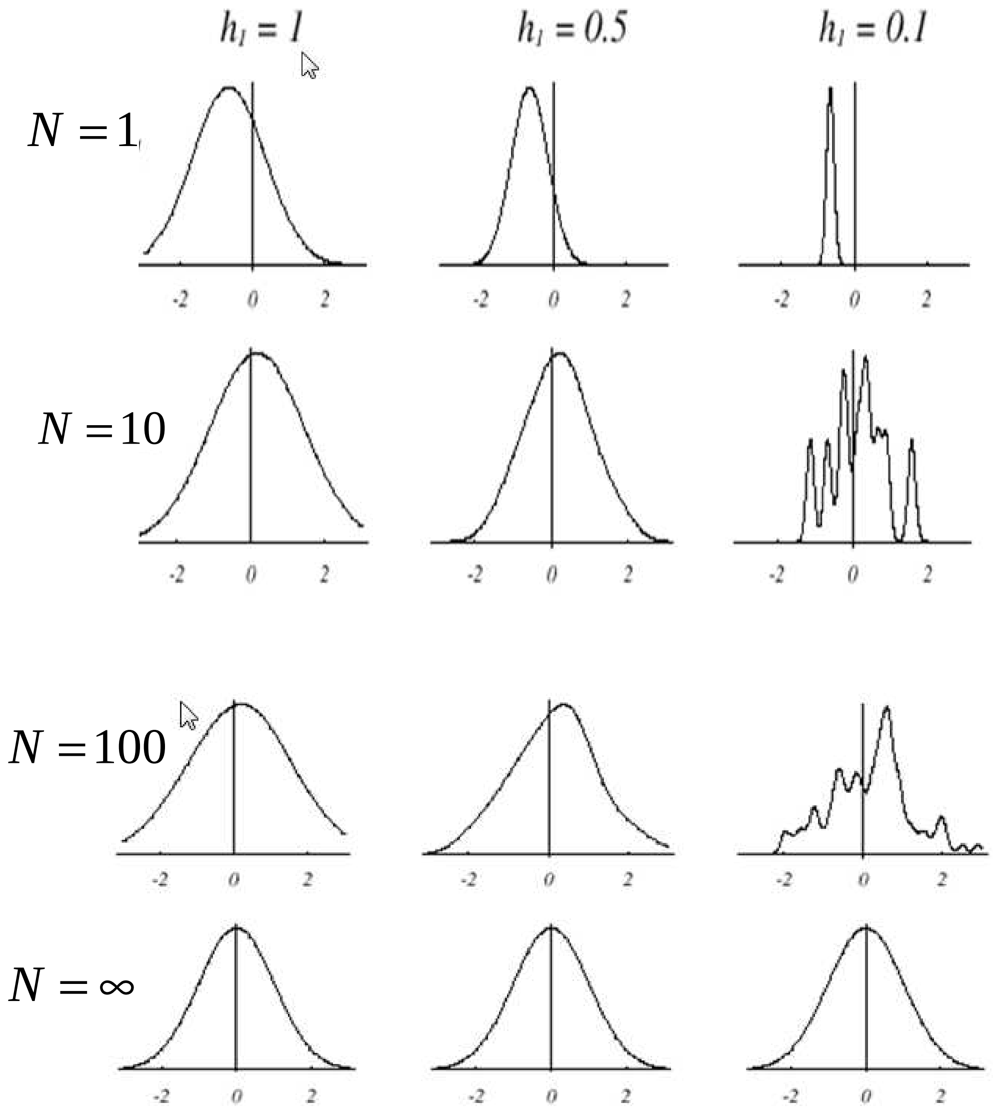
$$\phi_{\mathrm{E}} = \begin{cases} \frac{1}{2} V_n^{-1}(n+2)(1-r^2), & \text{if} \quad \mathrm{r} \le 1, \\ 0, & \text{else}, \end{cases}$$

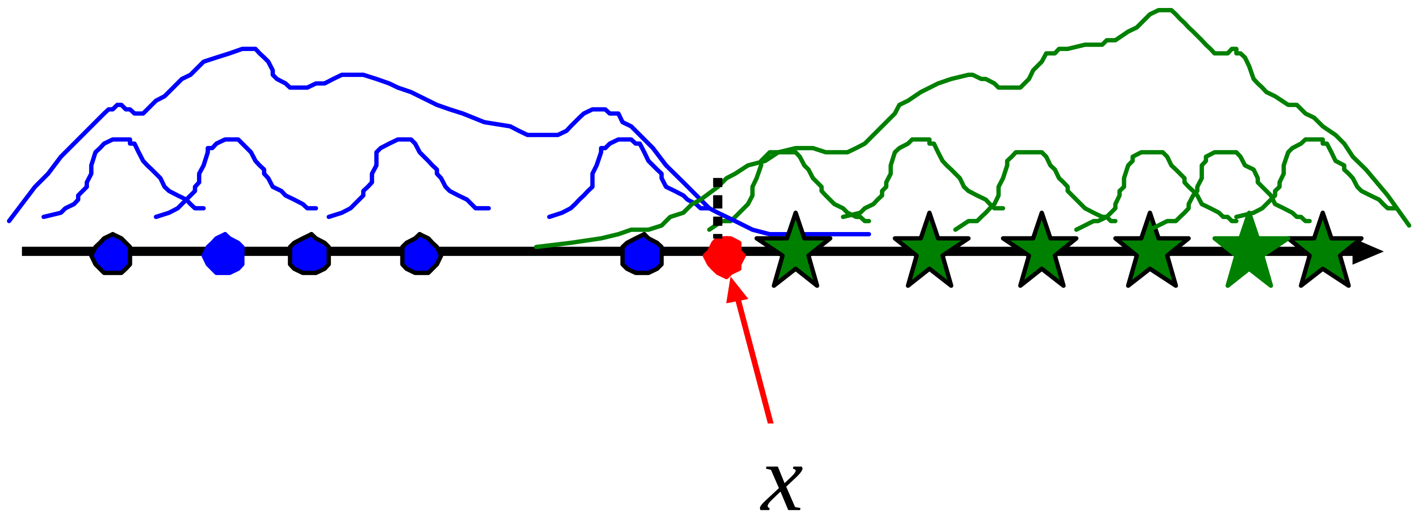
где V_n – объем n-мерной единичной сферы

Многомерное нормальное ядро:

$$\phi_N = (2\pi)^{-n/2} e^{-\frac{r^2}{2}}.$$

Пример (нормальное ядро)





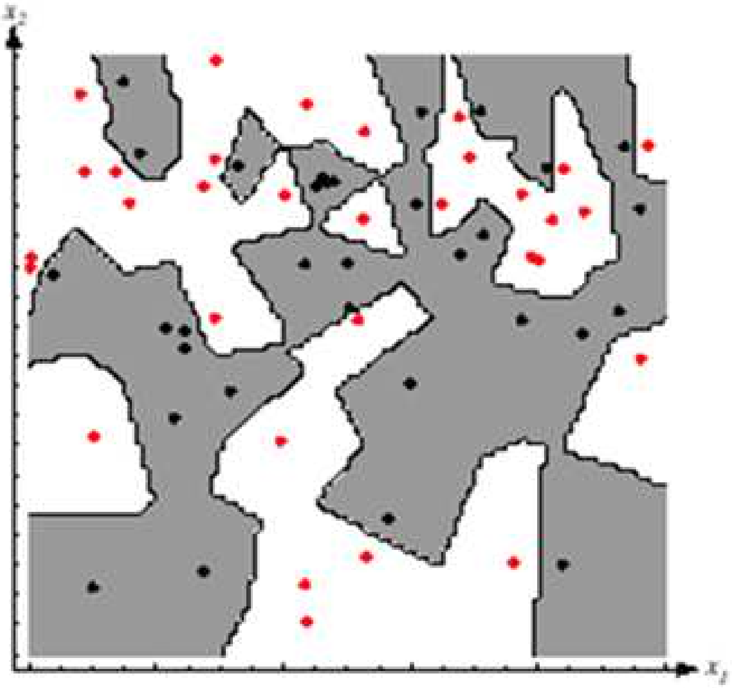
Оптимальная выборочная разделяющая функция:

$$\hat{f}_B(x) = \arg\max_{\omega} \hat{P}(\omega)\, p_{N,\omega}(x) =$$

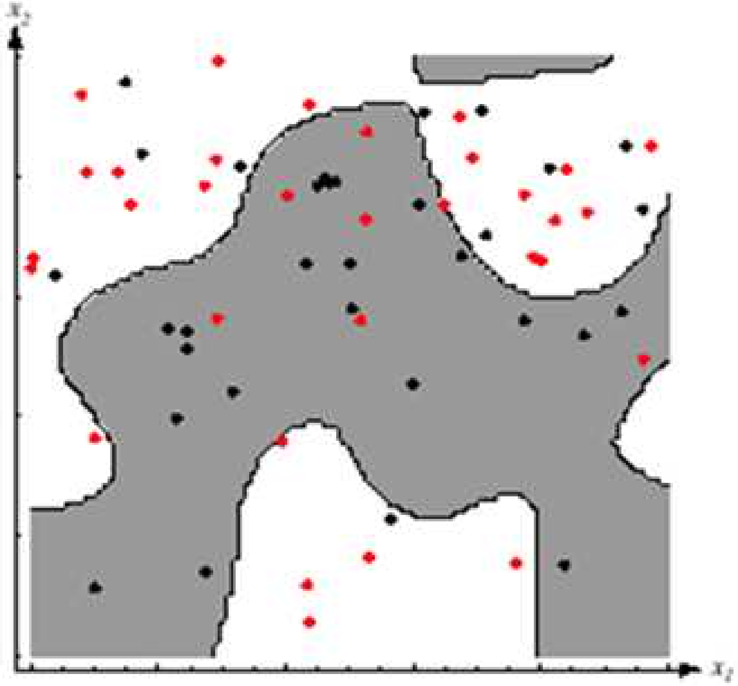
$$= \arg\max_{\omega} \frac{N(\omega)}{N} \frac{1}{N(\omega)h^n} \sum_{i:y^{(i)}=\omega} \phi\left(\frac{\| x - x^{(i)} \|}{h}\right) =$$

$$= \arg\max_{\omega} \sum_{i:y^{(i)}=\omega} \phi\left(\frac{\| x - x^{(i)} \|}{h}\right).$$

Влияние размера окна на решающую функцию



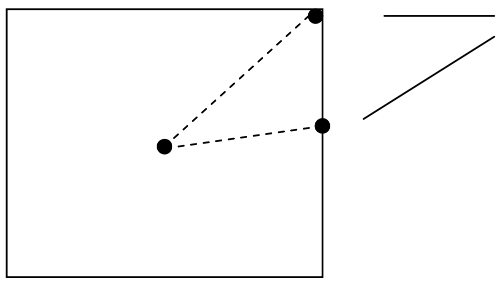
small h



large h

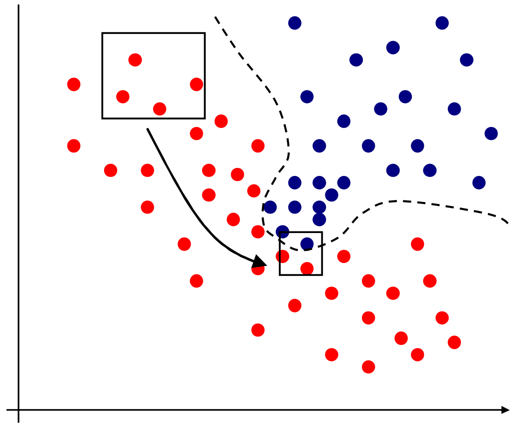
Недостатки окна Парзена

- для большой размерности $\rho(x, x')$ может быть существенно отличным от $\rho(x, x'')$, где x' тока в вершине куба, x'' точка на ребре куба.



equally dependent on density estimation, though distances to the center significantly differ

- для фиксированного объема V_N, число входящих объектов может существенно отличаться при разных x.



Метод k ближайших соседей kNN

Пусть V зависит от x: $V = V_N(x)$.

Решения для x зависит от k ближайших точек:
размер окна увеличивается пока $V_N(x)$ не будет содержать ровно k точек из N

$$(k = \sum_{\omega} k^{(\omega)}).$$

Оптимальная решающая функция:

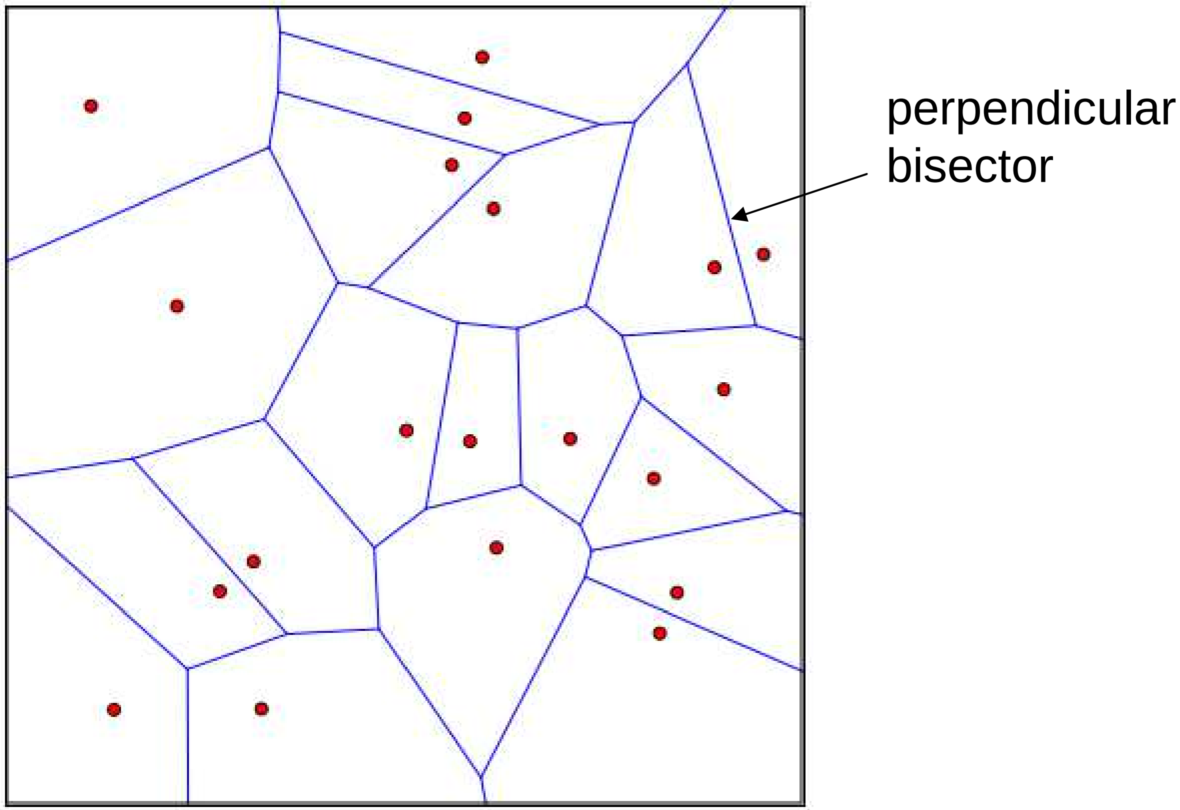
$$\hat{f}_B(x) = \arg\max_{\omega} \hat{P}(\omega) \frac{k^{(\omega)} / N(\omega)}{|V_N|} =$$

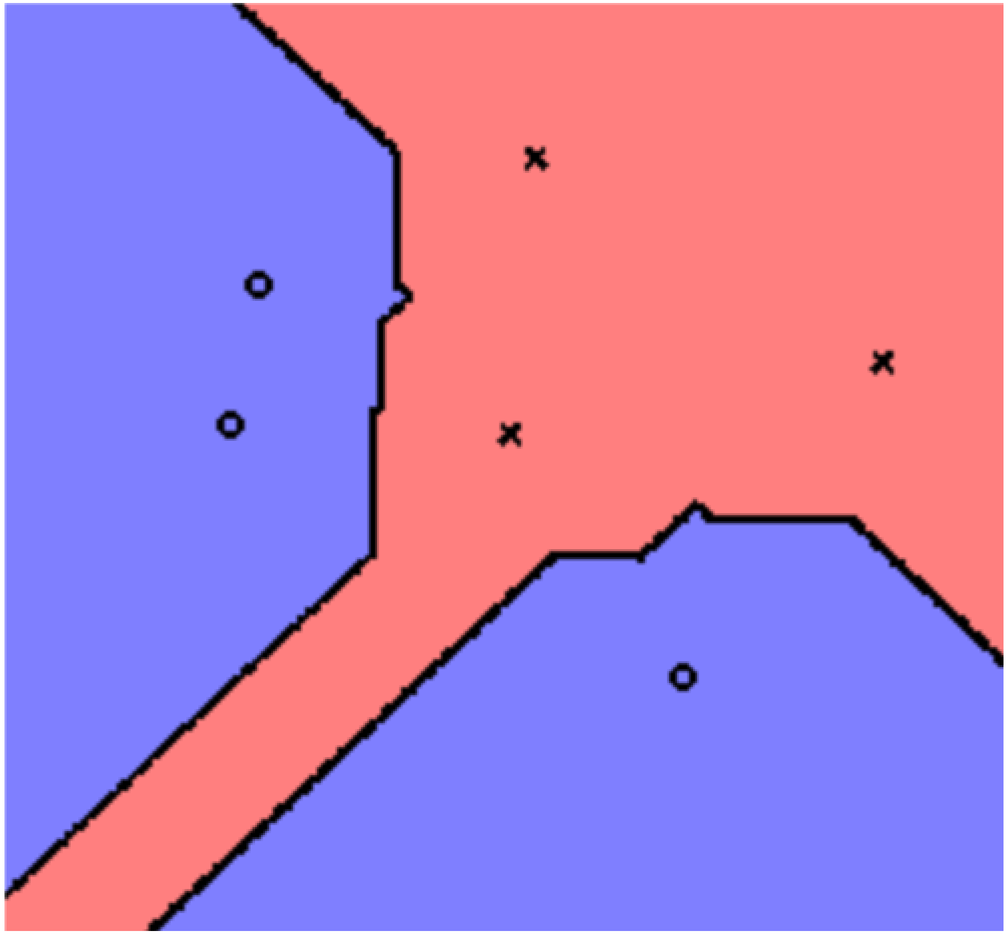
$$= \arg\max_{\omega} \frac{N(\omega)}{N} \frac{k^{(\omega)}}{N(\omega)} \frac{1}{|V_N|} = \arg\max_{\omega} k^{(\omega)}.$$

kNN запоминает все объекты обучающей выборки и затем сравнивает их с точкой x.

Метод ближайшего соседа $(k = 1)$:
каждой точке $x \in D$ соответствует класс ближайшей точки

Voronoi diagram





Region that corresponds to each class, is a union of polyhedrons in n-dimensional variable space

Вероятность ошибки для метода ближайшего соседа (для бинарной классификации)

Теорема. Вероятность ошибки для метода ближайшего соседа превосходит вероятность ошибки оптимального байесовского классификатора не более чем в 2 раза при достаточно большом объеме обучающей выборки.