

Intervalový odhad

Interval spoľahlivosti

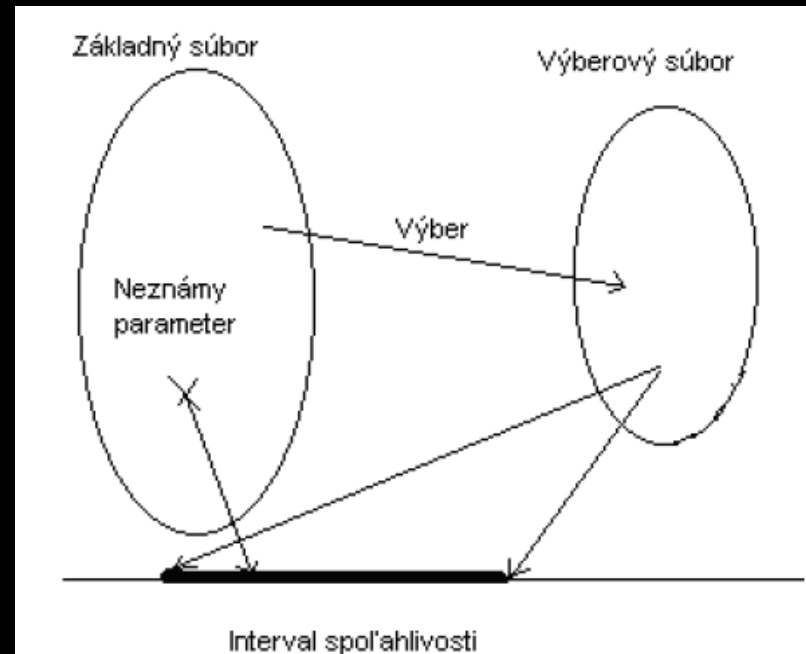
$$P(G_D < \theta < G_H) = 1 - \alpha$$

(G_D, G_H) – interval spoľahlivosti

$1 - \alpha$ – koeficient spoľahlivosti

θ – skutočná hodnota parametra

G_H, G_D –
kritická hodnota rozdelenia



Testovanie hypotéz

Hypotéza = tvrdenie o základnom súbore, ktoré môžeme overiť štatistickými metódami na základe údajov výberového súboru

$$H_0: \theta = \theta_0$$

Každá hypotéza implikuje svoj protiklad alebo alternatívu

$$H_1: \theta = \theta_1$$

$$H_1: \theta > \theta_0$$

$$H_1: \theta < \theta_0$$

$$H_1: \theta \neq \theta_0$$

Čo rieši testovanie hypotéz?

je vhodné zamietnuť testovanú - tzv. *nulovú hypotézu* H_0 - v prospech *alternatívnej hypotézy* H_1 alebo nie?

Na testovanie hypotézy H_0 oproti hypotéze H_1 použijeme vhodné *testovacie kritérium* (tzv. testovaciu štatistiku) g .
Testovacia štatistika g je funkciou náhodného výberu



Testovacie štatistiky

Pre priemer

1. Poznáme σ
2. Nepoznáme σ a $n > 30$
3. Nepoznáme σ a $n < 30$

$$u = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

1

$$u = \frac{\bar{X} - \mu_0}{\frac{s_1}{\sqrt{n}}}$$

2

$$t = \frac{\bar{X} - \mu_0}{\frac{s_1}{\sqrt{n}}}$$

3

Testovacie štatistiky

Pre rozptyl

1. Test zhody s konštantou σ_0
2. Test zhody dvoch rozptylov

$$\chi^2 = \frac{(n-1) \cdot s_1^2}{\sigma_0^2}$$

1

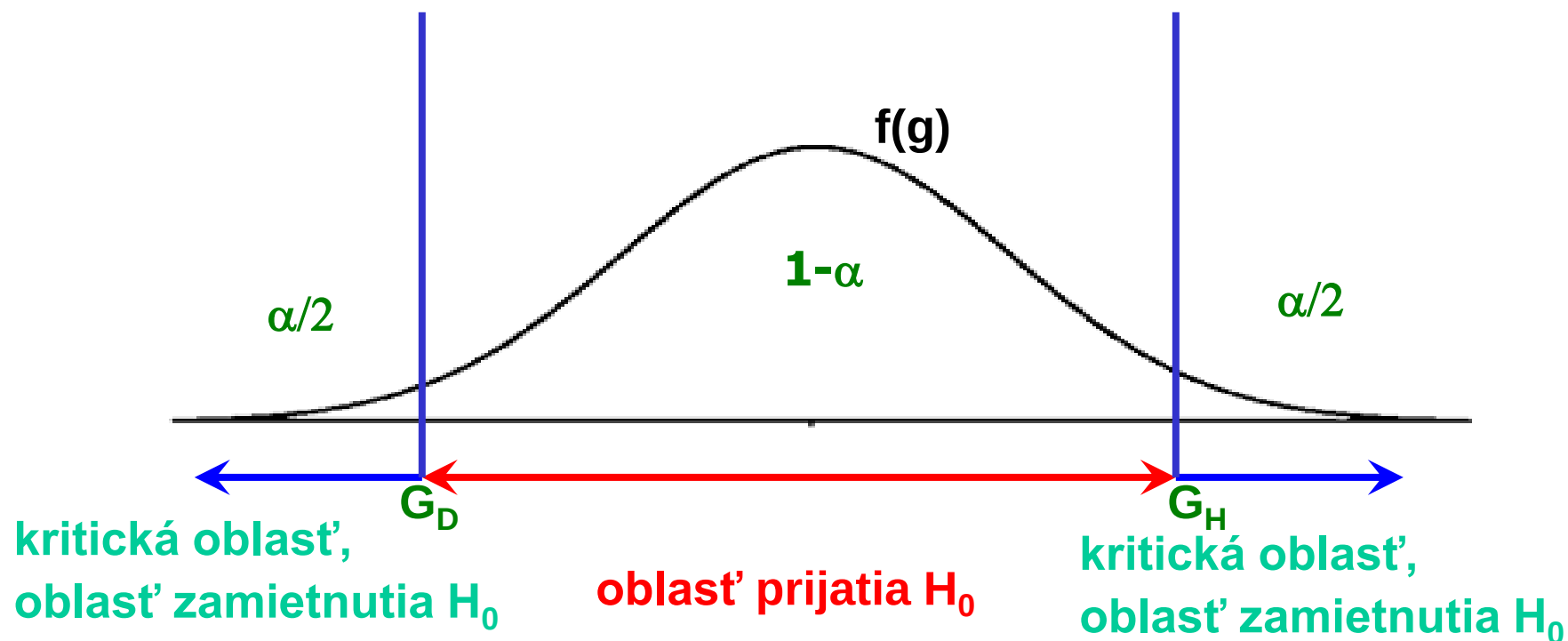
$$F = \frac{s_1^2}{s_2^2} \quad s_1^2 > s_2^2$$

2

Použitie kritickej oblasti

Ak hodnota testovacej štatistiky g padne do kritickej oblasti W_α , zamietame testovanú hypotézu H_0 a za správnu prijímame alternatívnu hypotézu H_1

Body oddeľujúce oblasť prijatia nulovej hypotézy od oblasti jej zamietnutia označujeme názvom *kritické hodnoty*



Postup

1. Formulujeme nulovú hypotézu H_0 a alternatívnu hypotézu H_1 (obojsmernú alebo jednostrannú)
2. Zvolíme testovaciu štatistiku a určíme jej rozdelenie;
3. Vypočítame hodnotu testovacej štatistiky;
4. Zvolíme úroveň (hladinu) významnosti α ;
5. Nájdeme (v štatistických tabuľkách) kritické hodnoty (resp. kvantily) rozdelenia testovacej štatistiky pre zvolenú hodnotu α , ktoré závisia od alternatívnej hypotézy H_1 a vymedzujú kritickú oblasť;
6. Urobíme konečné rozhodnutie t.j. zamietneme alebo nezamietneme H_0 na danej úrovni významnosti α podľa toho, či sa vypočítaná hodnota testovacej štatistiky nachádza alebo nenachádza v kritickej oblasti.

$$U = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n}$$

Možné výsledky

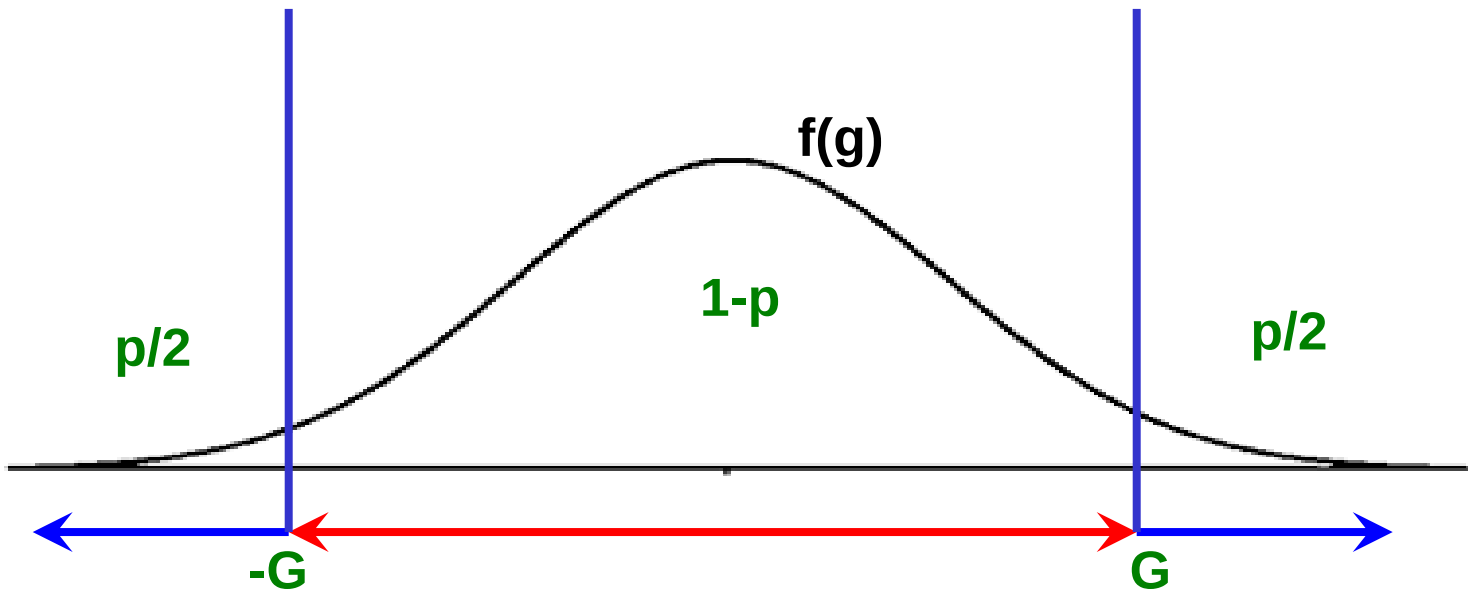
		<i>Skutočná situácia</i>	
		H_0 je pravda	H_0 nie je pravda
<i>Naše rozhodnutie</i>	H_0 je pravda	<i>správne rozhodnutie</i> pravdepodobnosť $1-\alpha$	<i>chyba typu II</i> pravdepodobnosť β
	H_0 nie je pravda	<i>chyba typu I</i> pravdepodobnosť α hladina významnosti	<i>správne rozhodnutie</i> pravdepodobnosť $1 - \beta$ sila testu

$$\alpha = P(g \in W_\alpha \mid H_0)$$

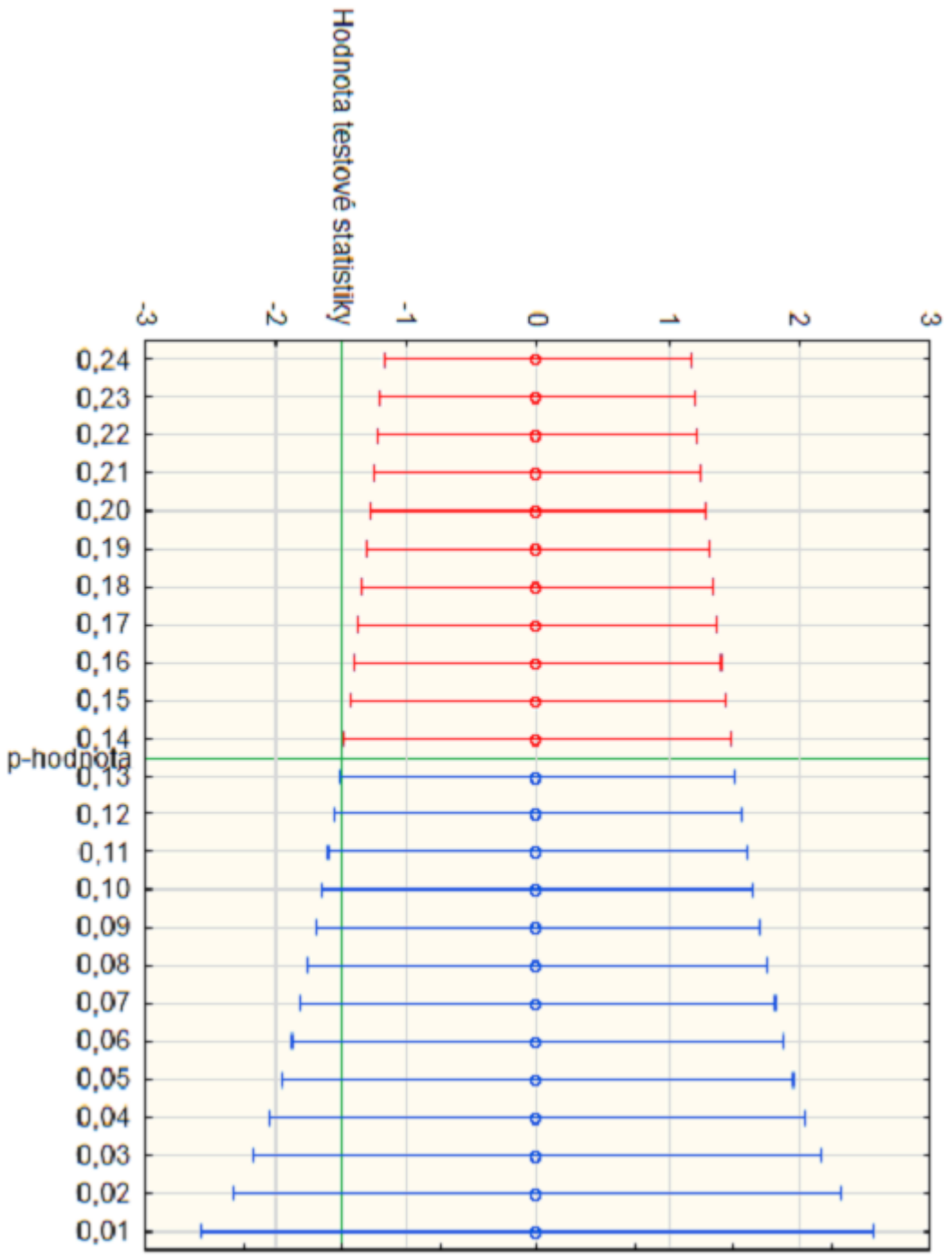
$$\beta = P(g \notin W_\alpha \mid H_1)$$

Použitie p-hodnoty

$$p = 2 \min\{P(g \leq G), P(g \geq G)\}$$



Ak je p-hodnota $\leq \alpha$, potom H_0 zamietame na hladine významnosti α



príklad

náhodný výber 100 výrobkov

priemerná váha 2,2 kg

smerodajná odchýlka váhy výrobkov $\sigma = 0,6$ kg

$H_0: \mu=2$

$H_1: \mu \neq 2$

Hladina významnosti: $\alpha=0,05$

$$U = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n} = \frac{2,2 - 2}{0,6} \sqrt{100} = \frac{10}{3} = 3,33 > 1,96$$

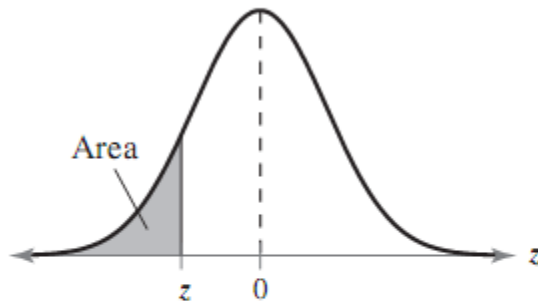
Zamietame H_0 na hladine významnosti 0,05

$X \sim N(0,1), P(|X| > u_\alpha) = \alpha$

α	0,01	0,02	0,05	0,1	0,2
u_α	2,5758	2,3263	1,9599	1,6448	1,299

$$U = 3,33$$

Table 4—Standard Normal Distribution



<i>z</i>	.09	.08	.07	.06	.05	.04	.03	.02	.01	.00
-3.4	.0002	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003
-3.3	.0003	.0004	.0004	.0004	.0004	.0004	.0004	.0005	.0005	.0005
-3.2	.0005	.0005	.0005	.0006	.0006	.0006	.0006	.0006	.0007	.0007
-3.1	.0007	.0007	.0008	.0008	.0008	.0008	.0009	.0009	.0009	.0010
-3.0	.0010	.0010	.0011	.0011	.0011	.0012	.0012	.0013	.0013	.0013
-2.9	.0014	.0014	.0015	.0015	.0016	.0016	.0017	.0017	.0018	.0019

$$p=0,0008 < 0,05$$

Zamietame H_0 na hladine významnosti 0,05

<http://mi21.vsb.cz/modul/vybrane-kapitoly-z-pravdepodobnosti>

Štatistické metódy a ich aplikácie

Dagmar Markechová, Beáta Stehlíková, Anna Tirpáková

Introduction to Probability Theory and Statistics

Javier R. Movellan

Pravdepodobnost a matematicka statistika

Mirko Navara

Vlastné čísla, vlastné vektory

Číslo λ nazývame vlastným číslom matice A , ak sústava lineárnych rovníc tvaru

$$A\mathbf{x} = \lambda\mathbf{x}$$

má okrem triviálneho riešenia aj nenulové riešenie. Príslušné nenulové riešenie sa nazýva sa vlastný vektor matice A odpovedajúci vlastnému číslu λ

- Vlastné čísla nájdeme vyriešením

$$\det(A - \lambda I) = 0$$

Potom pre každú λ hodnotu vyriešime

$$(A - \lambda I)\mathbf{x} = 0$$

príklad

$$\mathbf{A} = \begin{bmatrix} 2 & -4 \\ -1 & -1 \end{bmatrix}$$

$$\begin{aligned} p(\lambda) &= \begin{vmatrix} 2-\lambda & -4 \\ -1 & -1-\lambda \end{vmatrix} = \\ &= (2-\lambda)(-1-\lambda) - (-4)(-1) = \\ &= \lambda^2 - \lambda - 6 = \\ &= (\lambda - 3)(\lambda + 2) \end{aligned}$$

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0}$$

$$\begin{bmatrix} 2-3 & -4 \\ -1 & -1-3 \end{bmatrix} \begin{bmatrix} v_1^1 \\ v_2^1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$-v_1^1 - 4v_2^1 = 0$$

$$-v_1^1 - 4v_2^1 = 0$$

$$v_1^1 = -4v_2^1$$

$$\begin{bmatrix} 2-(-2) & -4 \\ -1 & -1-(-2) \end{bmatrix} \begin{bmatrix} v_1^2 \\ v_2^2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$4v_1^2 - 4v_2^2 = 0$$

$$-v_1^2 + v_2^2 = 0$$

$$v_1^2 = v_2^2$$

$$\mathbf{v}_1 = \begin{bmatrix} 4 \\ -1 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

príklad

$$A = \begin{bmatrix} -1 & 2 & 0 \\ 0 & 3 & 4 \\ 0 & 0 & 7 \end{bmatrix}$$

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0$$

$$\det\left(\begin{bmatrix} -1 & 2 & 0 \\ 0 & 3 & 4 \\ 0 & 0 & 7 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right) = 0$$

$$\det\left(\begin{bmatrix} -1-\lambda & 2 & 0 \\ 0 & 3-\lambda & 4 \\ 0 & 0 & 7-\lambda \end{bmatrix}\right) = 0$$

$$(-1-\lambda)(3-\lambda)(7-\lambda) = 0$$

$$\lambda = -1, \quad \lambda = 3, \quad \lambda = 7$$

príklad

$$\lambda = -1$$

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{v} = \mathbf{0}$$

$$\left(\begin{bmatrix} -1 & 2 & 0 \\ 0 & 3 & 4 \\ 0 & 0 & 7 \end{bmatrix} - \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \right) \begin{bmatrix} v_1^1 \\ v_2^1 \\ v_3^1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 2 & 0 \\ 0 & 4 & 4 \\ 0 & 0 & 8 \end{bmatrix} \begin{bmatrix} v_1^1 \\ v_2^1 \\ v_3^1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$0 + 2v_2^1 + 0 = 0$$

$$0 + 4v_2^1 + 4v_3^1 = 0$$

$$0 + 0 + 8v_3^1 = 0$$

$$v_1^1 = 1, \quad v_2^1 = 0, \quad v_3^1 = 0$$

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} 1 \\ 4 \\ 4 \end{bmatrix}$$

Singulárne číslo a singulárne vektory

$$\begin{aligned} \mathbf{A}\mathbf{v} &= \sigma\mathbf{u}, \\ \mathbf{A}^H\mathbf{u} &= \sigma\mathbf{v}, \end{aligned}$$

$\mathbf{v}$ je vlastný vektor matice $\mathbf{A}^T\mathbf{A}$

$\mathbf{u}$ je vlastný vektor matice $\mathbf{A}\mathbf{A}^T$

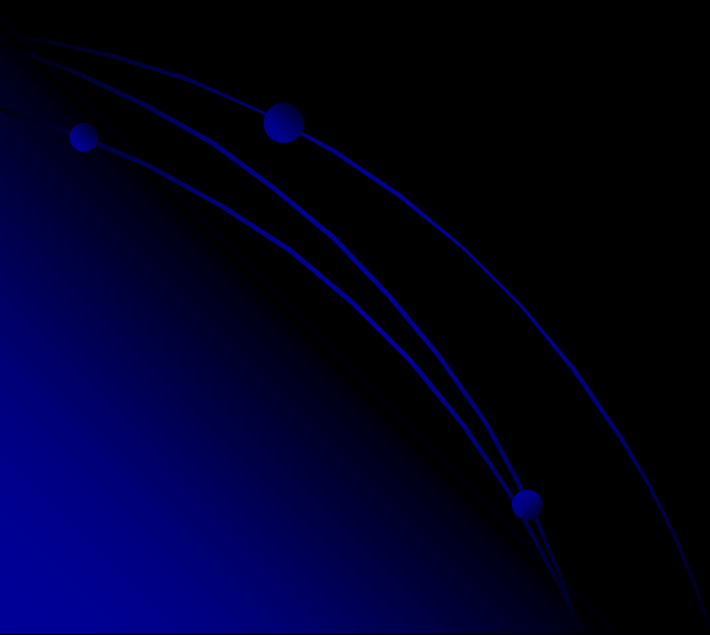
prislúchajúce vlastnému číslu σ^2

$$\begin{aligned} \mathbf{A}\mathbf{V} &= \mathbf{U}\mathbf{S}, \\ \mathbf{A}^T\mathbf{U} &= \mathbf{V}\mathbf{S}^T \end{aligned}$$

Singulárny rozklad matice

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

$$[\mathbf{A}]_{N \times D} = [\mathbf{U}]_{N \times N} [\mathbf{S}]_{N \times D} [\mathbf{V}^T]_{D \times D}$$

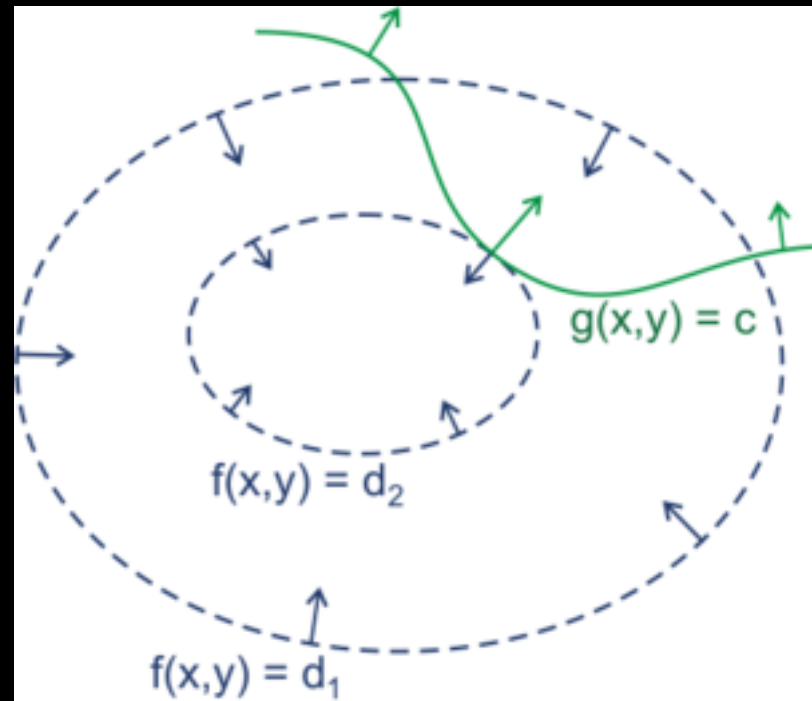


Nelineárne programovanie

$$\begin{array}{ll} \min & f(\mathbf{x}) \\ \text{aby} & g_i(\mathbf{x}) \leq c_i \quad i = 1, \dots, n \end{array}$$

$$\begin{array}{ll} \min & f(\mathbf{x}) \\ \text{aby} & g_i(\mathbf{x}) - c_i \leq 0 \quad i = 1, \dots, n \end{array}$$

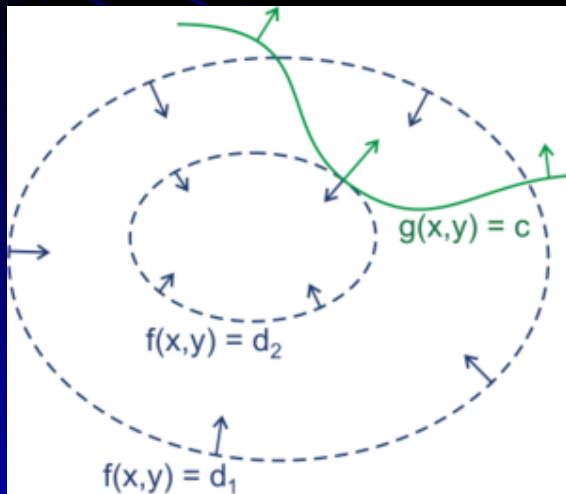
Lagrangeova metóda
neurčitých koeficientov



Lagrangeova funkcia

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{i=1}^n \lambda_i g_i(\mathbf{x})$$

Kuhn-Tuckerove podmienky



$$\frac{\partial L(\mathbf{x}, \boldsymbol{\lambda})}{\partial x_j} = \frac{\partial f(\mathbf{x})}{\partial x_j} + \sum_{i=1}^n \lambda_i \frac{\partial g_i(\mathbf{x})}{\partial x_j} = 0$$

$$\frac{\partial L(\mathbf{x}, \boldsymbol{\lambda})}{\partial \lambda_i} = g_i(\mathbf{x}) \leq 0$$

$$\lambda_i \frac{\partial L(\mathbf{x}, \boldsymbol{\lambda})}{\partial \lambda_i} = \lambda_i g_i(\mathbf{x}) = 0$$

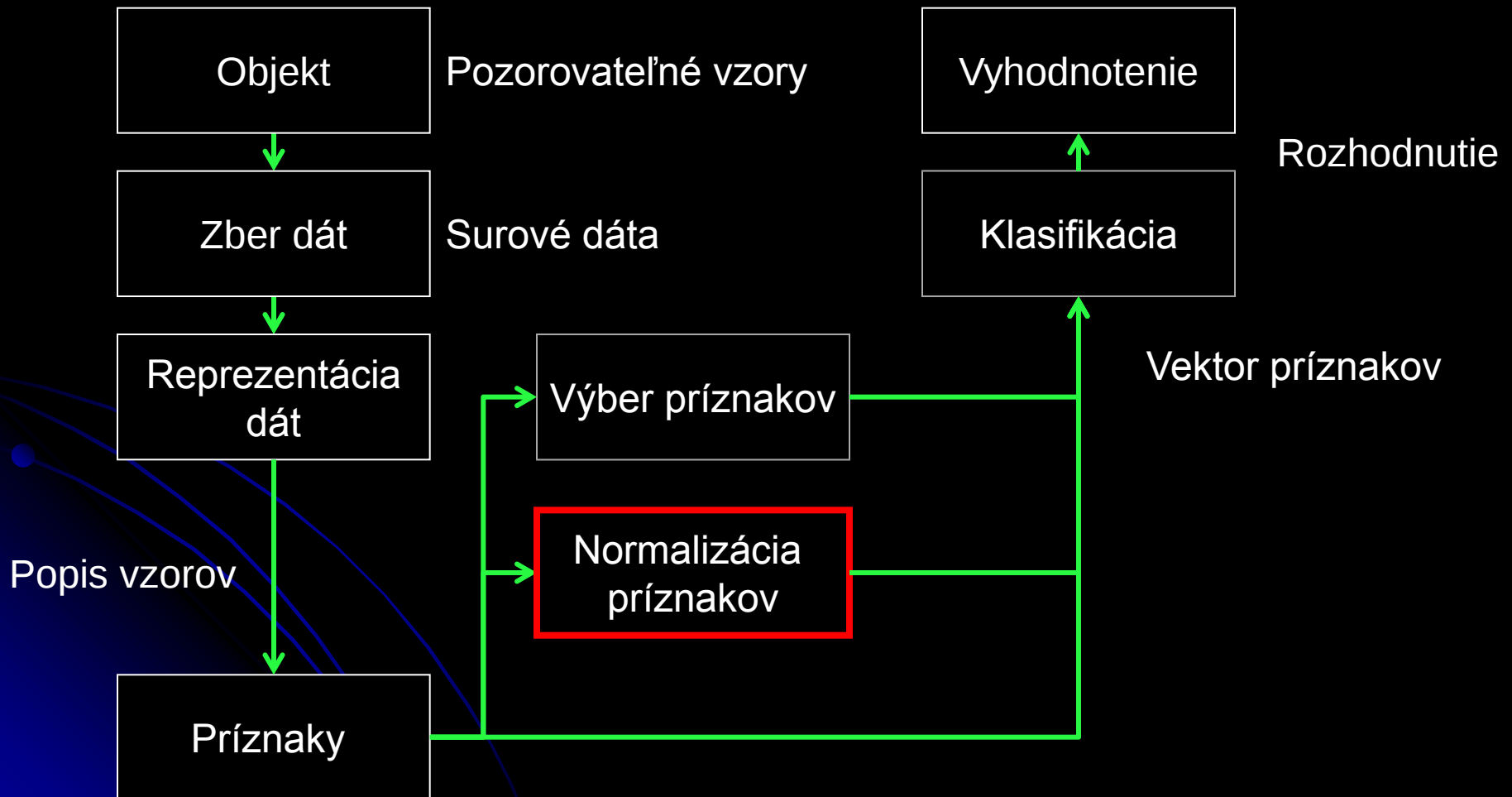
$$\lambda_i \geq 0$$

Koniec odbočky



ale nie koniec matiky

Rozpoznávanie



Príznaky bez rozmeru

Príznaky relatívne k určitej referenčnej hodnote

Príklad:

Skutočná výška hráčov v tíme (cm, m)

Bezrozmerný príznak výšky:

Vezmeme referenčnú hodnotu (výška najvyššieho, najnižšieho hráča, 1m, 100cm, ...)

$$(\text{Bezrozmerný príznak výšky}) = \frac{(\text{skutočná výška})}{(\text{referenčná hodnota})}$$

Normalizácia

Lineárne škálovanie na jednotkový interval

$$\tilde{x}_i = \frac{x_i - l}{u - l}$$

u – maximálna hodnota

l – minimálna hodnota

Lineárne škálovanie na jednotkovú dĺžku vektora

$$\tilde{x}_i = \frac{x_i}{\|\mathbf{x}\|}$$

Normalizácia

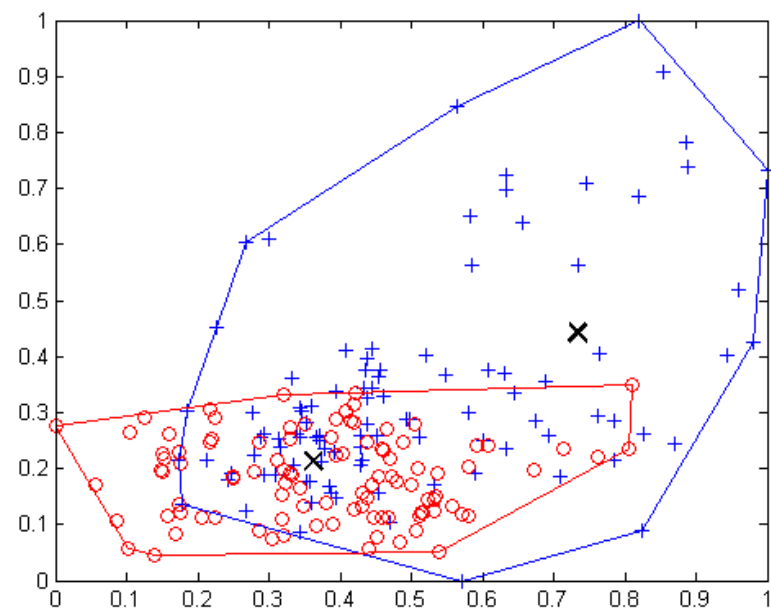
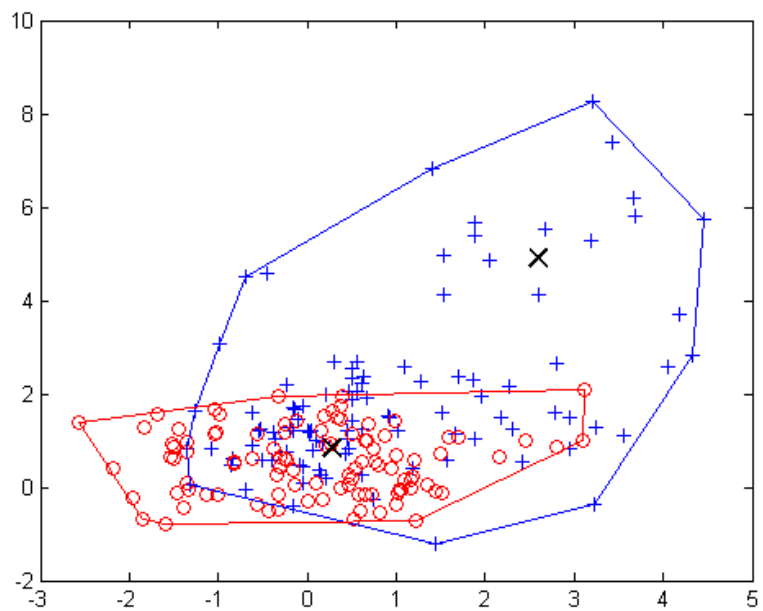
Lineárne škálovanie na nulový priemer
a jednotkovú varianciu

$$\tilde{x}_i = \frac{x_i - \bar{x}}{s}$$

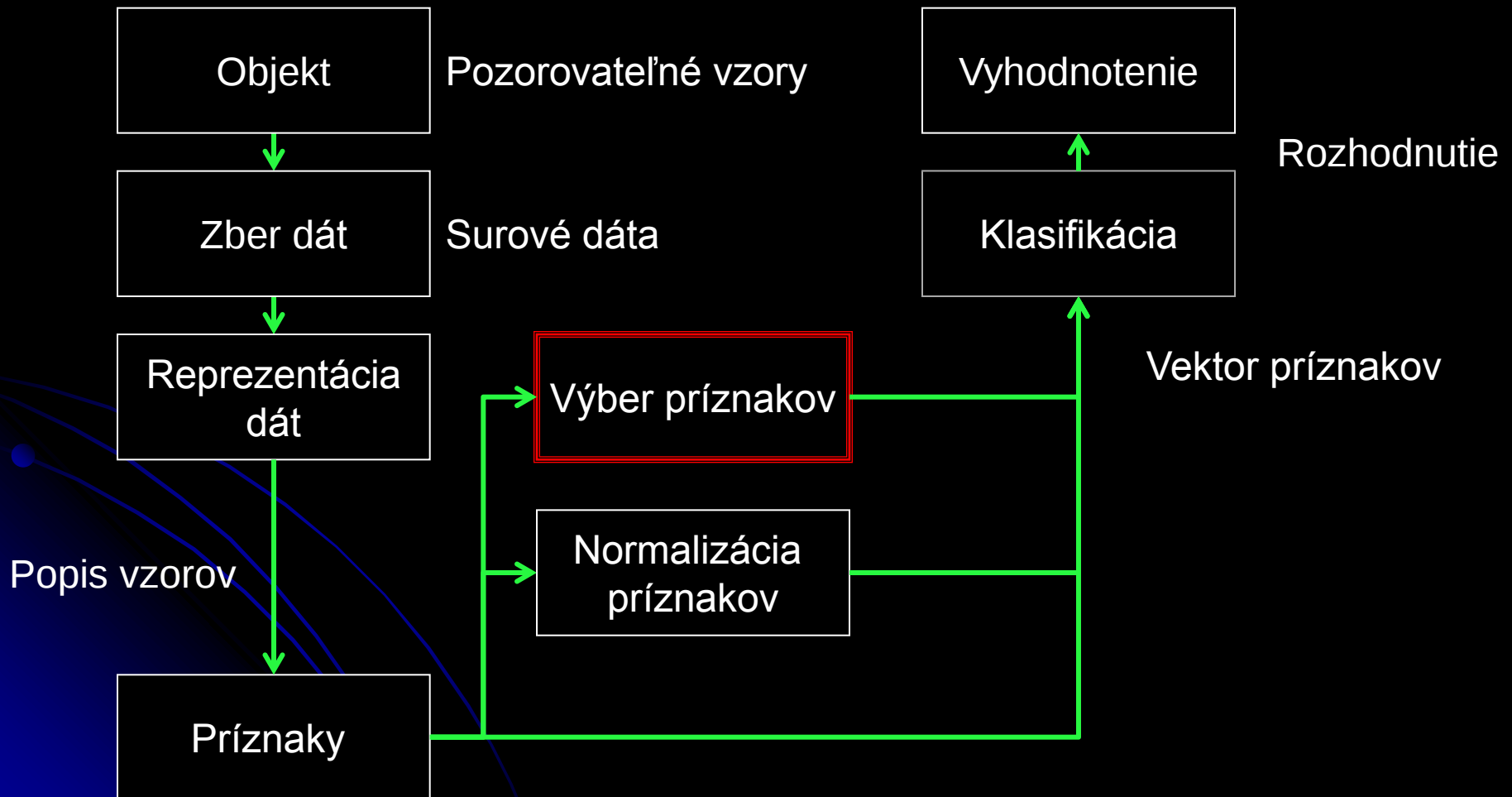
Normálne rozdelenie

$$\tilde{x}_i = \frac{\frac{x_i - \bar{x}}{3s} + 1}{2}$$

99% dát v [0,1]



Rozpoznávanie



Zníženie počtu príznakov

F_1	F_2	F_3	F_4	F_5	C
0	0	1	0	1	0
0	1	0	0	1	1
1	0	1	0	1	1
1	1	0	0	1	1
0	0	1	1	0	0
0	1	0	1	0	1
1	0	1	1	0	1
1	1	0	1	0	1

2 prístupy

Výber príznakov:

vyberieme podmnožinu z originálnych príznakov

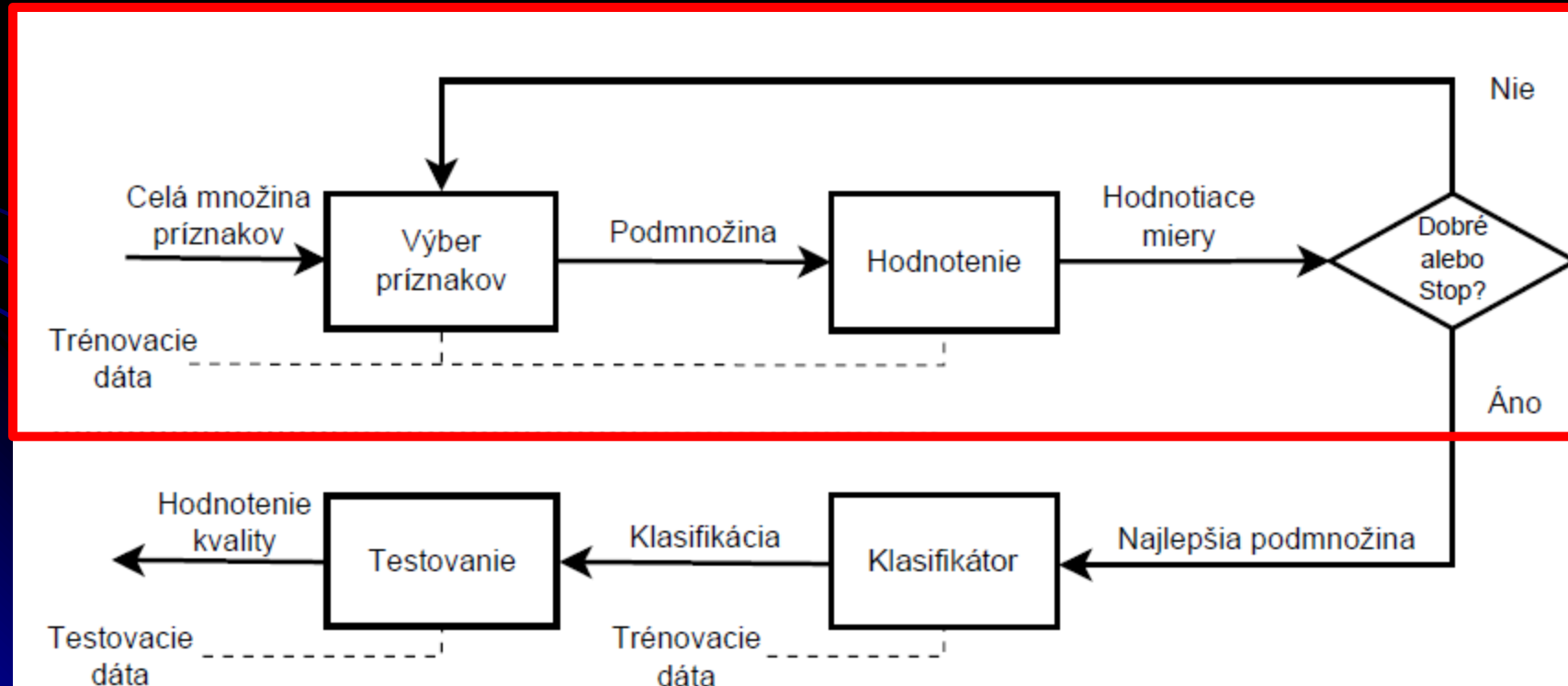
Redukcia príznakov:

transformujeme pôvodnú množinu do menej-dimenzionálnej

Výber príznačkov

Filter

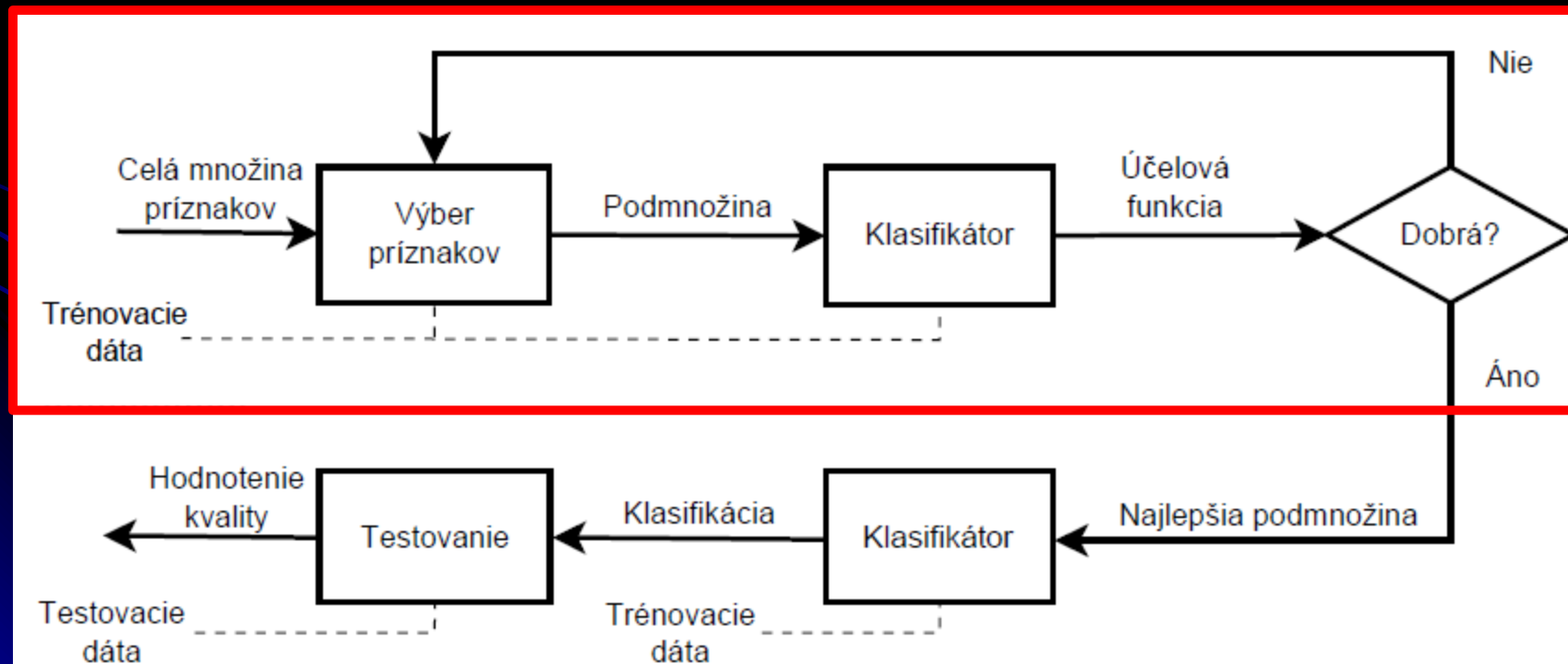
oddelenie procesu výberu od klasifikátora
charakteristika dát (*informácia, vzdialenosť, závislosť, konzistencia*)



Výber príznačkov

Obálka

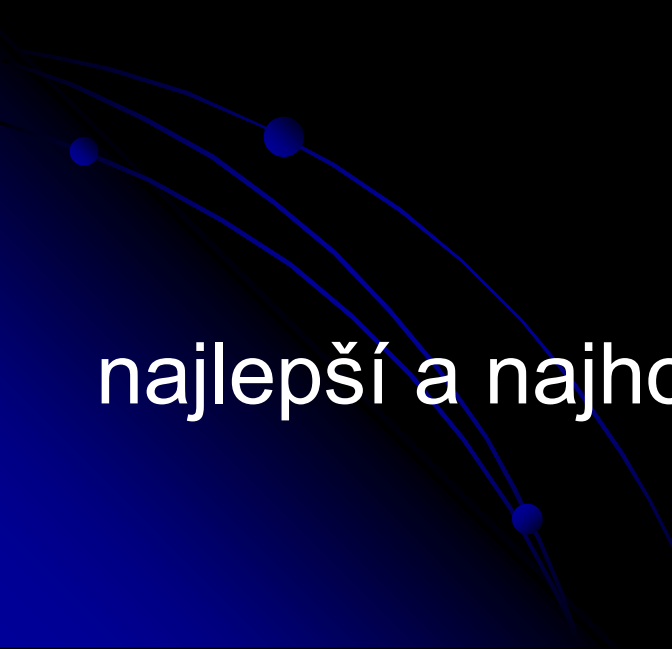
závisí od určeného klasifikátora
očakávaná presnosť klasifikátora
presné, výpočtovo náročné



Výber vhodných príznakov

Dopredný výber (jednokrokový)

K príznakov s najvyšším skóre



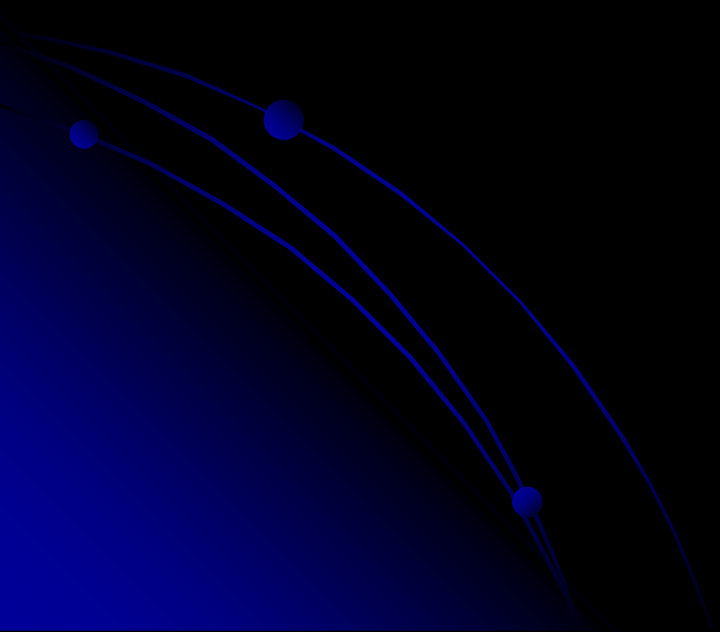
najlepší a najhorší spolu?

Iteratívny prístup

začni prázdnu množinou $\tilde{X} = \emptyset$
opakuj
 pre každý príznak $x_i \in X \setminus \tilde{X}$
 prerátaj skóre $\tilde{X} \cup \{x_i\}$
 vyber príznak s najvyšším skóre
kým nevyberieš K príznakov

Spätný výber (jednokrokový)

Z množiny príznačkov odstráň D-K príznačkov
s najnižším skóre



Iteratívny prístup

začni celou množinou príznakov $\tilde{X} = X$

opakuj

pre každý príznak $x_i \in \tilde{X}$

prerátaj skóre $\tilde{X} \setminus \{x_i\}$

odstráň príznak, ktorý maximalizuje skóre
kým neodstrániš D-K príznakov

Kombinovaný výber

- $L > R$: začni prázdnu množinou
opakuj
 - iteratívne pridaj L príznakov
 - iteratívne odober R príznakovkým nemáš K príznakov
- $L < R$: začni celou množinou príznakov
opakuj
 - iteratívne odober R príznakov
 - iteratívne pridaj L príznakovkým nemáš K príznakov

Hodnotiace miery

Miery vhodnosti príznakov

Filter:

Konzistencia

Medzitriedna vzdialenosť

Štatistická závislosť

Informačno-teoretické miery

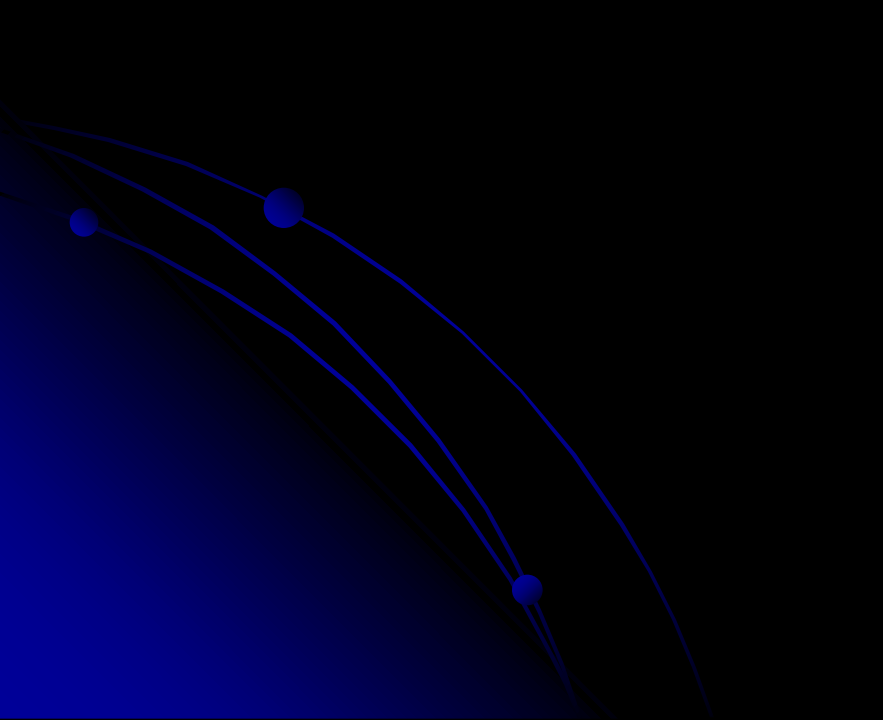
Obálka:

Dosiahnutá chyba klasifikátora

Konzistencia

Podmnožina príznačov musí separovať triedy tak konzistentne ako celá množina

Nekonzistencia, ak objekty s rovnakými príznačkami patria rôznym triedam



	Hair	Height	Weight	Lotion	Result
i_1	1	2	1	0	1
i_2	1	3	2	1	0
i_3	2	1	2	1	0
i_4	1	1	2	0	1
i_5	3	2	3	0	1
i_6	2	3	3	0	0
i_7	2	2	3	0	0
i_8	1	1	1	1	0

Sunburn data

M rovnakých vzoriek $\mathbf{x}$

m_i počet týchto vzoriek patriacich do triedy ω_i

$$\sum_{i=1}^C m_i = M$$

$$NJ(\mathbf{x}) = M - \max_i m_i$$

$$J(\tilde{X}) = 1 - \frac{\sum_{\mathbf{x} \in \text{Unique}(\tilde{X})} NJ(\mathbf{x})}{N}$$

Štatistická závislosť

Pearsonov (lineárny) korelačný koeficient dvoch premenných X a Y

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

nadobúda hodnoty $\langle -1, 1 \rangle$

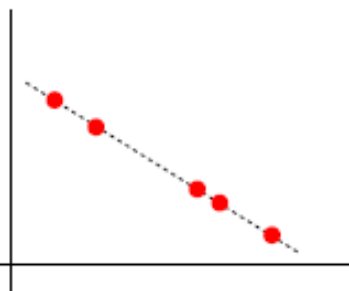
$\rho_{X,Y} = \pm 1$, ak premenné sú lineárne závislé

$\rho_{X,Y} = 0$, ak sú nekorelované

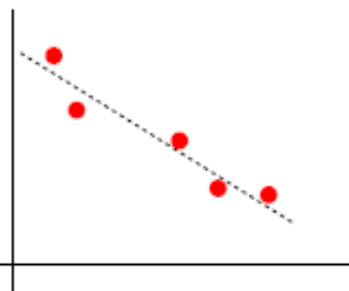
Nekorelovanosť $\neq$ nezávislosť

Iba ak X a Y majú normálne rozdelenie

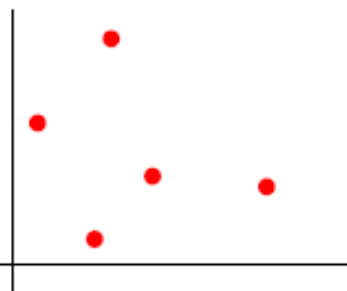
Závislosť $\rightarrow$ štatistická nadbytočnosť zdrojových dát



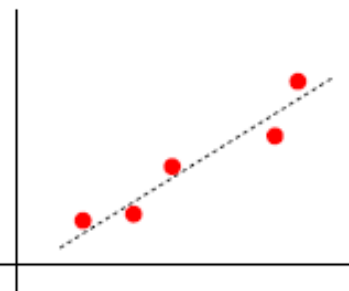
(a) $\rho = -1$



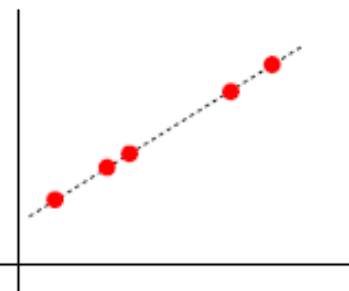
(b) $\rho \in (-1, 0)$



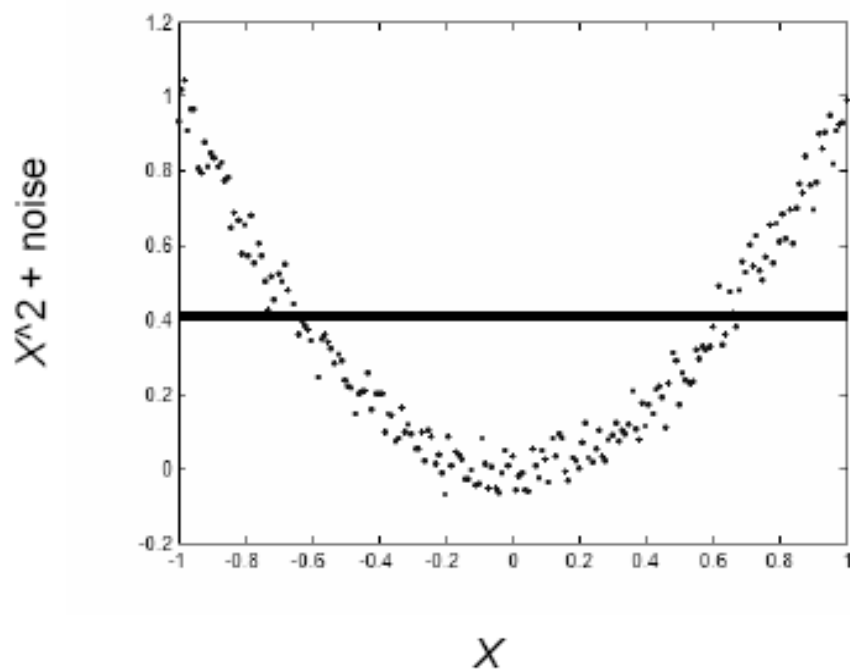
(c) $\rho = 0$



(d) $\rho \in (0, 1)$

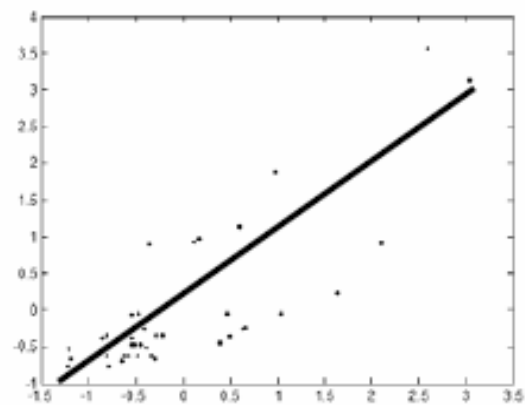


(e) $\rho = 1$



$$\rho_{XX^2} \approx 0$$

Number of murders



Number of churches

Informačno-teoretické miery

Entropia (neistota)

$$\begin{aligned} H(Y) &= -E(\log_2(P(Y = y))) = \\ &= -\sum_{y \in \Omega} P(Y = y) \cdot \log_2(P(Y = y)) \end{aligned}$$

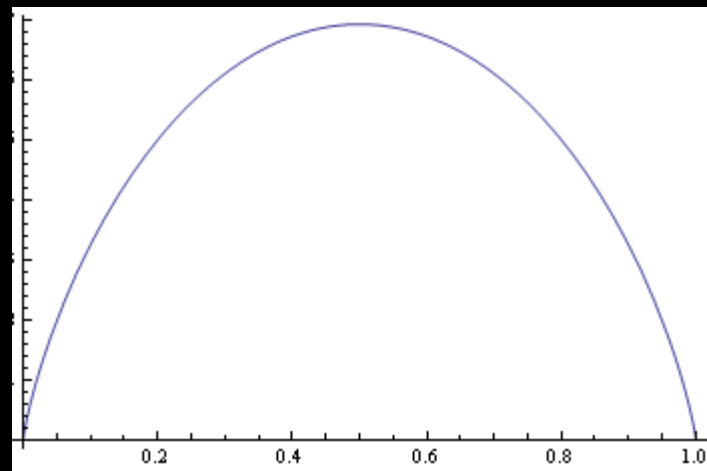
$$\Omega = \{0, 1\}$$

príklad

$$P(Y = 1) = p$$

$$P(Y = 0) = 1 - p$$

$$\begin{aligned} H(Y) &= -P(Y = 1) \cdot \log_2(P(Y = 1)) - P(Y = 0) \cdot \log_2(P(Y = 0)) \\ &= -p \cdot \log_2(p) - (1 - p) \cdot \log_2(1 - p) \end{aligned}$$



Entropy

X = College Major

Y = Likes "XBOX"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

$$\begin{aligned} H(Y) &= -E(\log_2(P(Y = y))) = \\ &= -\sum_{y \in \Omega} P(Y = y) \cdot \log_2(P(Y = y)) \end{aligned}$$

$$H(X) = 1.5$$

$$H(Y) = 1$$

Špecifická podmienená entropia

X = College Major

Y = Likes "XBOX"

$H(Y | X=v)$ = entropia len tých Y , $X=v$

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

$$H(Y|X=Math) = 1$$

$$H(Y|X=History) = 0$$

$$H(Y|X=CS) = 0$$

Podmienená entropia

X = College Major

Y = Likes "XBOX"

$H(Y|X)$ = priemerná špecifická
podmienená entropia Y

$$H(Y | X) = \sum_{x \in \Omega_X} P(X = x) \cdot H(Y | X = x)$$

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

x	P (X=x)	H(Y X = x)
Math	0.5	1
History	0.25	0
CS	0.25	0

$$H(Y|X) = .5$$

Vzájomná informácia

Ako sa znížia nároky (počet bitov) na prenos informácie Y , ak odosielateľ aj prijímateľ poznajú X ?

$$I(Y; X) = H(Y) - H(Y|X)$$

$$I(Y; X) = - \sum_{x \in \Omega_x} \sum_{y \in \Omega_y} P[Y = y, X = x] \log \frac{P[Y = y, X = x]}{P[Y = y]P[X = x]}$$

If X and Y are independent, $I(Y; X) = 0$

$$I(Y; Y) = H(Y)$$

$I(Y; X)$ is always non negative and less than $\min(H(Y), H(X))$

Vzájomná informácia

X = College Major

Y = Likes "XBOX"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

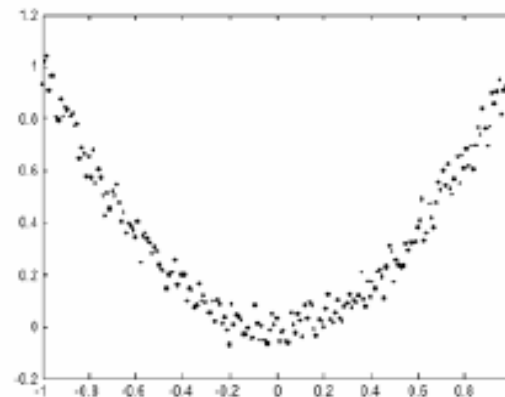
$$H(Y) = 1$$

$$H(Y|X) = 0.5$$

$$I(Y; X) = 0.5$$

Mutual information identifies nonlinear relationships between variables

X uniformly distributed over $[-1, 1]$
 Z uniformly distributed over $[-1, 1]$
 Z and X are independent
 $Y = X^2 + \text{noise}$



X

1000 samples	Y, Y	X, Y	Z, Y
Correlation	1	0.0460	0.0522
Mutual information	2.2582	1.1996	0.0030

Vzdialenosť

Euklidovská

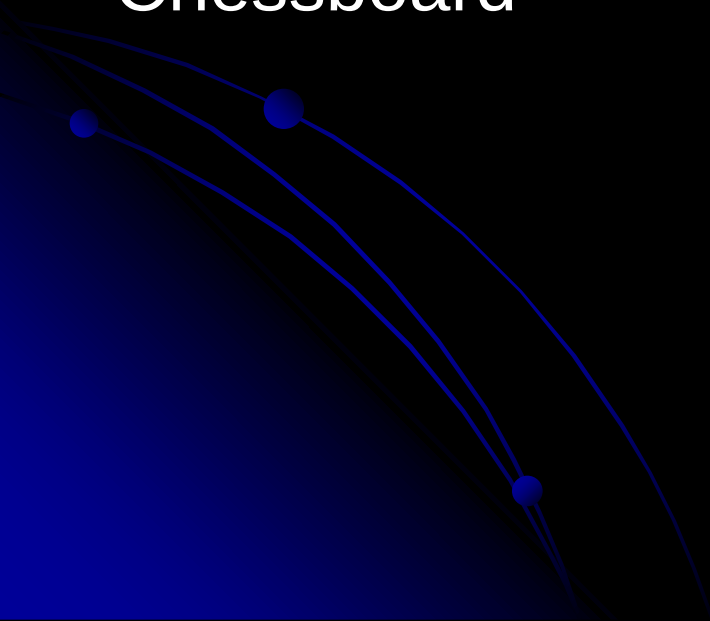
City Block (Manhattan)

Chessboard

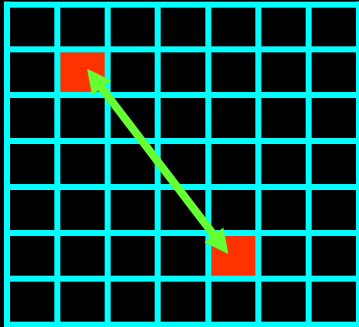
$$D_e(p_1, p_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

$$D_4(p_1, p_2) = |x_1 - x_2| + |y_1 - y_2|$$

$$D_8(p_1, p_2) = \max(|x_1 - x_2|, |y_1 - y_2|)$$



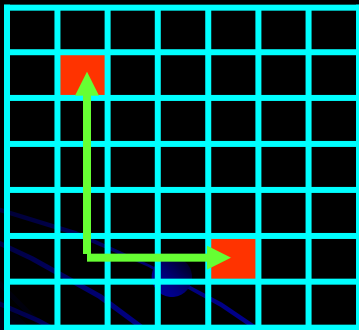
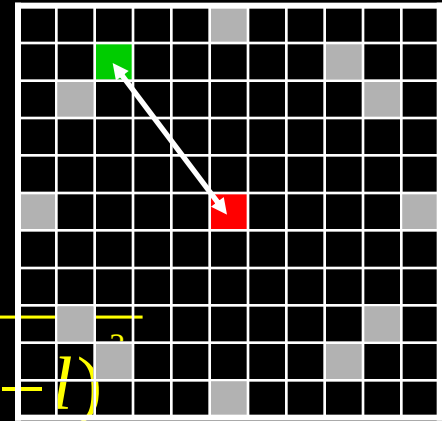
Metriky



Euklidovská

$$d([i, j], [k, l]) = \sqrt{(i - k)^2 + (j - l)^2}$$

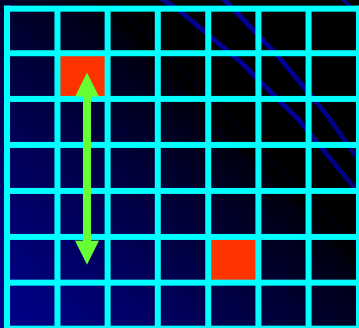
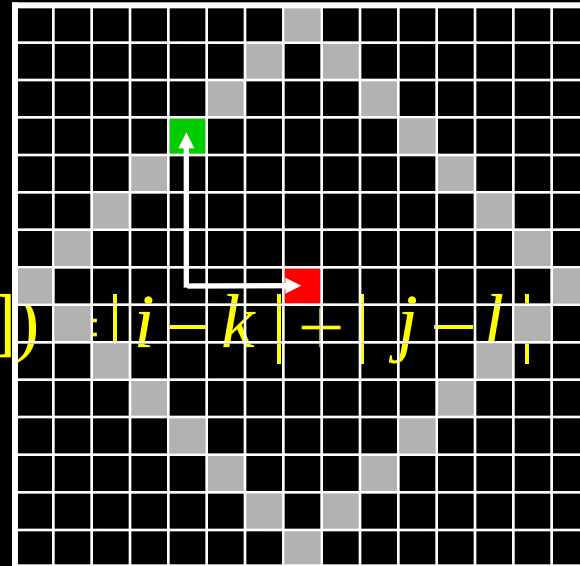
$d = 5$



City-block

$$d([i, j], [k, l]) = |i - k| + |j - l|$$

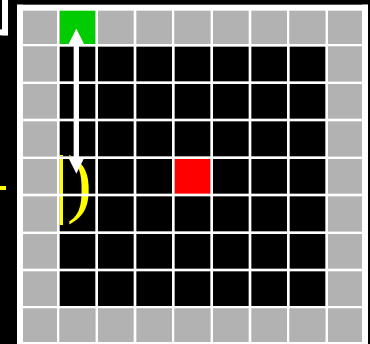
$d = 7$



Chessboard

$$d([i, j], [k, l]) = \max(|i - k|, |j - l|)$$

$d = 4$



Medzitriedna vzdialenosť

$$D_{\tilde{X}}(\omega_i, \omega_j) = \frac{1}{|\omega_i| |\omega_j|} \sum_{\mathbf{x} \in \omega_i} \sum_{\mathbf{y} \in \omega_j} d_{\tilde{X}}(\mathbf{x}, \mathbf{y})$$

$$J(\tilde{X}) = \sum_{i=1}^C P(\omega_i) \sum_{j=i+1}^C P(\omega_j) D_{\tilde{X}}(\omega_i, \omega_j)$$

