Rozpoznávanie obrazcov šk.r. 2019-20

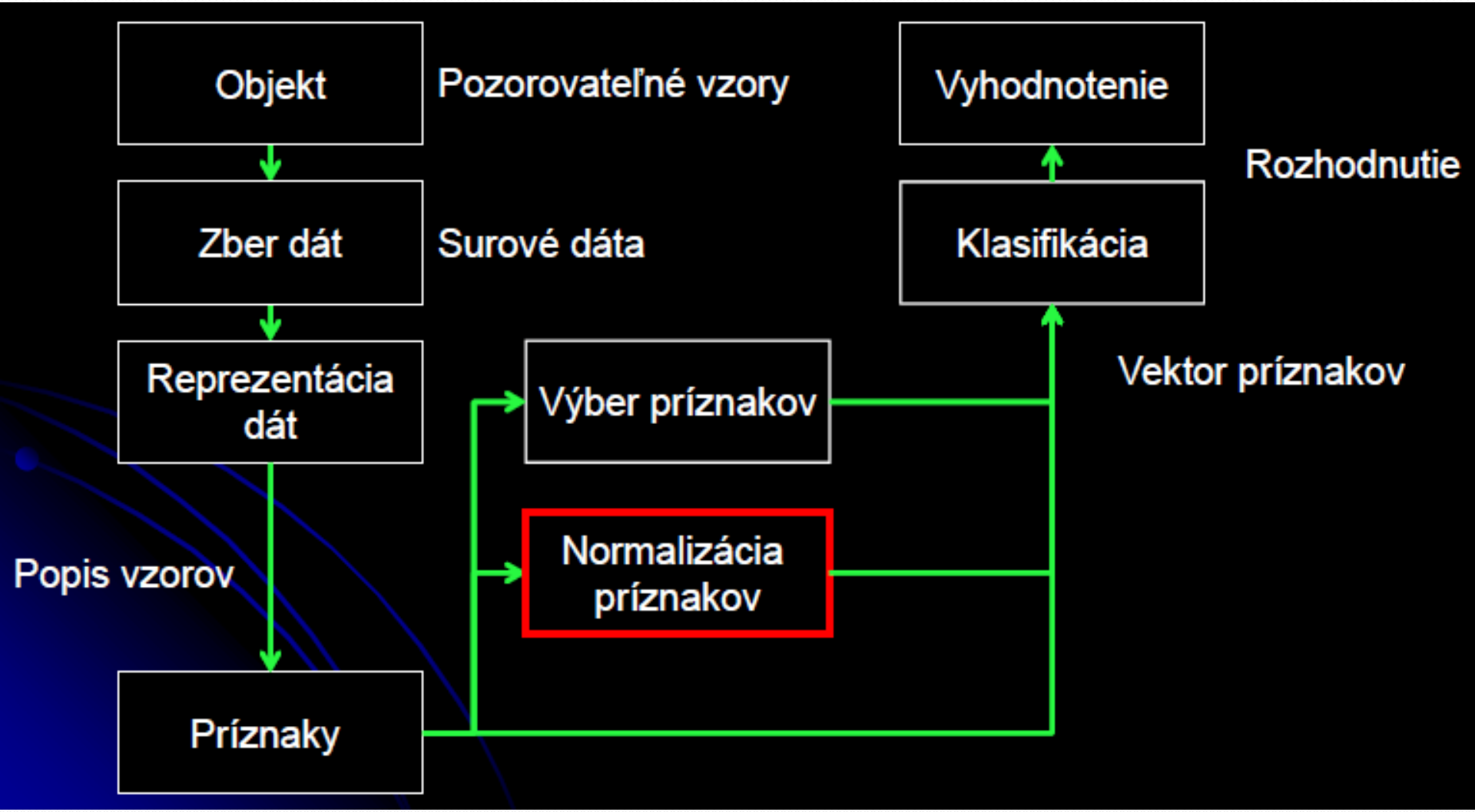
Príznaky

Doc. RNDr. Milan Ftáčnik, CSc.

Rekapitulácia

- Základný súbor, výberový súbor a štatistika ako funkcia náhodných premenných výberu
- Bodové odhady normálneho rozdelenia
- Intervalové odhady na hladine významnosti
- Testovacie štatistiky pre priemer a rozptyl
- Testovanie hypotéz a kritická oblasť
- Vlastné čísla, vlastné vektory a SVD
- Hľadanie viazaného extrému Lagrangeovou metódou

Práca s príznakmi



Príznaky bez rozmeru

- Príznaky relatívne k určitej referenčnej hodnote
- Príklad: Skutočná výška hráčov v tíme
- Bezrozmerný príznak výšky: vezmime referenčnú hodnotu (výška najmenšieho, najväčšieho, 1 m, 100 cm, ...) a vypočítame:
- Bezrozmerný príznak výšky = (skutočná výška) – (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 II

- Lineárne škálovanie na nulový priemer a jednotkový rozptyl

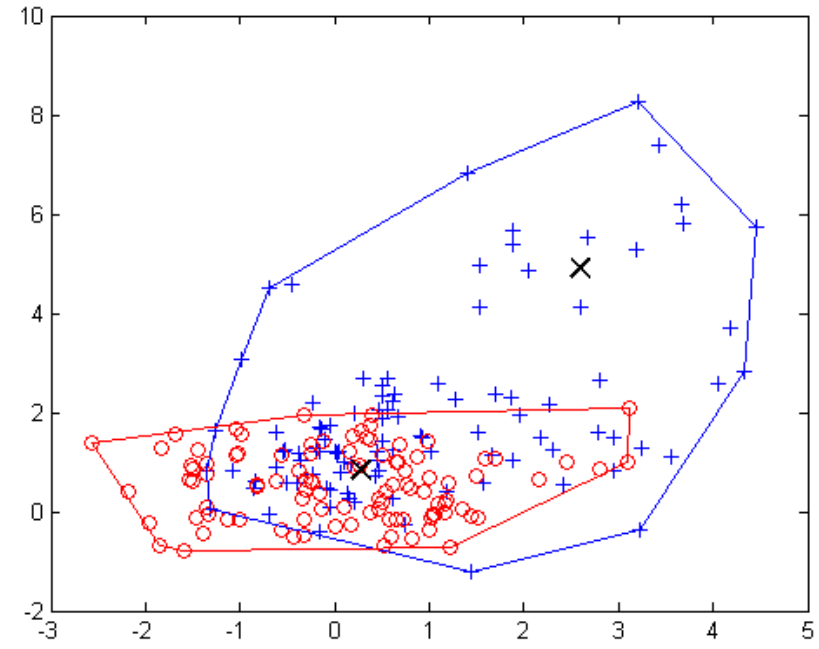
$$\tilde{x}_i = \frac{x_i - \bar{x}}{s}$$

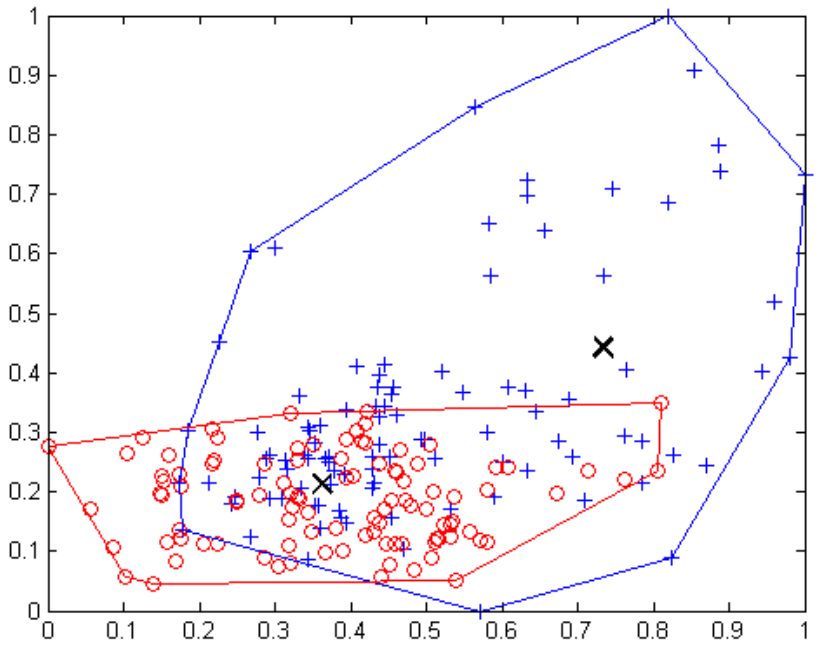
- Normálne rozdelenie s podmienkou

$$\tilde{x}_i = \frac{\frac{x_i - \mu}{3\sigma} + 1}{2}$$

Využitím pravidla 3σ dáta z normálneho rozdelenia naškálujeme tak, aby 99% dát ležalo v intervale [0,1]

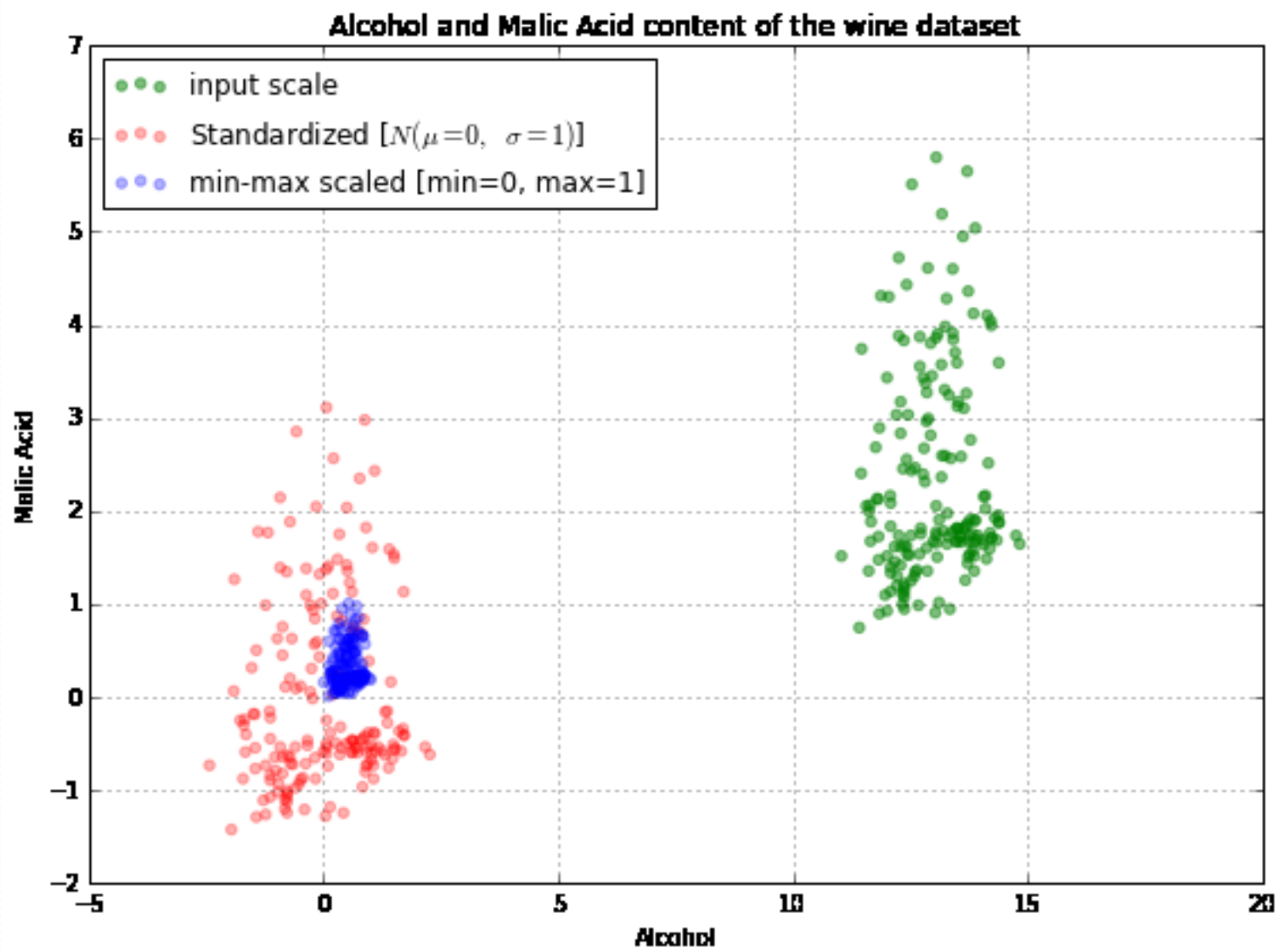
Príklad


10
8
6
4
2
0
-2
-3
-2
-1
0
1
2
3
4
5

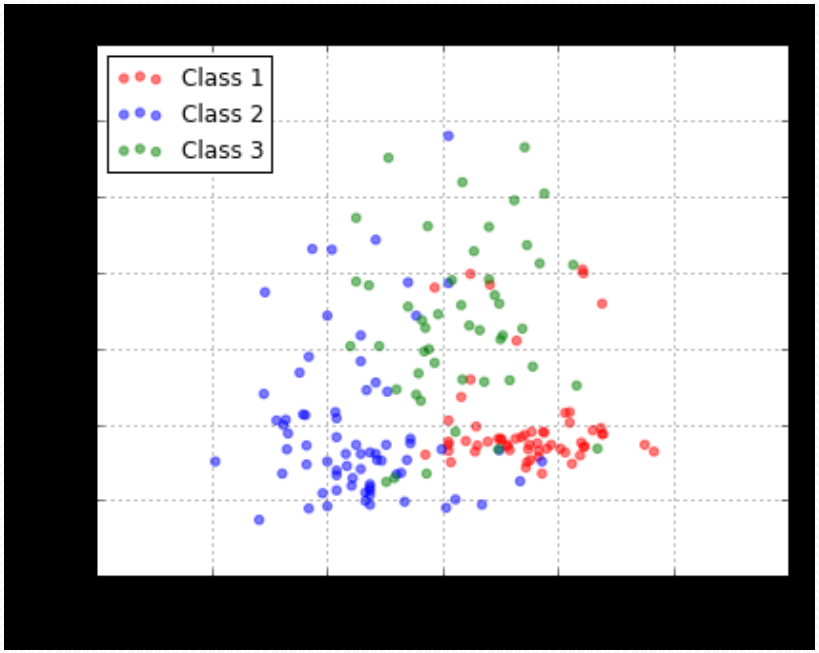

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

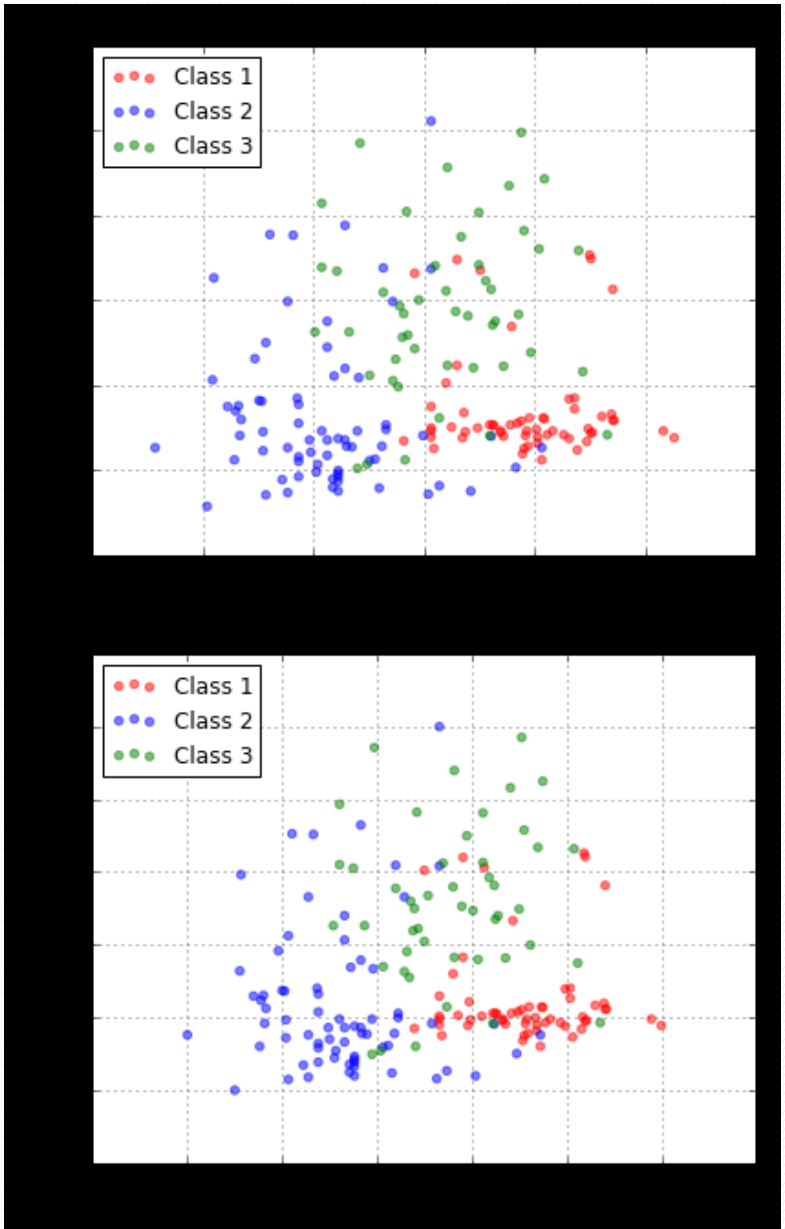
Druhý príklad

https://raw.githubusercontent.com/rasbt/pattern_classification/master/data/wine_data.csv



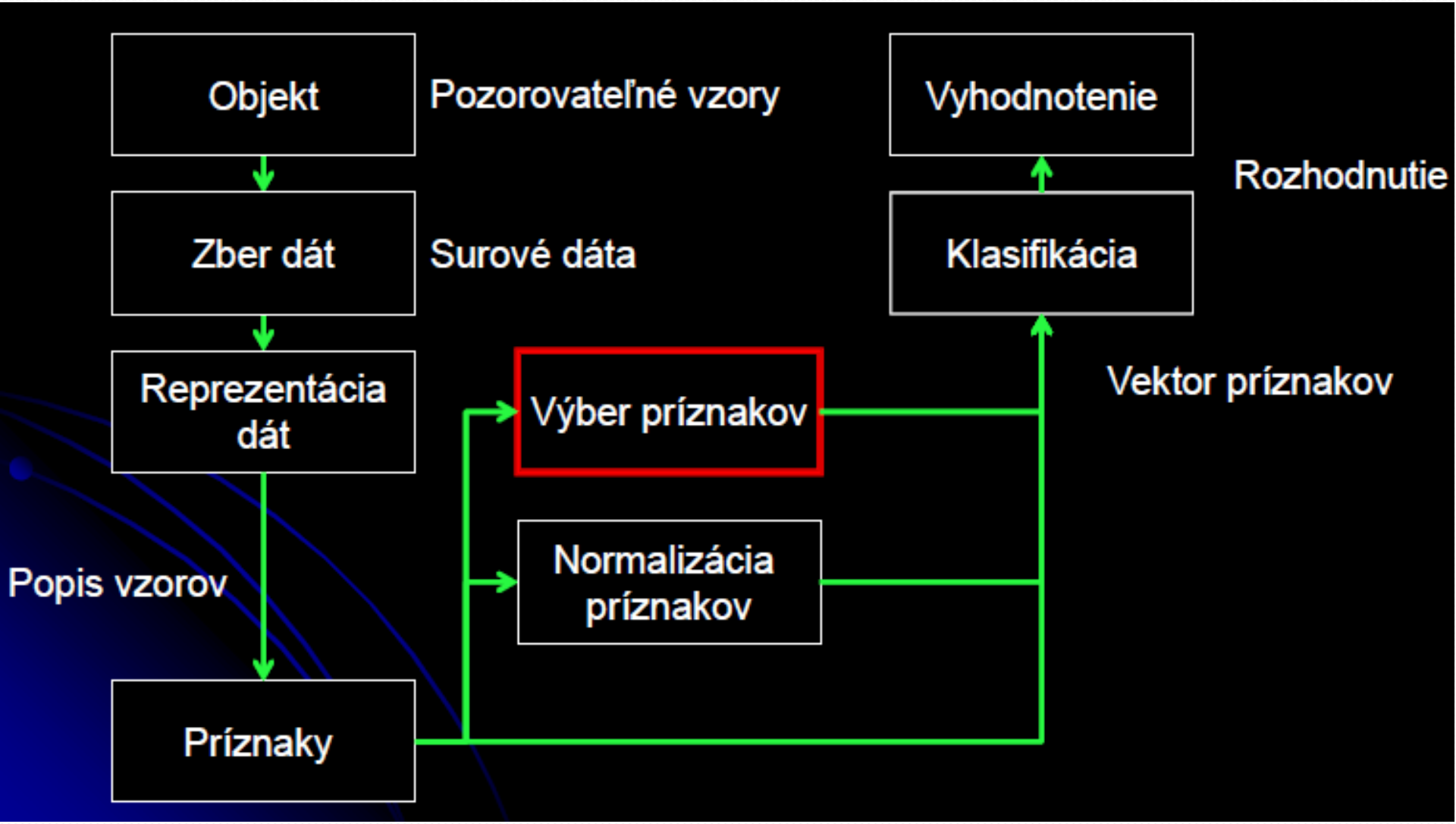
Druhý príklad II





- Treba zistiť, či sú normalizované príznaky vhodnejšie a rozhodnúť sa
- Niektoré klasifikátory nereagujú na normalizáciu

Výber príznakov



Zníženie počtu príznakov

- Zamyslieť sa nad príznakmi – nejaké očividné vzťahy? napr. F2 a F3 alebo F4 a F5

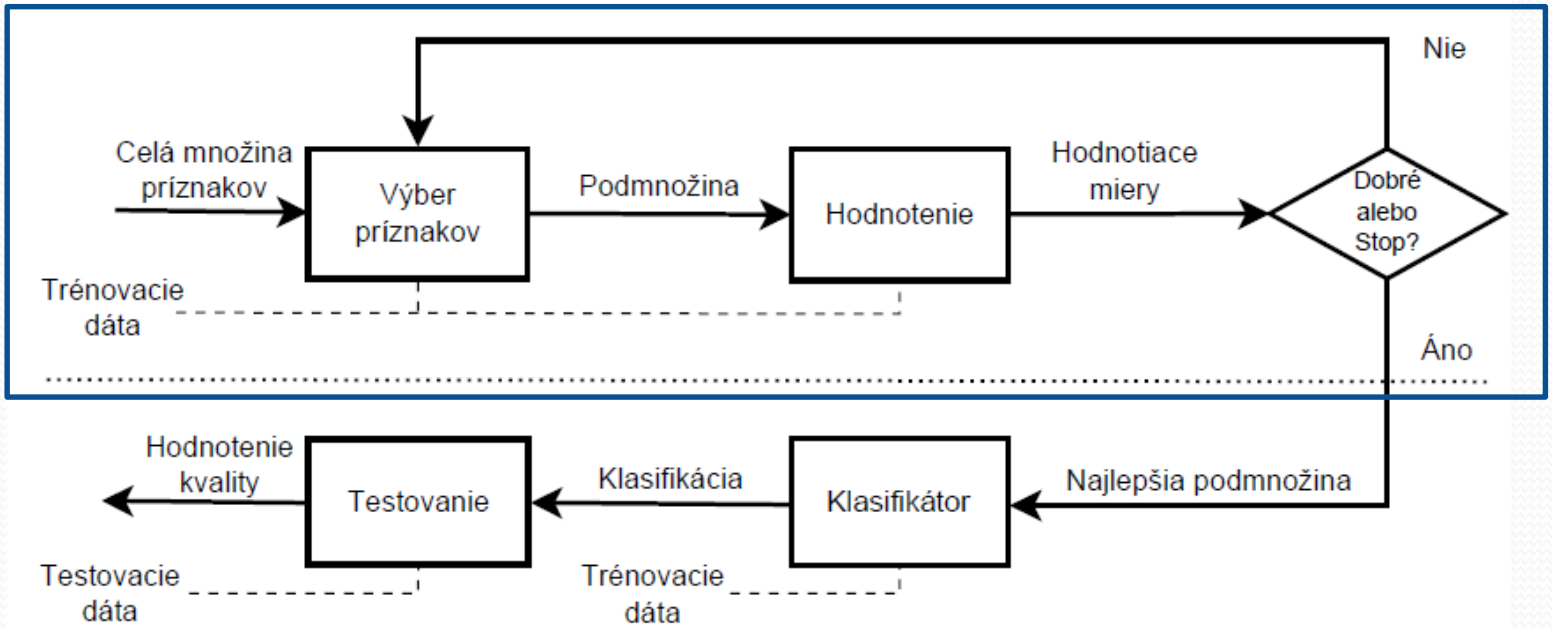
F_1	F_2	F_3	F_4	F_5	C
0	0	1	0	1	0
0	1	0	0	1	1
1	0	1	0	1	1
1	1	0	0	1	1
0	0	1	1	0	0
0	1	0	1	0	1
1	0	1	1	0	1
1	1	0	1	0	1

Dva prístupy

- **Výber príznakov:**
  - Vyberieme podmnožinu z originálnych príznakov
  - Feature selection (teda nie feature extraction)
- **Redukcia príznakov:**
  - Transformujeme pôvodnú množinu príznakov do menej-dimenzionálnej

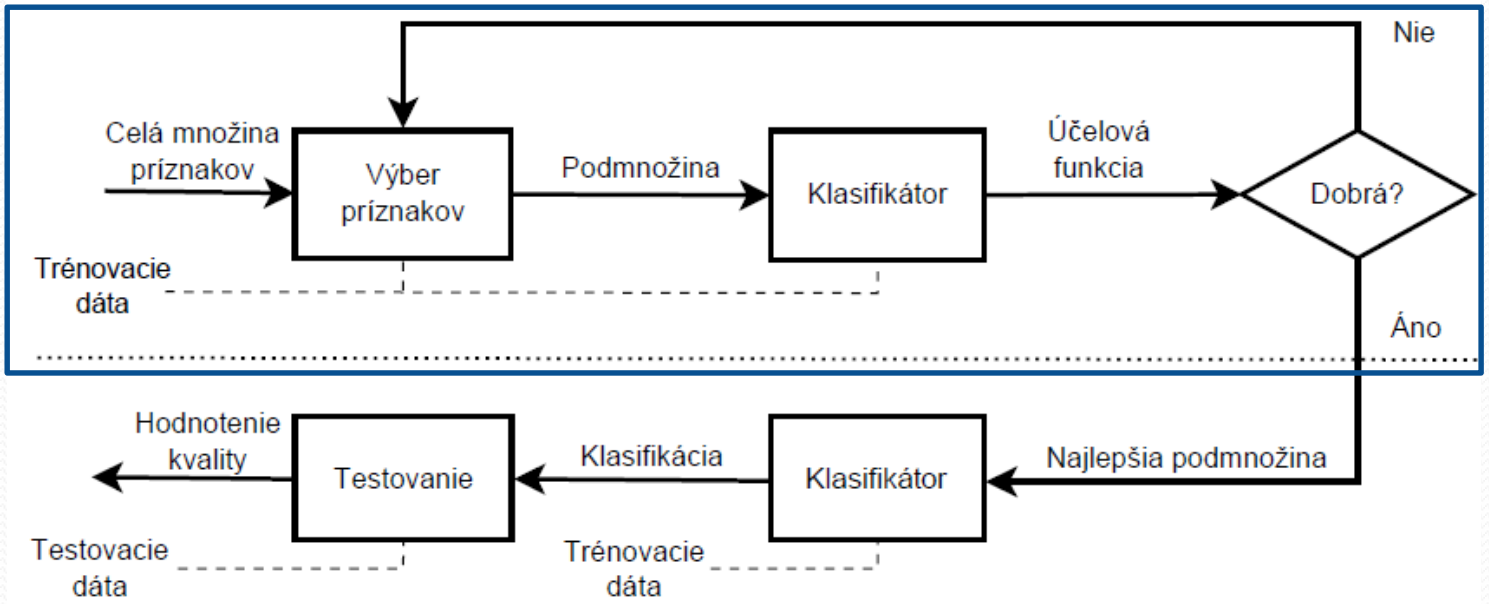
Výber príznakov

- Filter
  - Oddelenie procesu výberu od klasifikátora
  - Hodnotenie dát (miery: *vzájomná informácia, vzdialenosť, závislosť, konzistencia)*



Výber príznakov II

- Obálka
  - Závisí od určeného klasifikátora
  - Očakávaná presnosť klasifikátora
  - Presné, výpočtovo náročné



Výber príznakov III

- V oboch prípadoch môžeme hodnotiť príznaky samostatne alebo celé podmnožiny príznakov
- Jednotlivé príznaky – efektívne, ľahko implementovateľné, ale nemusíme odhaliť nadbytočnosť alebo koreláciu medzi príznakmi
- Taktiež je problém určiť správne počet príznakov, ktoré vyberieme do výslednej podmnožiny
- Hodnotenie podmnožín je náročnejšie. Pri D príznakoch máme 2^D podmnožín.

Výber vhodných príznakov

- Najjednoduchší výber je dopredný alebo spätný výber
- Sofistikovanejšie prístupy:
  - Kombinácia dopredného a spätného
  - Exponenciálne prehľadávanie (vetvenie a orezávanie, lúčové hľadanie)
  - Randomizované algoritmy (simulované žíhanie, genetické algoritmy, atď.)

Dopredný výber

- Jednokrokový: Ak chceme z *D* príznakov vybrať *K* $(K < D)$, pri tomto prístupe ohodnotíme všetky príznaky zvolenou mierou a priamo vyberieme *K* najlepších
- Tento postup málokedy funguje, pretože neberie do úvahy vzťahy
- Iná možnosť je iteratívny prístup

Dopredný výber II

- Iteratívny prístup:
- Začni s prázdnou 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: Odstraňujeme naraz alebo postupne príznaky.
- Ohodnotíme všetky príznaky zvolenou mierou a odstránime *D-K* najhorších
- Má podobné problémy ako jednokrokový dopredný
- Preto hľadáme aj iné riešenia

Spätný výber II

- Iteratívny:
- Začni s celou množinou príznakov $\tilde{X} = X$
- Opakuj

  pre každý príznak $x_i \in \tilde{X}$

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

  odstráň príznak, kt maximalizuje skóre
- kým neodstrániš *D-K* príznakov

Kombinovaný výber

- $L > R$: začni prázdnou množinou

  opakuj

  iteratívne pridaj *L* príznakov

  iteratívne odober *R* príznakov

  kým nemáš *K* príznakov

- $L < R$: začni celou množinou príznakov

  opakuj

  iteratívne odober *R* príznakov

  iteratívne pridaj *L* príznakov

  kým nemáš *K* príznakov

Zhodnotenie metód

- Jednokrokový výber – univariátne metódy, ohodnocujúce vždy jeden príznak
- Iteratívny prístup – multivariátne metódy, kde ohodnocujeme množinu príznakov
- Treba vhodné hodnotiace miery na ohodnotenie vhodnosti množiny príznakov

Hodnotiace miery

- Miery vhodnosti príznakov
- **Filter:**
  - Konzistencia
  - Medzitriedna vzdialenosť
  - Štatistická závislosť
  - Informačno-teoretické miery
- **Obálka:**
  - Dosiahnutá chyba klasifikátora

Konzistencia

- Podmnožina príznakov musí separovať triedy tak konzistentne ako celá množina
- Nekonzistencia: ak objekty s rovnakými príznakmi 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

Konzistencia II

- Nech v množine $\tilde{X}$ obsahujúcej *N* vzoriek reprezentovaných *K* vybranými príznakmi existuje *M* rovnakých vzoriek $\boldsymbol{x}$ a nech m_i je počet týchto vzoriek patriaci do triedy ω_i.
- Potom miera konzistencie $\tilde{X}$ je

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

Konzistencia III

- kde $Unique(\tilde{X})$ je množina všetkých vzájomne rôznych vzoriek z $\tilde{X}$
- $\sum_{i=1}^{C} m_i = M$
- $NJ(\boldsymbol{x}) = M - \max_i m_i$
- Prečo to nezávisí od $\boldsymbol{x}$?

Štatistická závislosť

- Skúma koreláciu a vychádza z toho, že príznak (alebo množina príznakov) vhodný na klasifikáciu je silno korelovaný s predikovanou triedou a zároveň nekorelovaný s inými príznakmi
- Korelácia sa určuje cez Pearsonov (lineárny) korelačný koeficient dvoch premenných X a Y určený takto:

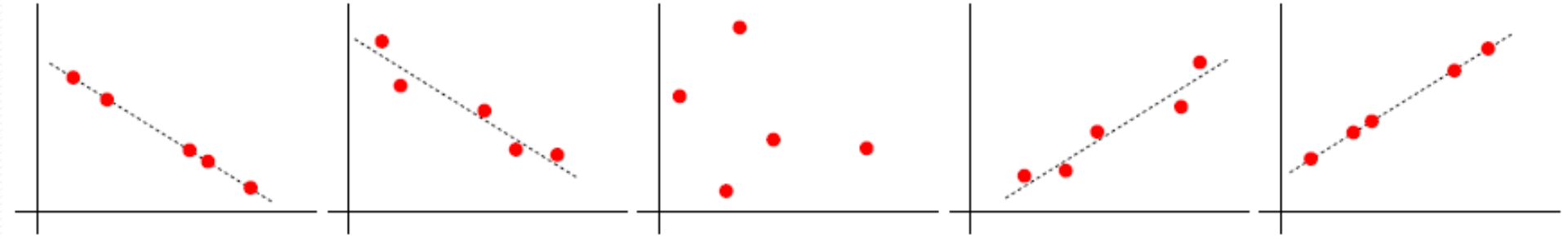
Štatistická závislosť II

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

- Koeficient nadobúda hodnoty $\langle -1,1 \rangle$
- $\rho_{X,Y} = \pm 1$, ak premenné sú lineárne závislé
- $\rho_{X,Y} = 0$, ak sú nekorelované
- Nekorelovanosť≠nezávislosť (iba ak majú *X* a *Y* normálne rozdelenie)
- Závislosť → štatistická nadbytočnosť dát

Štatistická závislosť III

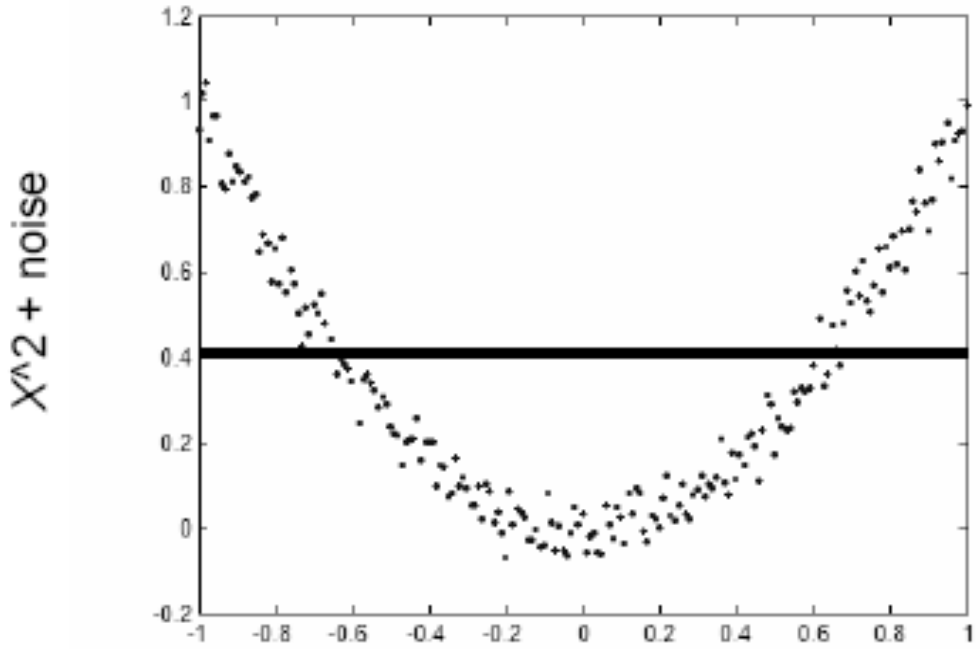
- Ukážka rôznych hodnôt korelačného koeficientu pri rozličnom usporiadaní zdrojových dát



(a) $\rho = -1$ (b) $\rho \in (-1, 0)$ (c) $\rho = 0$ (d) $\rho \in (0, 1)$ (e) $\rho = 1$

Štatistická závislosť IV

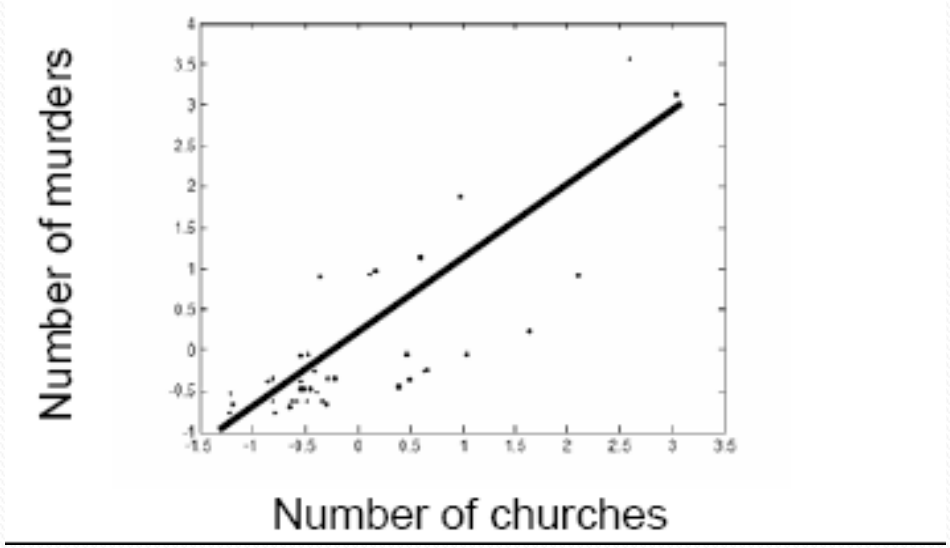
- Nízka korelácia neznamená absenciu vzťahu medzi dátami. Môže medzi nimi existovať nelineárny vzťah



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

Štatistická závislosť V

- Korelácia neznamená kauzalitu (príčinnú súvislosť)
- Kladná korelácia znamená, že sa dané javy vyskytujú často spolu, ale nie to, že jeden jav je príčinou druhého
- Príčinou môže byť iný (nepozorovaný) jav



Informačno-teoretické miery

- **Hartleyho miera informácie**
- Správa dĺžky *n*, počet rôznych symbolov *s*
- Množstvo informácie v správe je funkciou počtu možných správ $N = s^n$:
- $\Im = f(N)$
- Uvažujme dve správy dĺžky n_1 a n_2. Spojíme ich do jednej správy. Ktorá funkcia *f* spĺňa:

Informačno-teoretické miery II

$$\Im = \Im_1 + \Im_2$$

$$f(s^{n_1+n_2}) = f(s^{n_1}) + f(s^{n_2})$$

$$f(N_1.N_2) = f(N_1) + f(N_2)$$

- Tou funkciou je logaritmus
- Preto **Hartleyho miera informácie** je:
- $\Im = \log(N) = \log(s^n) = n.\log(s)$

Informačno-teoretické miery III

- **Shannonova miera informácie**
- Majme diskrétnu náhodnú premennú *A* s možnými výstupmi $\{a_1, \dots, a_n\}$ a pravdepodobnostnú funkciu $P(A = a_i) = p_i$
- Informácia, ktorú dostaneme, keď pozorujeme výstup a_i je
- $\Im(a_i) = -\log_2(P(A = a_i))$

Shannonova entropia

- Entropia (neistota)= stredná hodnota informácie

$$H(A) = E(\Im(A))$$
$$= -E(\log_2(P(A))) =$$
$$= -\sum_{a \in \Omega} P(A = a).\log_2(P(A = a))$$

- Pri Hartleym sme predpokladali rovnakú pravdepodobnosť výskytu znaku

Shannonova entropia II

- Vlastnosti Shannonovej entropie:

- $H(A) \leq \log(N)$

  $H(A) = \log(N) \leftrightarrow \forall i \; P(A = a_i) = 1/N$

- $H(A) \geq 0$

  $H(A) = 0 \leftrightarrow \exists k \; P(A = a_k) = 1$

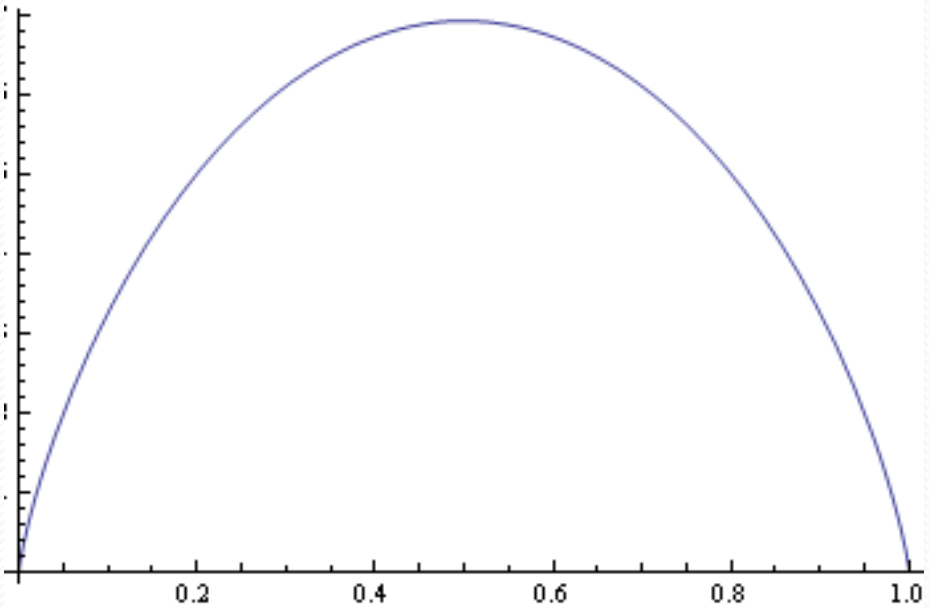
Shannonova entropia III

- Príklad

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

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

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



$$H(Y) = -P(Y = 1).\log_2(P(Y = 1)) - P(Y = 0).\log_2(P(Y = 0))$$

$$= -p.\log_2(p) - (1 - p).\log_2(1 - p)$$

Shannonova entropia IV

- X – predmet
- Y – má rád XBOX

$$H(Y) = -E(\log_2(P(Y = y))) =$$

$$= -\sum_{y \in \Omega} P(Y = y).\log_2(P(Y = y))$$

X	Y
Matematika	Áno
História	Nie
Informatika	Áno
Matematika	Nie
Matematika	Nie
Informatika	Áno
História	Nie
Matematika	Áno

H(X) = 1,5

H(Y) = 1

Špecifická podmienená entropia

- X – predmet
- Y – má rád XBOX

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

X	Y
Matematika	Áno
História	Nie
Informatika	Áno
Matematika	Nie
Matematika	Nie
Informatika	Áno
História	Nie
Matematika	Áno

$$H(Y|X = M) = 1$$
$$H(Y|X = H) = 0$$
$$H(Y|X = I) = 0$$

Podmienená entropia

- X – predmet
- Y – má rád XBOX

$H(Y|X)$ priemerná špeci-fická podmien. entropia

X	Y
Matematika	Áno
História	Nie
Informatika	Áno
Matematika	Nie
Matematika	Nie
Informatika	Áno
História	Nie
Matematika	Áno

$$H(Y \mid X) = \sum_{x \in \Omega_X} P(X = x).H(Y \mid X = x))$$

x	P(X=x)	$H(Y\|X = x)$
Matematika	0.5	1
História	0.25	0
Informatika	0.25	0

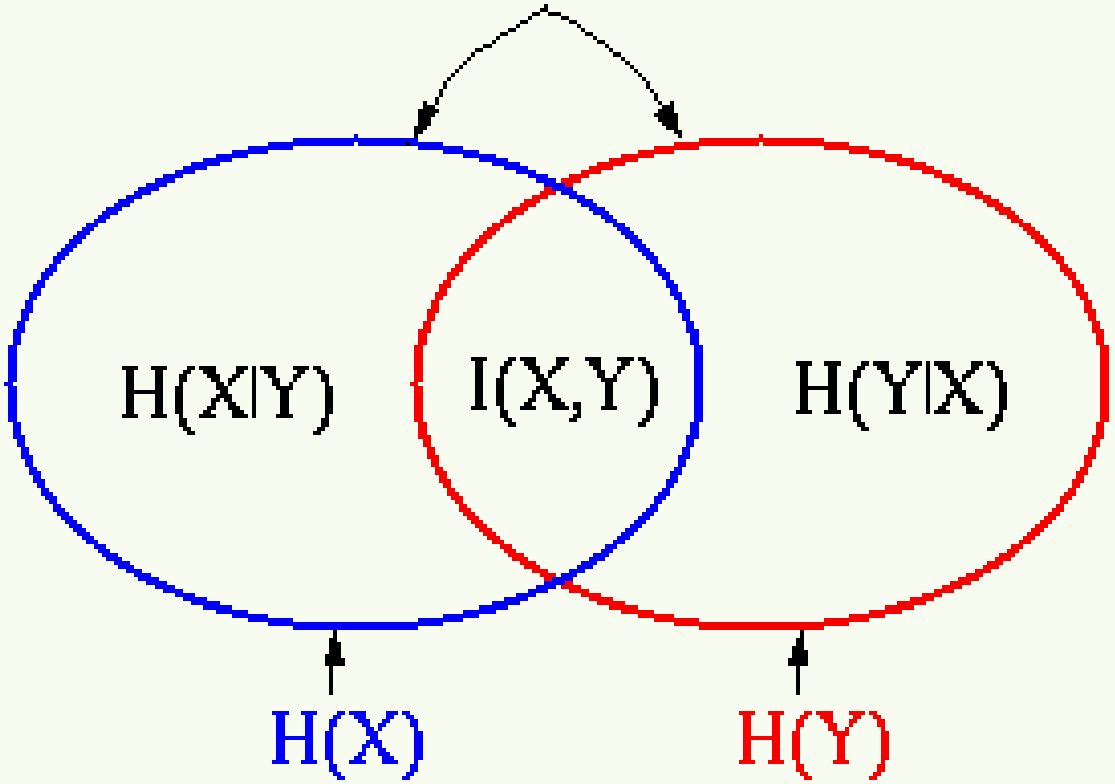
$$H(Y|X) = 0.5$$

Vzájomná informácia

- Ako sa znížia nároky (počet bitov) na prenos informácie *Y*, ak odosielateľ aj prijímateľ poznajú *X?*
- $I(Y;X) = H(Y) - H(Y|X)$
- $I(Y;X) = \sum_{x \in \Omega_x} \sum_{y \in \Omega_y} P[Y = y, X = x] \, log \frac{P[Y=y,X=x]}{P[Y=y]P[X=x]}$
- Ak *X* a *Y* sú nezávislé, tak $I(Y;X) = 0$
- $I(Y;Y) = H(Y)$
- $I(Y;X)$ je vždy nezáporné a $\leq \min(H(Y), H(X))$

Vzájomná informácia II

$$H(X,Y)=H(X|Y)+H(Y|X)+I(X,Y)$$



$$I(X,Y)=H(X)+H(Y)-H(X,Y)=H(X)-H(X|Y)=H(Y)-H(Y|X)$$

Vzájomná informácia III

- X – predmet
- Y – má rád XBOX

X	Y
Matematika	Áno
História	Nie
Informatika	Áno
Matematika	Nie
Matematika	Nie
Informatika	Áno
História	Nie
Matematika	Áno

$$H(Y) = 1$$

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

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

Vzájomná informácia IV

- Ohodnotenie množiny príznakov:
- $J(\tilde{X}) = I(\tilde{X}; y)$
- Celá podmnožina

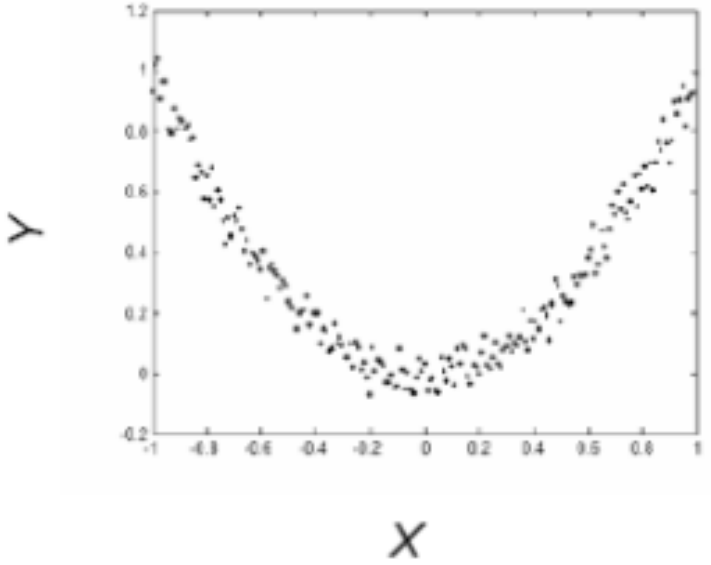
$$X^* = \arg\max_{X_S \subset X} I(X_S; y)$$

- Jednotlivé príznaky

$$x^* = \arg\max_{x_k \in X \setminus X_S} I(\{X_S, x_k\}; y)$$

Vzájomná informácia V

- Vzájomná informácia identifikuje nelineárne vzťahy medzi premennými
- *X* má rovnom. rozdelenie na $[-1,1]$
- *Z* má rovnom. rozdelenie na $[-1,1]$
- *Z* a *X* sú nezávislé
- $Y = X^2 + \check{s}um$



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

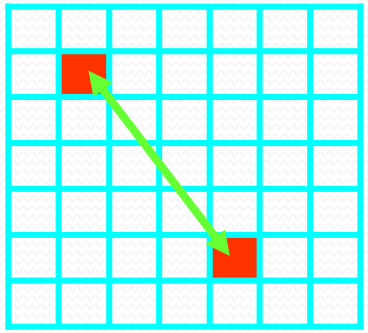
Medzitriedna vzdialenosť

- Ako definovať vzdialenosť?
- Pomocou metriky
- Euklidovská (bod – bod)
- Mahalanobisova (bod – množina)
- Bhattachayova (množina – množina)
- Hellingerova (množina – množina)
- …

Medzitriedna vzdialenosť II

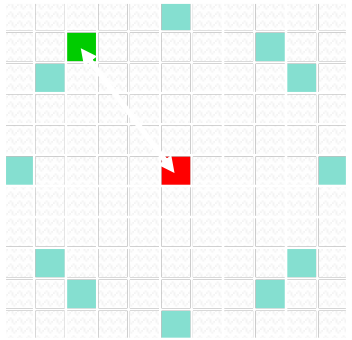
- Ako definovať vzdialenosť v diskrétnom priestore?
- Metriky:
- Euklidovská $D_e(p_1, p_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$
- Manhattanská $D_4(p_1, p_2) = |x_1 - x_2| + |y_1 - y_2|$
- Šachovnicová $D_8(p_1, p_2) = \max(|x_1 - x_2|, |y_1 - y_2|)$

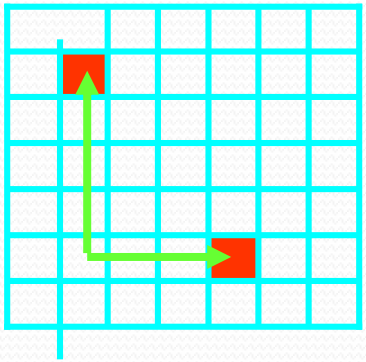
Medzitriedna vzdialenosť III



Euklidovská

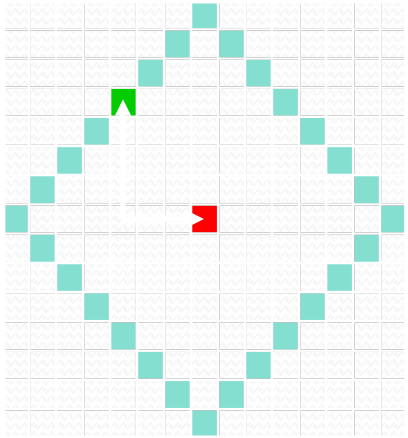
d = 5

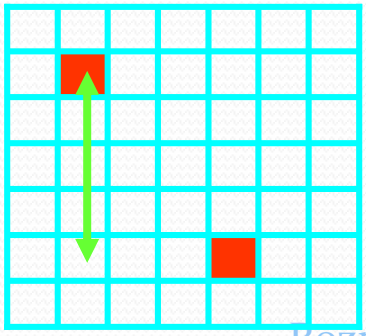




Manhattanská

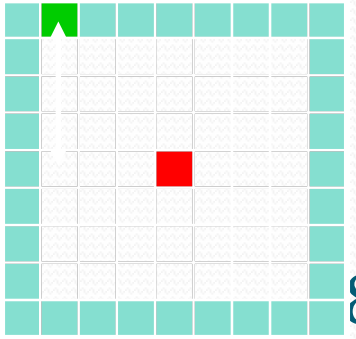
d = 7





Šachovnicová

d = 4



Medzitriedna vzdialenosť IV

- Príznak *i* budeme preferovať pred príznakom *j*, ak lepšie separuje klasifikačné triedy, t.j. ak ich vzdialenosť je väčšia pri použití príznaku *i*

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

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

Hodnotiace miery

- Miery vhodnosti príznakov
- **Filter:**
  - Konzistencia
  - Medzitriedna vzdialenosť
  - Štatistická závislosť
  - Informačno-teoretické miery
- **Obálka:**
  - Dosiahnutá chyba klasifikátora

Hľadanie optimálnej podmnožiny

- **Dopredný výber**
- Inicializuj $s = \{\ \ \}$
- Vykonaj:
- Pridaj príznak ku s,
- ktorý najviac zlepšuje OF*(s)*
- kým sa OF(s) dá zlepšiť

Spätný výber

Inicializuj $s = \{1,2, \dots, n\}$

Vykonaj:

Odober príznak z s,

ktorý najviac zlepšuje OF*(s)*

kým sa OF(s) dá zlepšiť

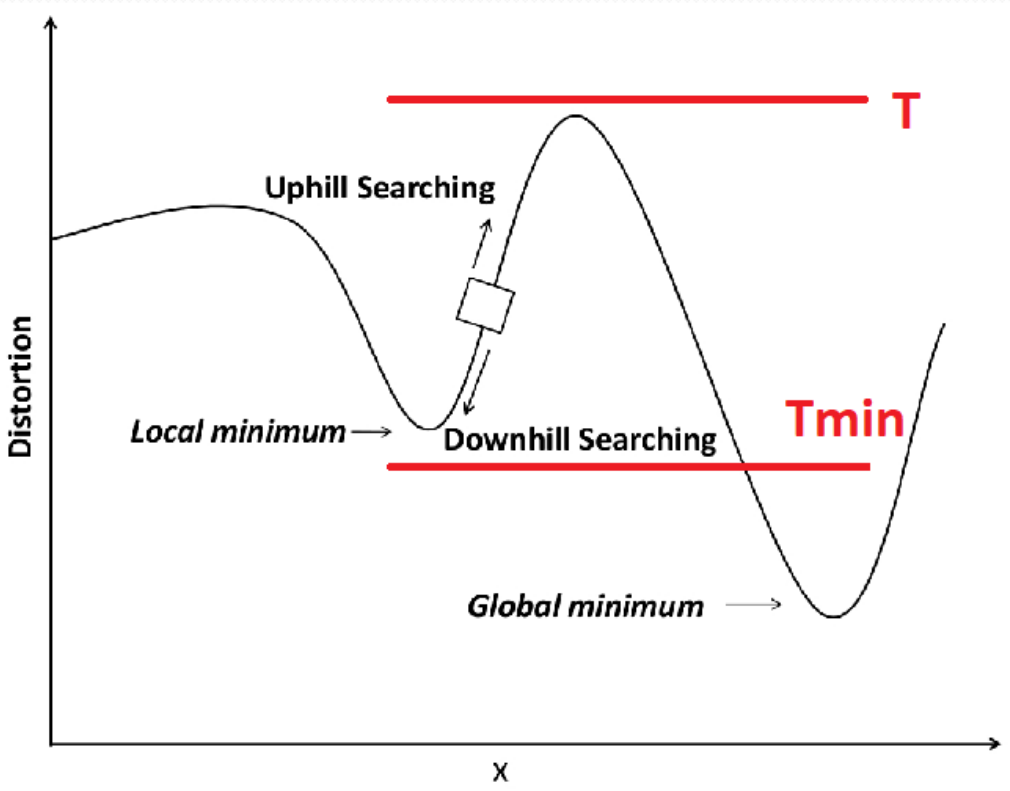
- Spätná eliminácia nachádza lepšie modely
- Problém môže byť pri veľkých dátach na začiatku procesu
- Obe môžu byť priveľmi pažravé (greedy)

Simulované žíhanie

- Vytvoril ho fyzik Vlado Černý – FMFI UK inšpiroval sa zohrievaním telesa a jeho postupným chladením (znižovaním teploty) s cieľom dosiahnuť optimálnu kryštalizáciu
- Je to pravdepodobnostná technika na nájdenie globálneho optima danej funkcie vo veľkom (často diskrétnom) prehľadávacom priestore pomocou nového parametra teplota

Simulované žíhanie II

- Hľadá riešenie v smere najväčšieho zlepšenia - gradientu (ako horolezecký algoritmus), ale umožňuje sa dostať z lokálneho optima tým, že zhorší hodnotiacu funkciu až do výšky parametra teplota, ktorá sa postupne znižuje



Simulované žíhanie III

- Určenie chladiacej funkcie T_i
- Určenie počiatočného riešenia $Y_0 = \{0,1\}^D$
- Kým $T_i > T_{min}$

    1. nájdenie nového riešenia Y_{i+1} lokálnym prehľadávaním okolia Y_i

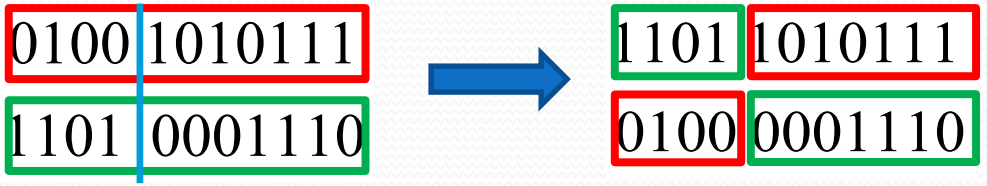
    2. Výpočet $\Delta E = J(Y_i) - J(Y_{i+1})$

    3. Ak $\Delta E < 0$, tak akceptovanie riešenia, inak akceptovanie s pravdepodobnosťou $P = \min(1, e^{-\frac{\Delta E}{T_i}})$

    4. $T_{i+1} = f(T_i)$

Genetické algoritmy

- Vytvorenie počiatočnej náhodnej populácie $\{Y_i\}_{i=1}^{N}$
- Ohodnotenie jedincov populácie $J(Y_i)$
- Opakuj
  1. Náhodný výber jedincov na báze ohodnotenia
  2. Vytvorenie nového jedinca krížením



  3. Mutácia jedincov

  4. Ohodnotenie novej populácie