

Bayesovský klasifikátor priradí objekt triede, kde $P(\omega_i|\mathbf{x})$ je maximálne

$$P(\omega_i|\mathbf{x}) = \frac{P(\mathbf{x}|\omega_i)P(\omega_i)}{P(\mathbf{x})}$$

$$\frac{P(\mathbf{x}|\omega_i)P(\omega_i)}{P(\mathbf{x})} \geq \frac{P(\mathbf{x}|\omega_j)P(\omega_j)}{P(\mathbf{x})}$$

Rozhodovacia funkcia

$$d_i(\mathbf{x}) = P(\mathbf{x}|\omega_i)P(\omega_i)$$

Ako určiť pravdepodobnosti

$$d_i(\mathbf{x}) = P(\mathbf{x}|\omega_i)P(\omega_i)$$

Pre *kategorické atribúty*

$$P(\mathbf{x} | \omega_i) = \prod_{k=1}^d P(x_k | \omega_i)$$

Naivný klasifikátor

$$P(x_k | \omega_i) = \frac{N^{i,k}}{N^i}$$

Pravdepodobnosť triedy ω_i je $\frac{N^i}{N}$

N # všetkých príkladov

N^i # tých príkladov, ktoré patria do triedy ω_i

$N^{i,k}$ # tých príkladov, ktoré patria do triedy ω_i a príznak k má hodnotu x_k

Príklad

$$d_i(\mathbf{x}) = P(\mathbf{x}|\omega_i)P(\omega_i)$$

Mám červený Touareg, ukradnú mi ho?

$P(\text{Yes}|\text{Red, SUV, Domestic})$

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

$P(\text{Yes})=P(\text{No})$

$P(\text{Red, SUV, Domestic}|\text{Yes}) \quad ? \quad P(\text{Red, SUV, Domestic}|\text{No})$

$$P(\mathbf{x} | \omega_i) = \prod_{k=1}^d P(x_k | \omega_i)$$

$P(\text{Red}|\text{Yes})$

$P(\text{SUV}|\text{Yes})$

$P(\text{Domestic}|\text{Yes})$

$P(\text{Red}|\text{No})$

$P(\text{SUV}|\text{No})$

$P(\text{Domestic}|\text{No})$

Do ktorej triedy patrí tento zákazník?

X =
(age <=30,
Income = medium,
Student = yes
Credit_rating = Fair)

age	income	student	credit rating	buys computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

Odhad parametrov

Parametrické metódy

Unimodálne rozdelenie – odhadneme parametre normálneho alebo iného rozdelenia

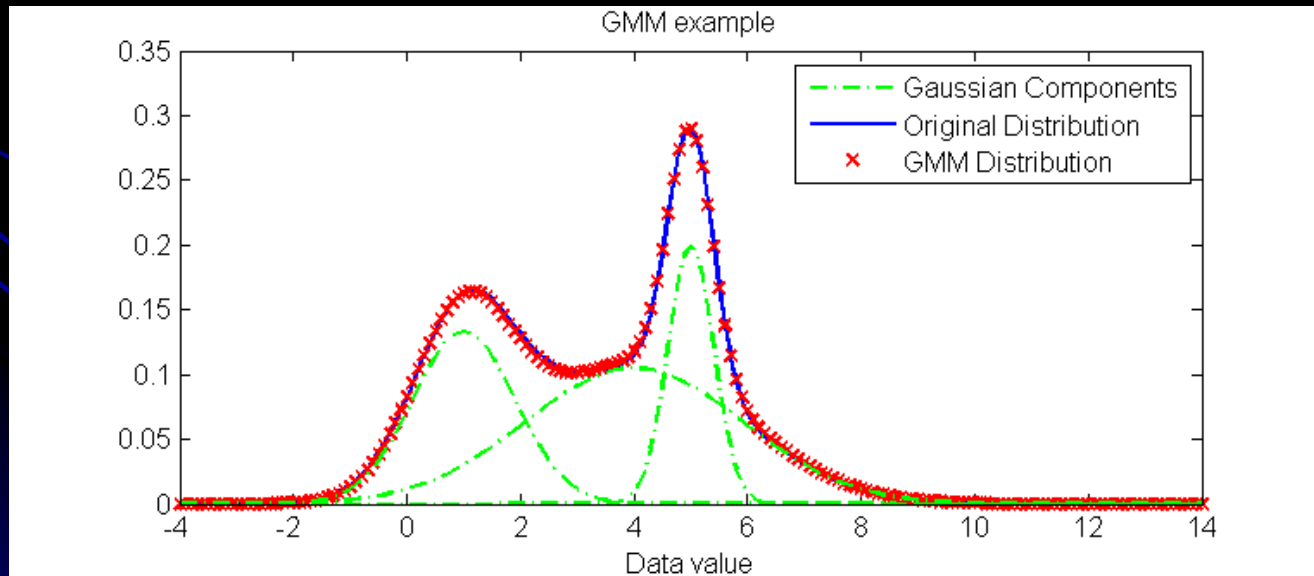
Výberový priemer

Výberová kovariancia

$$P(\mathbf{x} | \omega_i) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} \exp\left(-\frac{1}{2} (\mathbf{x} - \bar{\mathbf{x}}_i)^T \Sigma_i^{-1} (\mathbf{x} - \bar{\mathbf{x}}_i)\right)$$

Viacmodálne rozdelenie – GMM (Gaussian mixture model)

$$P(\mathbf{x} | \omega_i) = \sum_{m=1}^M \frac{w_m}{\sqrt{(2\pi)^d |\Sigma_m|}} \exp\left(-\frac{1}{2} (\mathbf{x} - \bar{\mathbf{x}}_m)^T \Sigma_m^{-1} (\mathbf{x} - \bar{\mathbf{x}}_m)\right)$$



Neparametrické metody

pravděpodobnost, že vybraný vektor bude z oblasti $\mathcal{R}$

$$P = \int_{\mathcal{R}} p(\mathbf{x}) d\mathbf{x}$$

pre dostatočne malé okolie $\mathcal{R}$

$$P = \int_{\mathcal{R}} p(\mathbf{x}) d\mathbf{x} \approx p(\mathbf{x}) \int_{\mathcal{R}} d\mathbf{x} = p(\mathbf{x}) V$$

kde V je obsah $\mathcal{R}$

máme n nezávislých výberov z $p(\mathbf{x})$

$$\mathbf{x}_1, \dots, \mathbf{x}_n$$

ak M z nich patrí do $\mathcal{R}$

$$P = M/n$$

$$\hat{p}(\mathbf{x}) = \frac{M/n}{V}$$

Problém, keď $V \rightarrow 0$

Parzenove okno

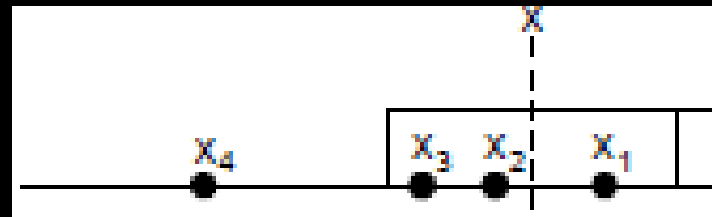
Oblasť $\mathcal{R}$

D -rozmerná jednotková hyperkocka

$$k(\mathbf{u}) = \begin{cases} 1 & |u_j| \leq \frac{1}{2}; \quad j = 1, \dots, D \\ 0 & \text{inak,} \end{cases}$$

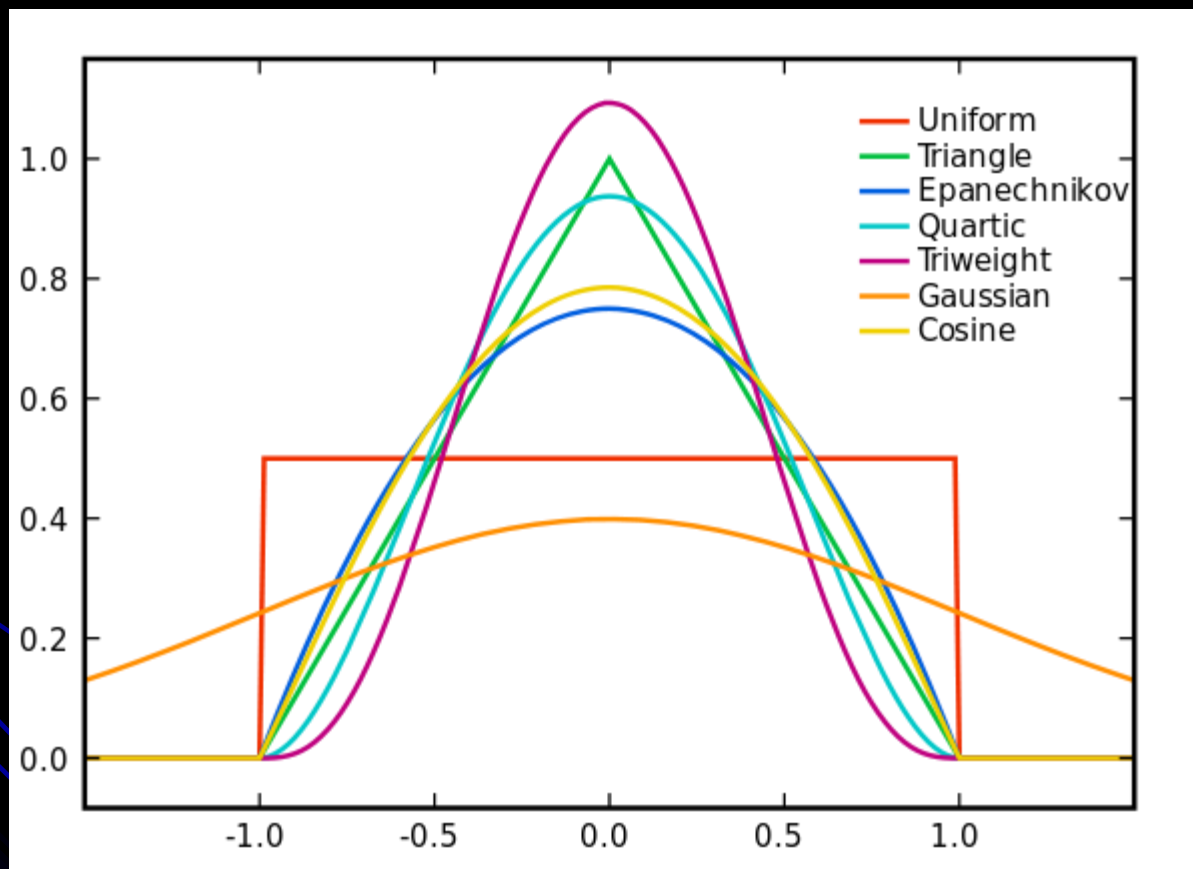
Počet vzoriek v kocke so stranou h a stredom v $\mathbf{x}$:

$$M = \sum_{i=1}^n k\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right)$$



$$\hat{p}(\mathbf{x}) = \frac{M}{nV} = \frac{1}{n} \sum_{i=1}^n \frac{1}{h^D} k\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right)$$

Spojité jádra

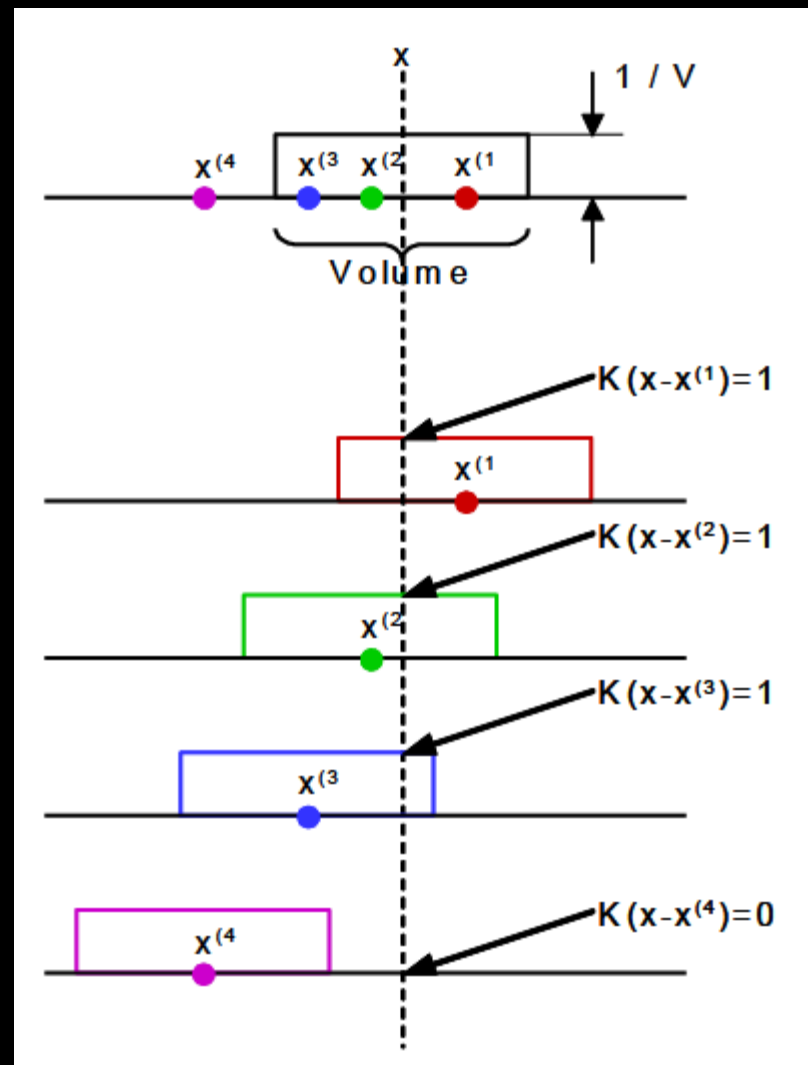


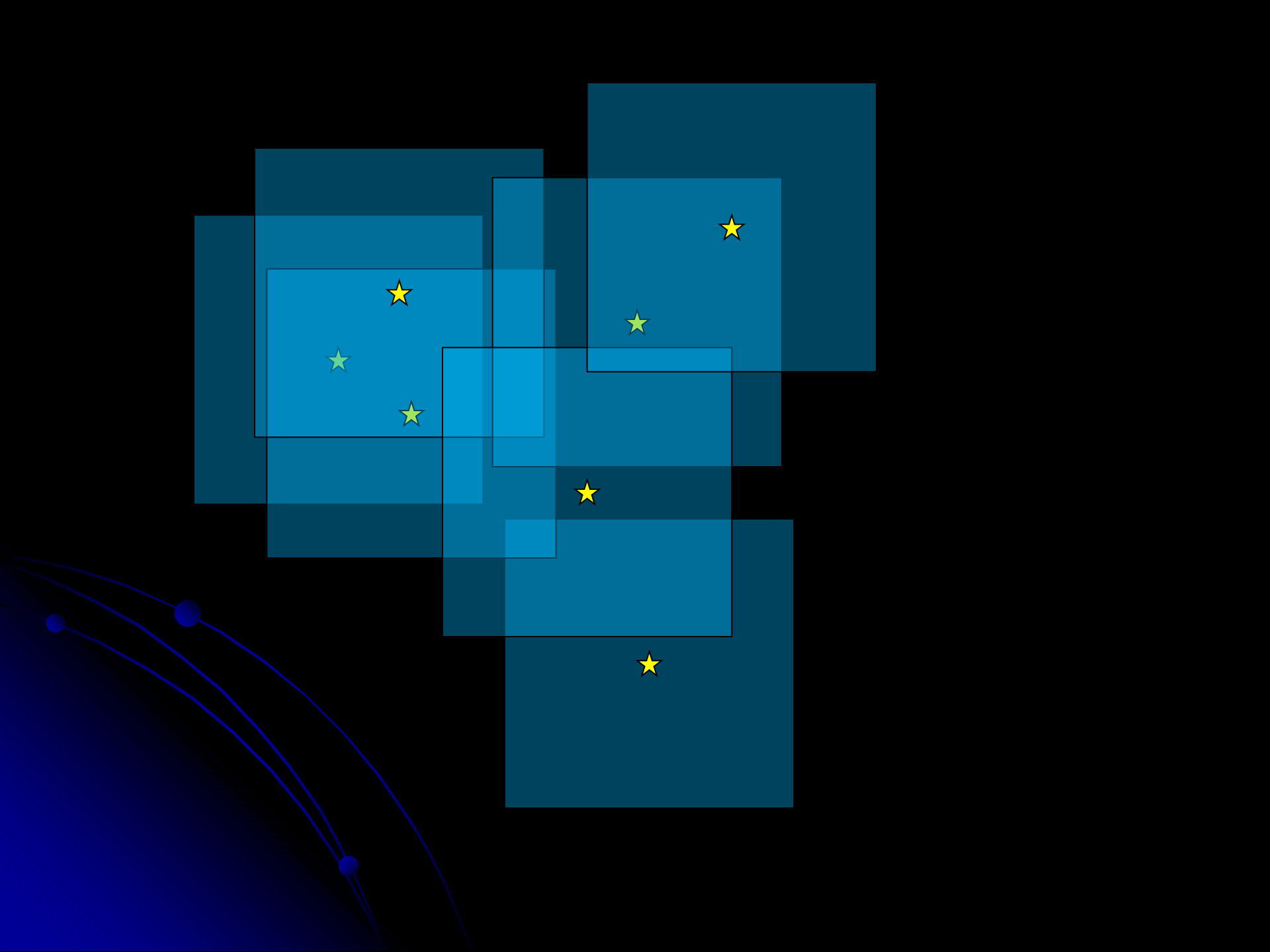
<http://en.wikipedia.org/wiki/File:Kernels.svg>

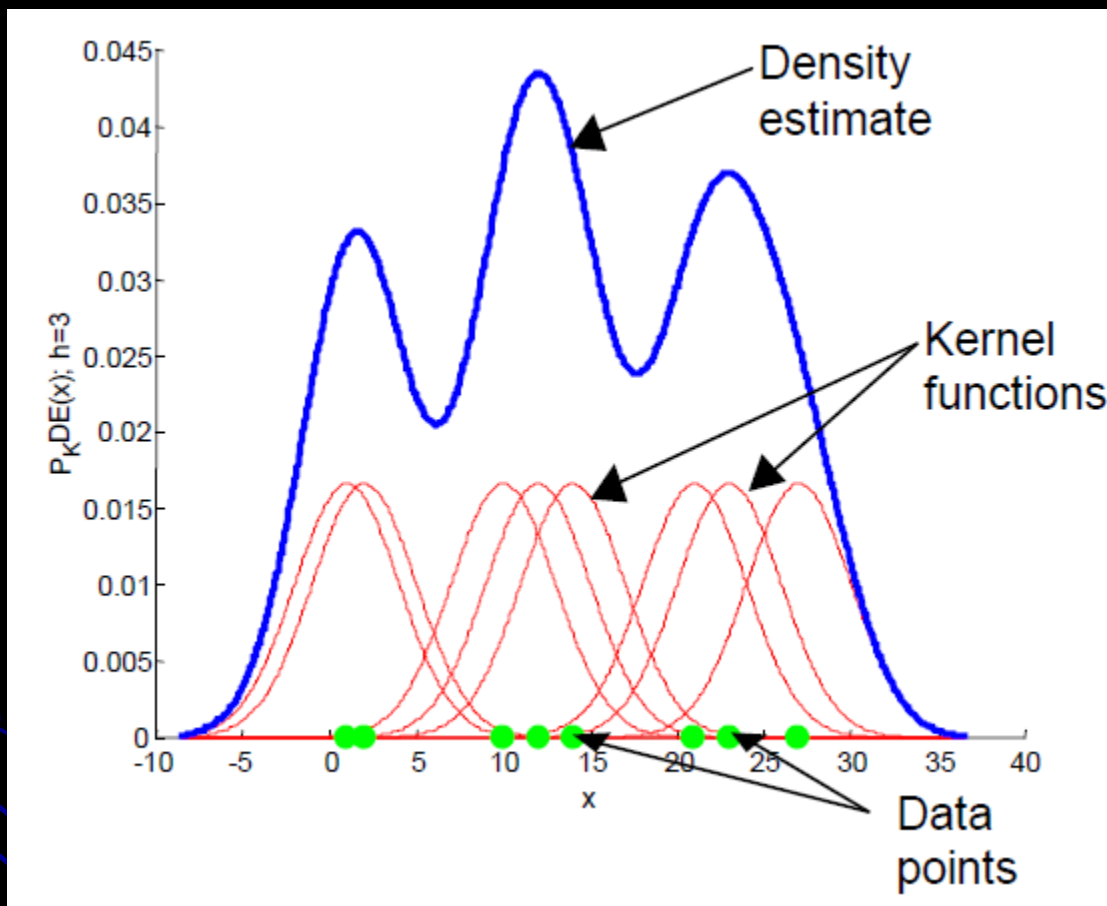
$$p(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h^D} k\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right)$$

Okná okolo vzoriek

Nespojitá funkcia







veľkosť okna ovplyvňuje amplitúdu aj šírku

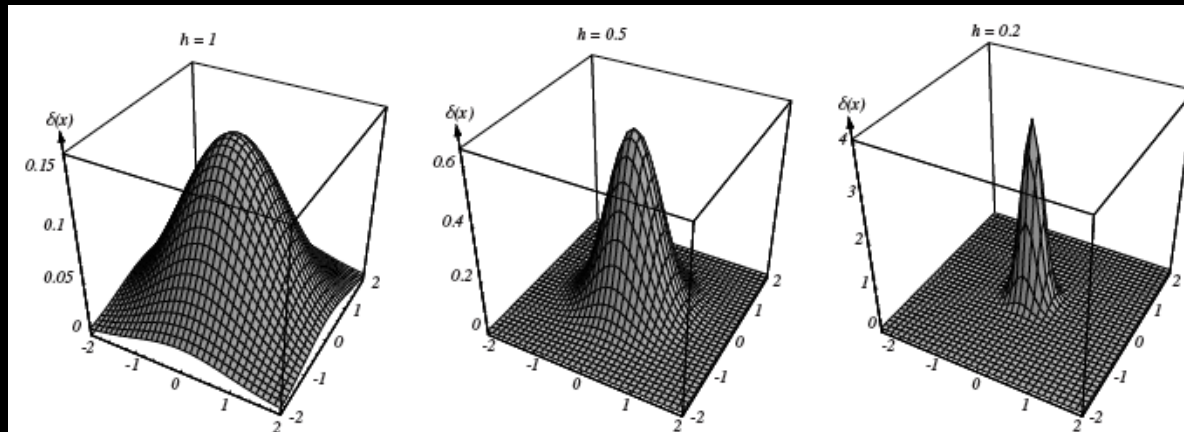


FIGURE 4.3. Examples of two-dimensional circularly symmetric normal Parzen windows for three different values of h . Note that because the $\delta(x)$ are normalized, different vertical scales must be used to show their structure. From: Richard O. Duda, Peter E.

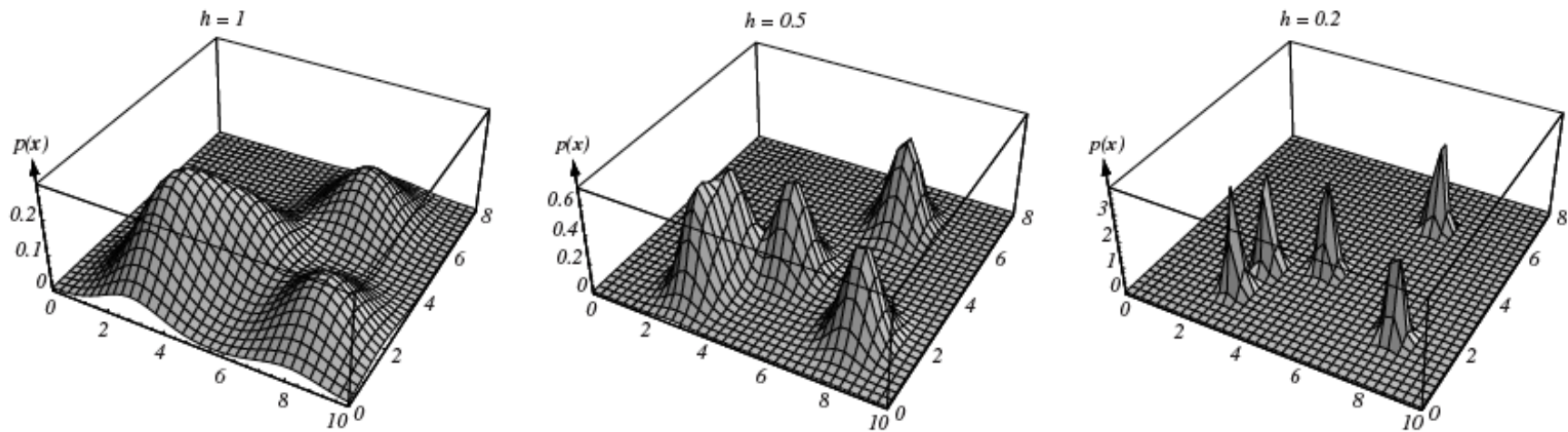
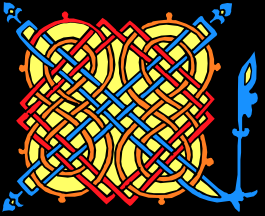


FIGURE 4.4. Three Parzen-window density estimates based on the same set of five samples, using the window functions in Fig. 4.3. As before, the vertical axes have been scaled to show the structure of each distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



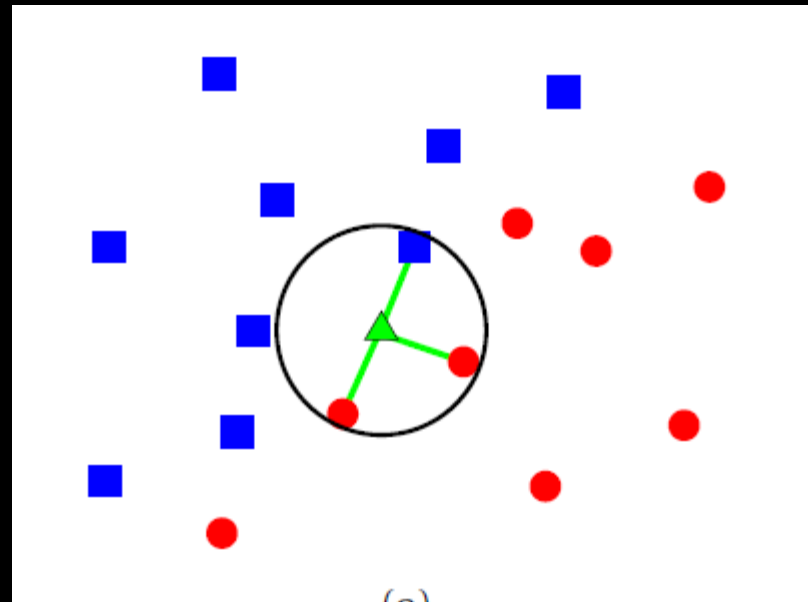
K najbližších susedov

Algoritmus

- pre každý testovaný bod x , nájdi K najbližších susedov z trénovacej množiny
- klasifikuj x podľa väčšiny

$K = 3$

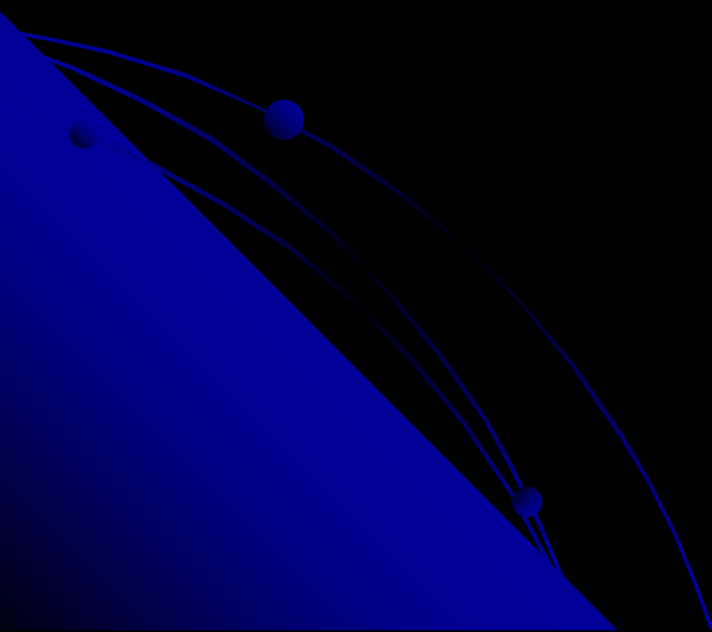
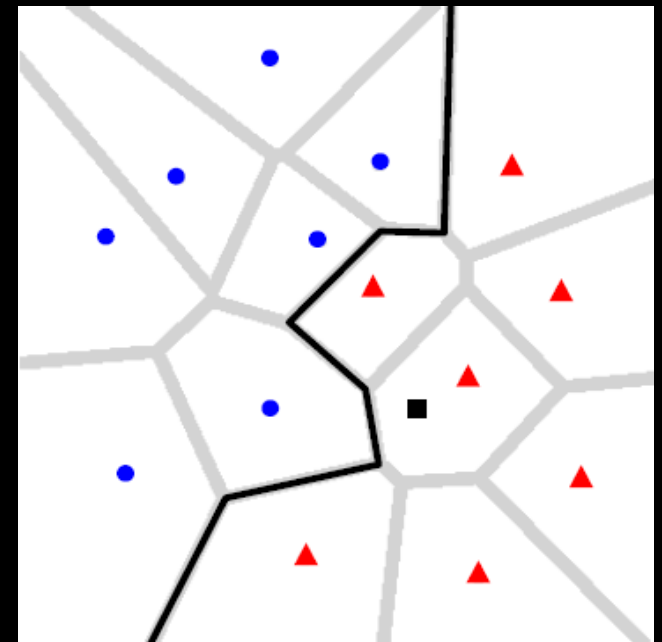
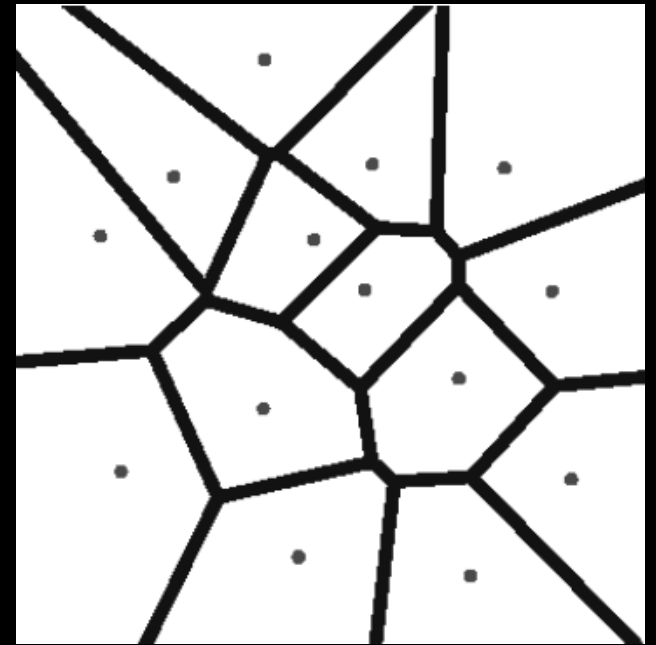
$K=1?$

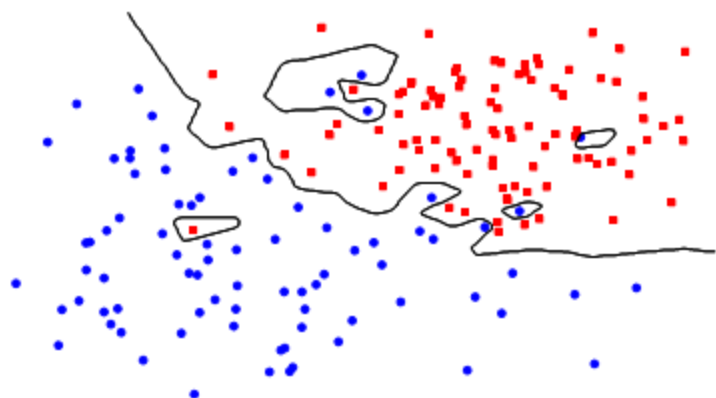


(a)

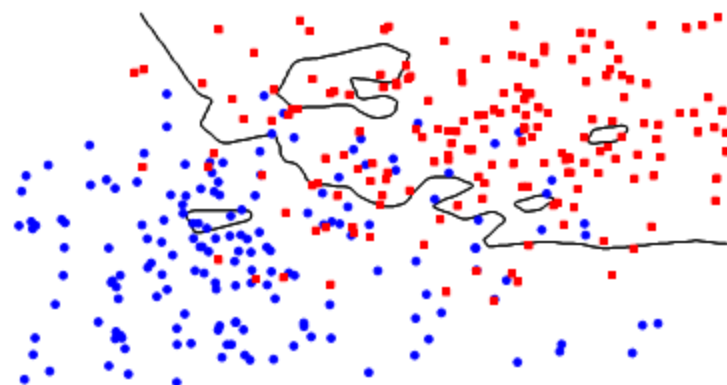
$K=1$

Voronoi diagram

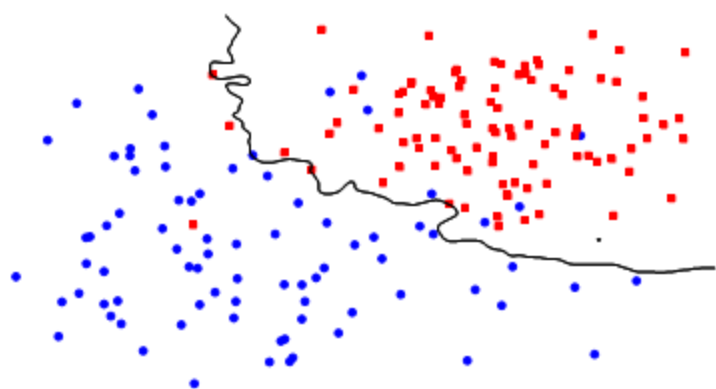




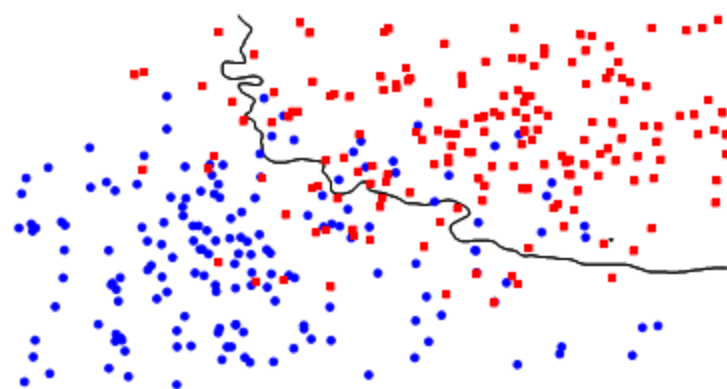
(a) $K = 1$, tréning, dosiahnutá chyba 0



(b) $K = 1$, testovanie, dosiahnutá chyba 0,1667



(c) $K = 7$, tréning, dosiahnutá chyba 0,0565



(d) $K = 7$, testovanie, dosiahnutá chyba 0,1637

zvyšovanie K:

- vyhladzuje sa hranica
- zvyšuje / znižuje sa klasifikačná chyba
- rastie výpočtová náročnosť

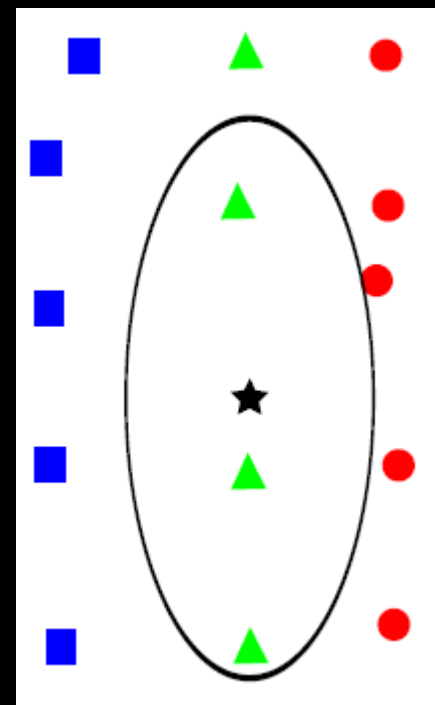
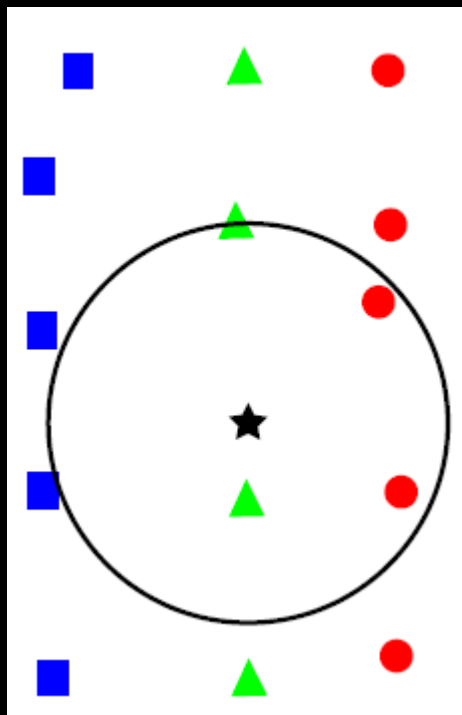
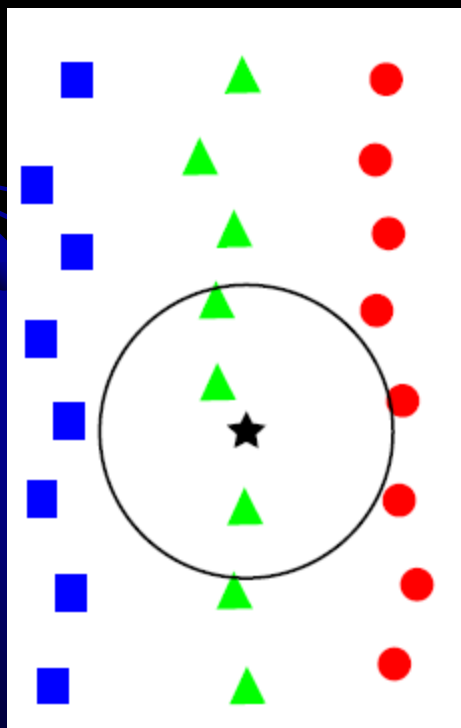
vzájomná validácia na určenie K

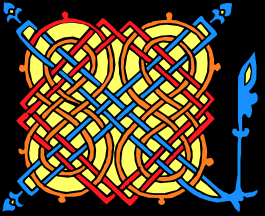
- trénovacia / testovacia množina
- priemer chyby na trénovacích množinách

Čo znamená najbližší? Metrika...

Výpočtová náročnosť: Musíme prehľadať celú tréningovú množinu na určenie susedov.

$K=3$





Rozhodovacie stromy

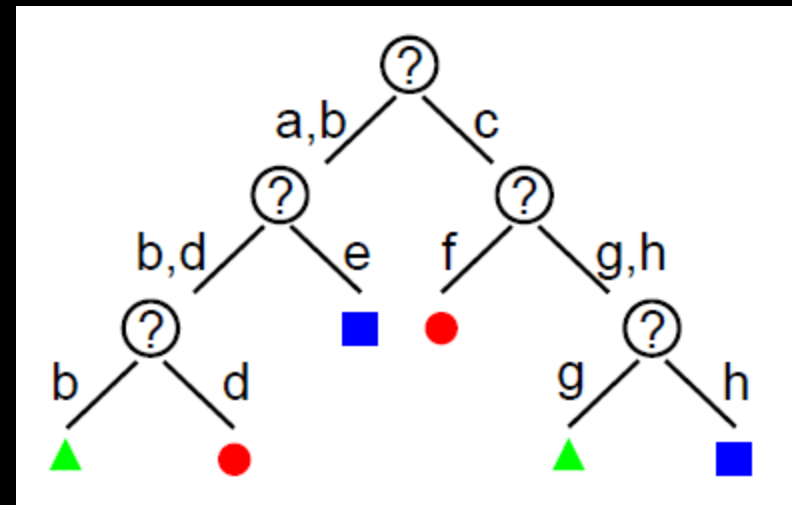
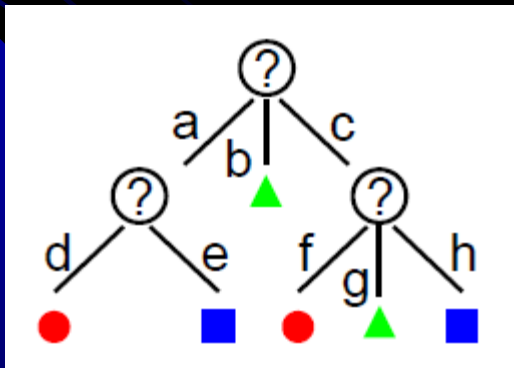
nominálne dáta, pojem vzdialenosti stráca zmysel

pravidlá

Strom:

uzol = test, vetvy zodpovedajú možným výsledkom

list = klasifikačná trieda



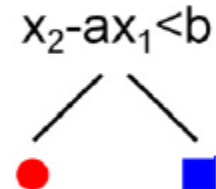
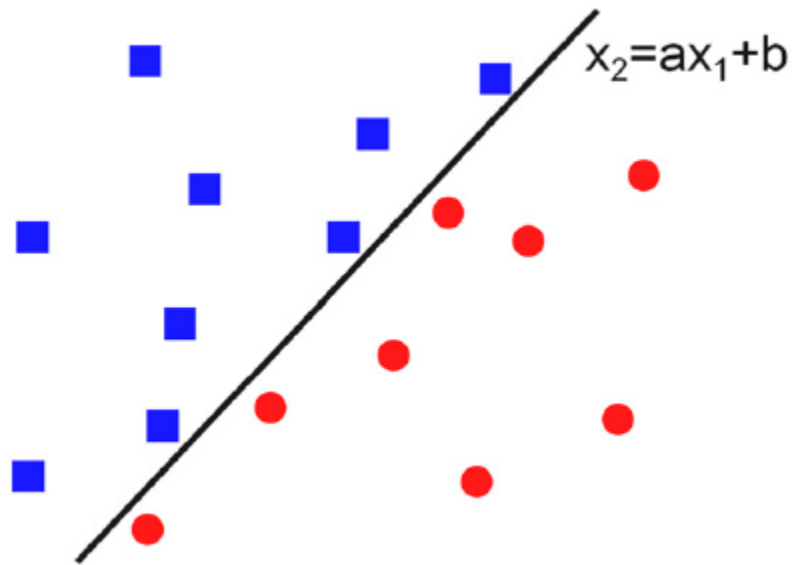
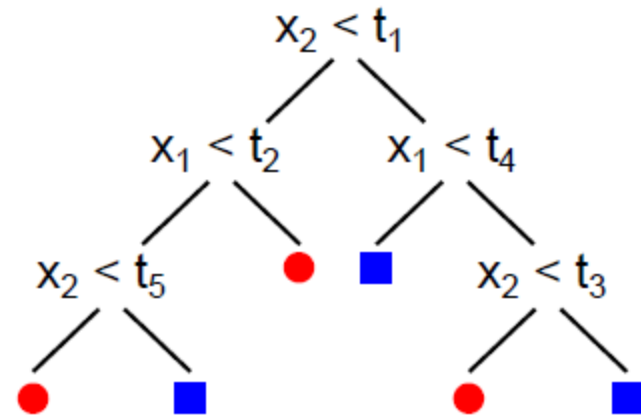
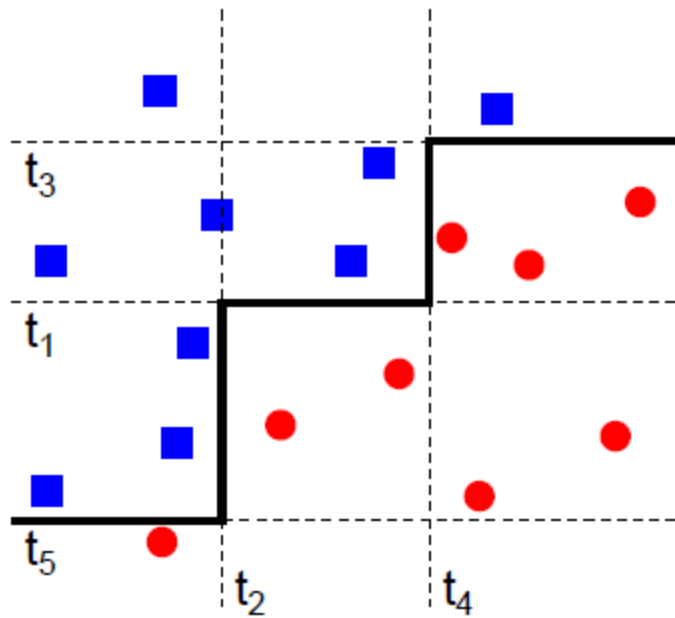
Rozhodovacie porovnanie

v každom uzle

aký je vzťah hodnoty určeného príznaku k zvolenej prahovej hodnote

uzol t skúma dáta podmnožiny X_t trénovacej množiny a rozdeľuje ju na dve podmnožiny X_{tA} , X_{tN}

$$\begin{aligned} X_{tA} \cap X_{tN} &= \emptyset, \\ X_{tA} \cup X_{tN} &= X_t \end{aligned}$$



Rozdeľujúce kritérium

určuje prahovú hodnotu a príznak, ktoré budú vystupovať v rozhodovacom porovnaní

Ukončujúce kritérium

riadi rast stromu

Pravidlá

určujú klasifikačnú triedu v listoch

Rozdeľujúce kritérium

Gini-index – ako často bude náhodný element nesprávne klasifikovaný, ak bude náhodne klasifikovaný podľa rozdelenia tried v množine

$$p_i = \frac{|\{\mathbf{x} \mid \mathbf{x} \in \omega_i\}|}{|U|}$$

Pravdepodobnosť nesprávnej klasifikácie

$$GI = \sum_{i=1}^m p_i(1 - p_i) = 1 - \sum_{i=1}^m p_i^2$$

Pravdepodobnosť výberu

ID3 (Iterative Dichotomiser 3) – minimalizácia entropie

C4.5 – maximalizácia MI

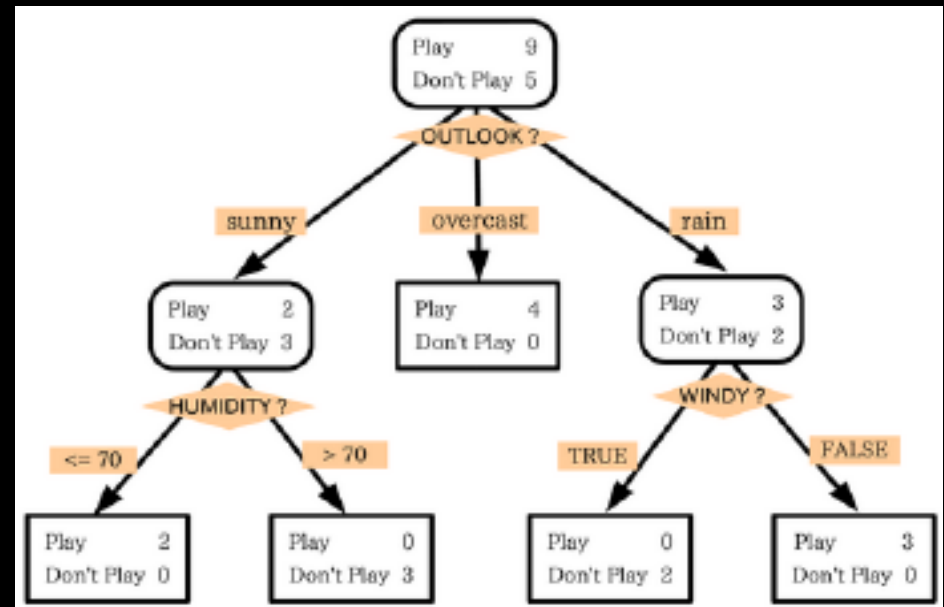
Ukončujúce kritérium

- Všetky objekty v skúmanej množine patria do jednej klasifikačnej triedy.
- Strom dosiahol maximálnu stanovenú hĺbku.
- Počet objektov klasifikovaných v danom uzle je menší ako stanovený prah.
- Ohodnotenie najlepšieho príznaku je nižšie ako stanovený prah.

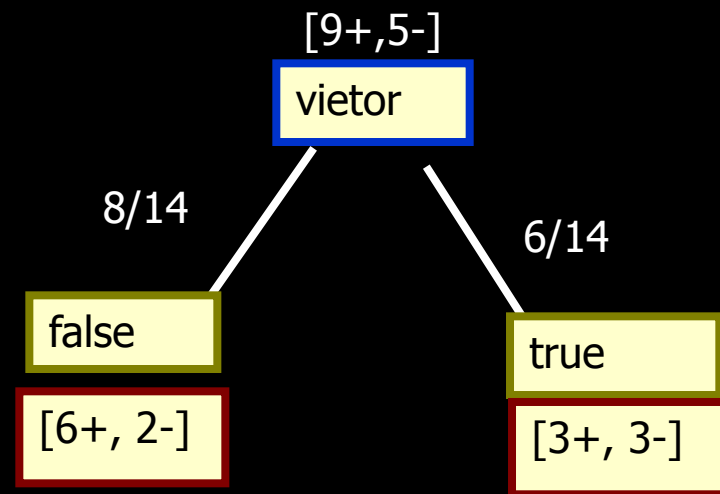
```
PROC RozhStrom (X)
  vytvor strom T s jediným vrcholom V
  IF je splnené ukončovacie kritérium THEN
    označ V ako list a zarad' do určenej triedy
  ELSE
    nájdí príznak P a prahovú hodnotu H pomocou
      rezdeľujúceho kritéria
    vyhodnoť rozhodovacie porovnanie
    urči množiny  $X_{tA}$ ,  $X_{tN}$ 
    TA=RozhStrom( $X_{tA}$ )
    TN=RozhStrom( $X_{tN}$ )
    spoj uzol V s TA hranou označenou A
    spoj uzol V s TN hranou označenou N
  END IF
  RETURN T
```

Príklad C4.5

Independent variables				Dep. var
OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	FALSE	Don't Play
sunny	80	90	TRUE	Don't Play
overcast	83	78	FALSE	Play
rain	70	96	FALSE	Play
rain	68	80	FALSE	Play
rain	65	70	TRUE	Don't Play
overcast	64	65	TRUE	Play
sunny	72	95	FALSE	Don't Play
sunny	69	70	FALSE	Play
rain	75	80	FALSE	Play
sunny	75	70	TRUE	Play
overcast	72	90	TRUE	Play
overcast	81	75	FALSE	Play
rain	71	80	TRUE	Don't Play



Outlook	Temp	Humidity	Windy	Play (positive) / Don't Play (negative)
sunny	69	70	FALSE	Play
overcast	81	75	FALSE	Play
overcast	83	78	FALSE	Play
rain	68	80	FALSE	Play
rain	75	80	FALSE	Play
sunny	85	85	FALSE	Don't Play
sunny	72	95	FALSE	Don't Play
rain	70	96	FALSE	Play
overcast	64	65	TRUE	Play
rain	65	70	TRUE	Don't Play
sunny	75	70	TRUE	Play
rain	71	80	TRUE	Don't Play
overcast	72	90	TRUE	Play
sunny	80	90	TRUE	Don't Play



$I(S; \text{vietor})$:

$$= 0.940 - (8/14) * 0.811 - (6/14) * 1.0$$

$$= 0.048$$

$$I(Y; X) = H(Y) - H(Y|X) = - \sum p_i \cdot \log_2(p_i) - \sum P(X \in \omega) H(Y|X \in \omega)$$

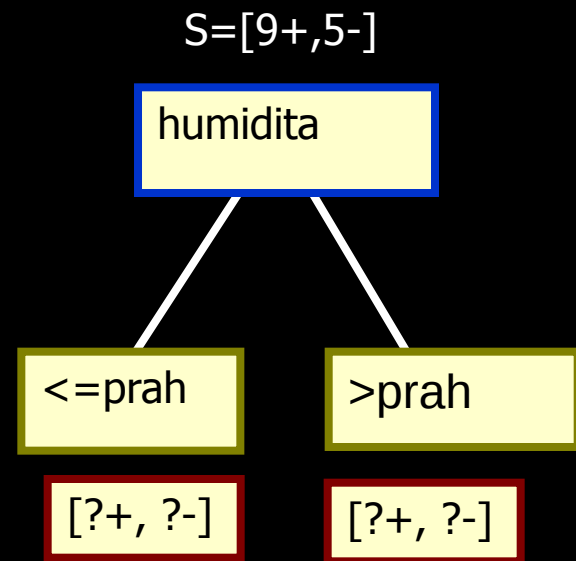
Prah
$$h = \frac{v_i + v_{i+1}}{2}$$

$\{v_1, v_2, \dots, v_j\}$ a $\{v_{j+1}, v_{j+2}, \dots, v_k\}$

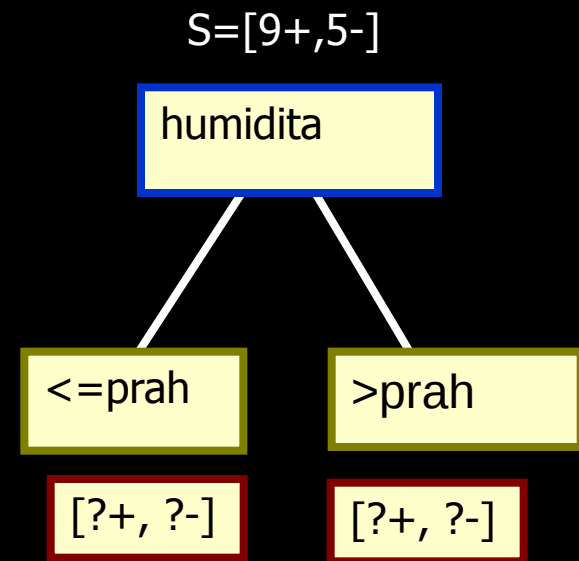
Takýchto možných rozdelení je $k-1$.

hodnotiaca funkcia rozdelenia

rozdelenie s maximálnym ohodnotením.



Outlook	Temp	Humidity	Windy	Play (positive) / Don't Play (negative)
overcast	64	65	TRUE	Play
rain	65	70	TRUE	Don't Play
sunny	69	70	FALSE	Play
sunny	75	70	TRUE	Play
overcast	81	75	FALSE	Play
overcast	83	78	FALSE	Play
rain	68	80	FALSE	Play
rain	71	80	TRUE	Don't Play
rain	75	80	FALSE	Play
sunny	85	85	FALSE	Don't Play
overcast	72	90	TRUE	Play
sunny	80	90	TRUE	Don't Play
sunny	72	95	FALSE	Don't Play
rain	70	96	FALSE	Play



8 možných prahov

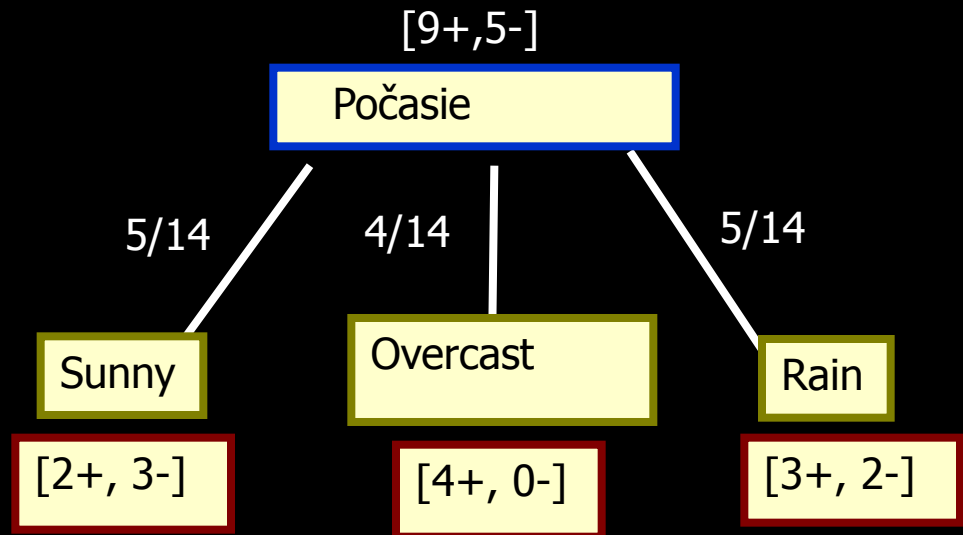
Najlepší: 82,5

Obdobne teplota: prah=70,5

$$I(S, \text{Humidity}) = 0.102$$

$$I(S, \text{Teplota}) = 0.045$$

Outlook	Temp	Humid	Windy	Play (positive) / Don't Play (negative)
overcast	64	65	TRUE	Play
overcast	81	75	FALSE	Play
overcast	83	78	FALSE	Play
overcast	72	90	TRUE	Play
rain	65	70	TRUE	Don't Play
rain	68	80	FALSE	Play
rain	71	80	TRUE	Don't Play
rain	75	80	FALSE	Play
rain	70	96	FALSE	Play
sunny	69	70	FALSE	Play
sunny	75	70	TRUE	Play
sunny	85	85	FALSE	Don't Play
sunny	80	90	TRUE	Don't Play
sunny	72	95	FALSE	Don't Play



$I(S; \text{počasie})$

$$= 0.940 - (5/14) * 0.971 - (4/14) * 0.0 - (5/14) * 0.971$$

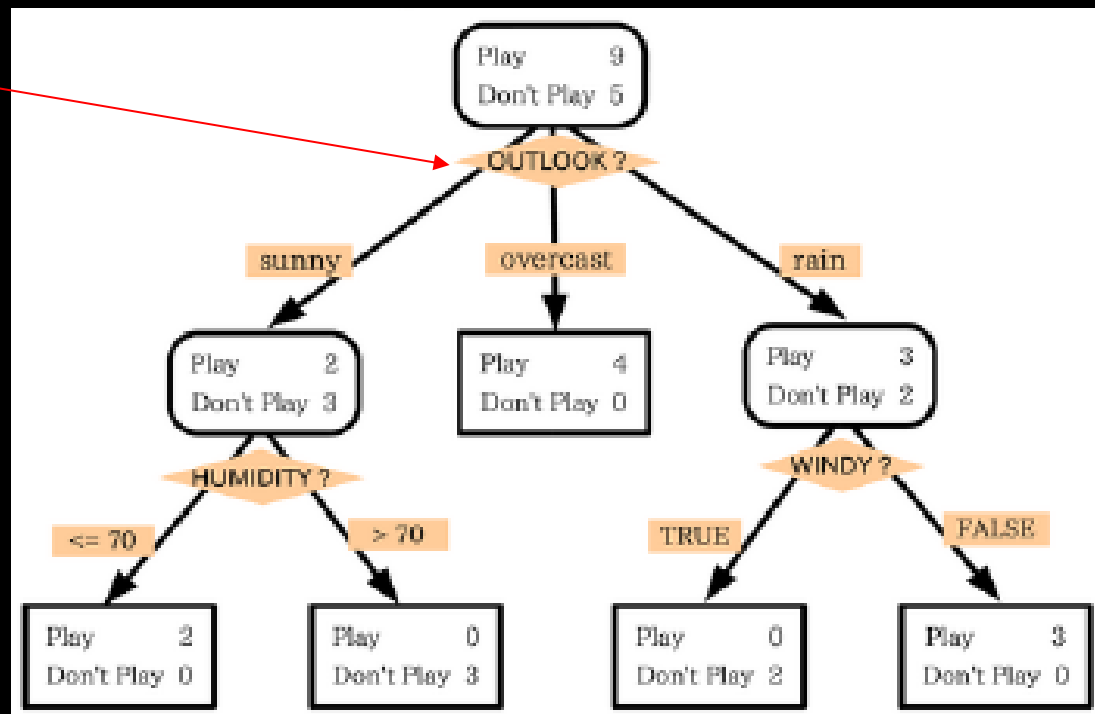
$$= 0.247$$

$I(S; \text{počasie}) = 0.247$

$I(S; \text{vietor}) = 0.048$

$I(S; \text{humidity}) = 0.102$

$I(S; \text{teplota}) = 0.045$



sunny	69	70	FALSE	Play
sunny	75	70	TRUE	Play
sunny	85	85	FALSE	Don't Play
sunny	80	90	TRUE	Don't Play
sunny	72	95	FALSE	Don't Play

rain	65	70	TRUE	Don't Play
rain	68	80	FALSE	Play
rain	71	80	TRUE	Don't Play
rain	75	80	FALSE	Play
rain	70	96	FALSE	Play

Orezávanie rozhodovacích stromov

počas generovania rozhodovacieho stromu

Ak hodnota zvolenej miery významnosti pre skúmaný uzol nepresiahne stanovený prah, ďalšie vetvenie sa zastaví.

po vygenerovaní rozhodovacieho stromu

Toto orezávanie odstraňuje vetvy z už vygenerovaného stromu takým spôsobom, že porovnáva veľkosť očakávanej chyby klasifikácie pre daný podstrom a jeho náhradu listovým uzlom. Ak sa chyba po náhrade listovým uzlom nezväčší, daný podstrom je možné odrezat'.

Výhody

- Jednoduchá interpretácia
- Dáta netreba predspracovať
- Numerické aj kategorické dáta
- Biela skrinka (boolovská logika)
- Nízka výpočtová náročnosť

Nevýhody

- Optimálny strom – NP úplný problém
- Heuristiky nezaručujú optimálny strom
- Možnosť preučenia