ANALÝZA A KLASIFIKACE DAT

prof. Ing. Jiří Holčík, CSc.

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

IV.
KLASIFIKACE PODLE MINIMÁLNÍ VZDÁLENOSTI

PRINCIPY KLASIFIKACE

- pomocí diskriminačních funkcí – funkcí, které určují míru příslušnosti k dané klasifikační třídě;
- pomocí definice hranic mezi jednotlivými třídami a logických pravidel;
- pomocí vzdálenosti od reprezentativních obrazů (etalonů) klasifikačních tříd;
- pomocí ztotožnění s etalony;

PRINCIPY KLASIFIKACE

- pomocí diskriminačních funkcí – funkcí, které určují míru příslušnosti k dané klasifikační třídě;
- pomocí definice hranic mezi jednotlivými třídami a logických pravidel;
- pomocí vzdálenosti od reprezentativních obrazů (etalonů) klasifikačních tříd;
- pomocí ztotožnění s etalony;

METRIKA - VZDÁLENOST

Metrika ρ na X je funkce $\rho: X \times X \rightarrow R$, kde R je množina reálných čísel, taková, že:

$$\exists \rho_0 \in R: -\infty < \rho_0 \leq \rho(x,y) < +\infty, \forall x,y \in X$$

$$\rho(x,x) = \rho_0, \forall x \in X$$

a

$$\rho(x,y) = \rho(y,x), \forall x,y \in X. \textit{(symetrie)}$$

Když dále

$$\rho(x, y) = \rho_0 \text{ když a jen když } x = y \textit{ (totožnost)}$$

a $$\rho(x, z) \leq \rho(x, y) + \rho(y, z), \forall x,y,z \in X. \textit{(}\Delta\textit{ nerovnost)}$$

Prostor X, ve kterém metrika ρ definována, nazýváme metrickým prostorem.

Vzdálenost je hodnota určená podle metriky.

METRIKA PODOBNOSTI - PODOBNOST

Metrická míra podobnosti s na X je funkce $s: X \times X \rightarrow R$, kde R je množina reálných čísel, taková, že:

$$\exists s_0 \in R: -\infty < s(x,y) \leq s_0 < +\infty, \forall x,y \in X$$

$$s(x,x) = s_0, \forall x \in X$$

a

$$s(x,y) = s(y,x), \forall x,y \in X. \textit{(symetrie)}$$

Když dále

$$s(x,y) = s_0 \text{ když a jen když } x = y \textit{ (totožnost)}$$

a $$s(x,y).s(y,z) \leq [s(x,y) + s(y,z)].s(x,z), \forall x,y,z \in X.$$

MÍRY PODOBNOSTI VS. NEPODOBNOSTI

Vzdálenostní míry (míry nepodobnosti) mohou být transformovány na podobnostní míry různými transformacemi, např.

$$s_{ij} = 1/\rho_{ij}$$

$$s_{ij} = 1/(1+ \rho_{ij})$$

$$s_{ij} = c - \rho_{ij},\ c \geq \max \rho_{ij},\ \forall i,j$$

MÍRY PODOBNOSTI VS. NEPODOBNOSTI

Vzdálenostní míry (míry nepodobnosti) mohou být transformovány na podobnostní míry různými transformacemi, např.

$$s_{ij} = 1/\rho_{ij}$$

$$s(x,y).s(y,z) \leq [s(x,y) + s(y,z)].s(x,z), \forall x,y,z \in X$$

$$s_{ij} = 1/(1+ \rho_{ij})$$

$$s_{ij} = c - \rho_{ij}, c \geq \max \rho_{ij}, \forall i,j$$

MÍRY PODOBNOSTI VS. NEPODOBNOSTI

Vzdálenostní míry (míry nepodobnosti) mohou být transformovány na podobnostní míry různými transformacemi, např.

$$s_{ij} = 1/\rho_{ij}$$

$$s(x,y).s(y,z) \leq [s(x,y) + s(y,z)].s(x,z), \forall x,y,z \in X$$

$$s_{ij} = 1/(1+ \rho_{ij})$$

$$s_{ij} = c - \rho_{ij},\ c \geq \max \rho_{ij},\ \forall i,j$$

MÍRY PODOBNOSTI VS. NEPODOBNOSTI

Vzdálenostní míry (míry nepodobnosti) mohou být transformovány na podobnostní míry různými transformacemi, např.

$$s_{ij} = 1/\rho_{ij}$$

$$s(x,y).s(y,z) \leq [s(x,y) + s(y,z)].s(x,z), \forall x,y,z \in X$$

$$s_{ij} = 1/(1+ \rho_{ij})$$

$$s(x,y).s(y,z) \leq [s(x,y) + s(y,z) - s(x,y).s(y,z)].s(x,z), \forall x,y,z \in X$$

$$s_{ij} = c - \rho_{ij},\ c \geq \max \rho_{ij},\ \forall i,j$$

MÍRY PODOBNOSTI VS. NEPODOBNOSTI

Vzdálenostní míry (míry nepodobnosti) mohou být transformovány na podobnostní míry různými transformacemi, např.

$$s_{ij} = 1/\rho_{ij}$$

$$s(x,y).s(y,z) \leq [s(x,y) + s(y,z)].s(x,z), \forall x,y,z \in X$$

$$s_{ij} = 1/(1+ \rho_{ij})$$

$$s(x,y).s(y,z) \leq [s(x,y) + s(y,z) - s(x,y).s(y,z)].s(x,z), \forall x,y,z \in X$$

$$s_{ij} = c - \rho_{ij},\ c \geq \max \rho_{ij}, \forall i,j$$

$$s(x,z) \geq s(x,y) + s(y,z) - c$$

TYPY MĚR VZDÁLENOSTI (PODOBNOSTI)

- dle typu příznaků (numerické hodnoty, nominální či ordinální hodnoty, binární hodnoty);
- dle objektů, jejichž vztah hodnotíme – obrazy (vektory), množiny obrazů (vektorů), rozdělení
- deterministické (nepravděpodobnostní) vs. pravděpodobnostní míry

MÍRY VZDÁLENOSTI

obecné poznámky:

- ☑ výběr konkrétní metriky závisí na použití

  kritéria:

  - ➔ optimální výsledky (klasifikační chyby, ztráta, …)
  - ➔ výpočetní nároky
  - ➔ charakter rozložení dat

- ☑ obecně nelze doporučit vhodnou metriku pro určité standardní situace

METRIKY PRO URČENÍ VZDÁLENOSTI MEZI DVĚMA OBRAZY S KVANTITATIVNÍMI PŘÍZNAKY

EUKLIDOVA METRIKA

metrika zřejmě s nejnázornější geometrickou interpretaci

$$\rho_E(\mathbf{x}_1, \mathbf{x}_2) = \left[\sum_{i=1}^{n} (x_{1i} - x_{2i})^2 \right]^{1/2}$$

- ☑ geometrickým místem bodů s toutéž Euklidovou vzdáleností od daného bodu je hyperkoule (kruh ve dvourozměrném prostoru);
- ☑ dává větší důraz na větší rozdíly mezi souřadnicemi (žádoucí nebo nežádoucí? – volba i podle toho, jak chceme zdůrazňovat rozdíly mezi jednotlivými souřadnicemi)
- ☑ čtverec euklidovské vzdálenosti (lépe se počítá) je stále mírou nepodobnosti, ale není metrikou

EUKLIDOVA METRIKA

metrika zřejmě s nejnázornější geometrickou interpretaci

$$\rho_E(\mathbf{x}_1, \mathbf{x}_2) = \left[\sum_{i=1}^{n} (x_{1i} - x_{2i})^2 \right]^{1/2}$$

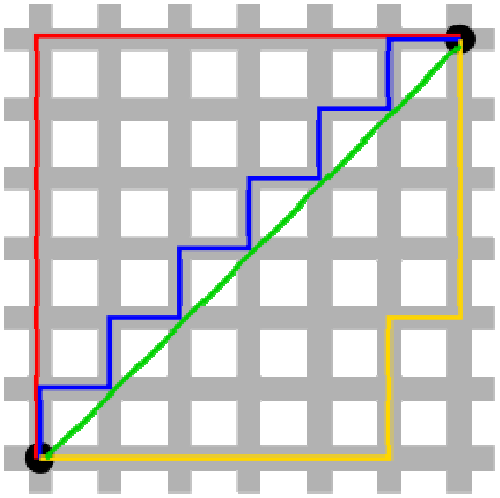
- **Sokalova metrika**

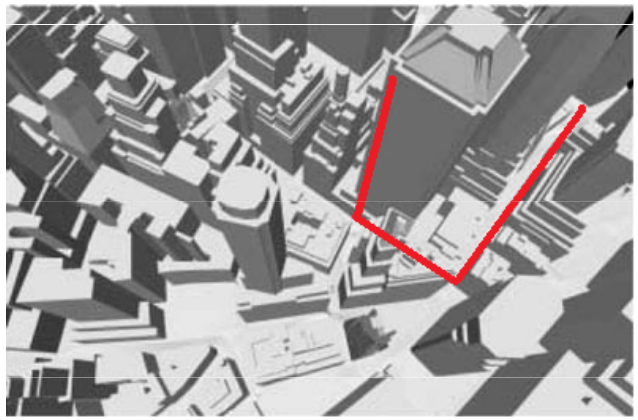
$$\rho_S(\mathbf{x}_1, \mathbf{x}_2) = \left\{ \frac{\rho_E^2(\mathbf{x}_1, \mathbf{x}_2)}{n} \right\}^{1/2}$$

HAMMINGOVA METRIKA

(metrika Manhattan, manhattanská metrika, city-block m., taxi driver m. – taxikářská metrika)

$$\rho_H(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{n} \left| x_{1i} - x_{2i} \right|$$





HAMMINGOVA METRIKA

(metrika Manhattan, manhattanská metrika, city-block m., taxi driver m. – taxikářská metrika)

$$\rho_H(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{n} \left| x_{1i} - x_{2i} \right|$$

- ☑ geometrickým místem bodů ve dvou rozměrném prostoru je čtverec postavený na špičku;
- ☑ nižší výpočetní nároky než E.m. $\Rightarrow$ použití v úlohách s vysokou výpočetní pracností

MINKOVSKÉHO METRIKA

$$\rho_M(\mathbf{x}_1, \mathbf{x}_2) = \left[\sum_{i=1}^{n} \left| x_{1i} - x_{2i} \right|^m \right]^{1/m}$$

- zobecnění Euklidovy a Hammingovy metriky;
- volba m záleží na míře důrazu – čím větší m, tím větší váha na velké rozdíly mezi příznaky,

  pro $m \to \infty$ metrika konverguje k Čebyševově metrice

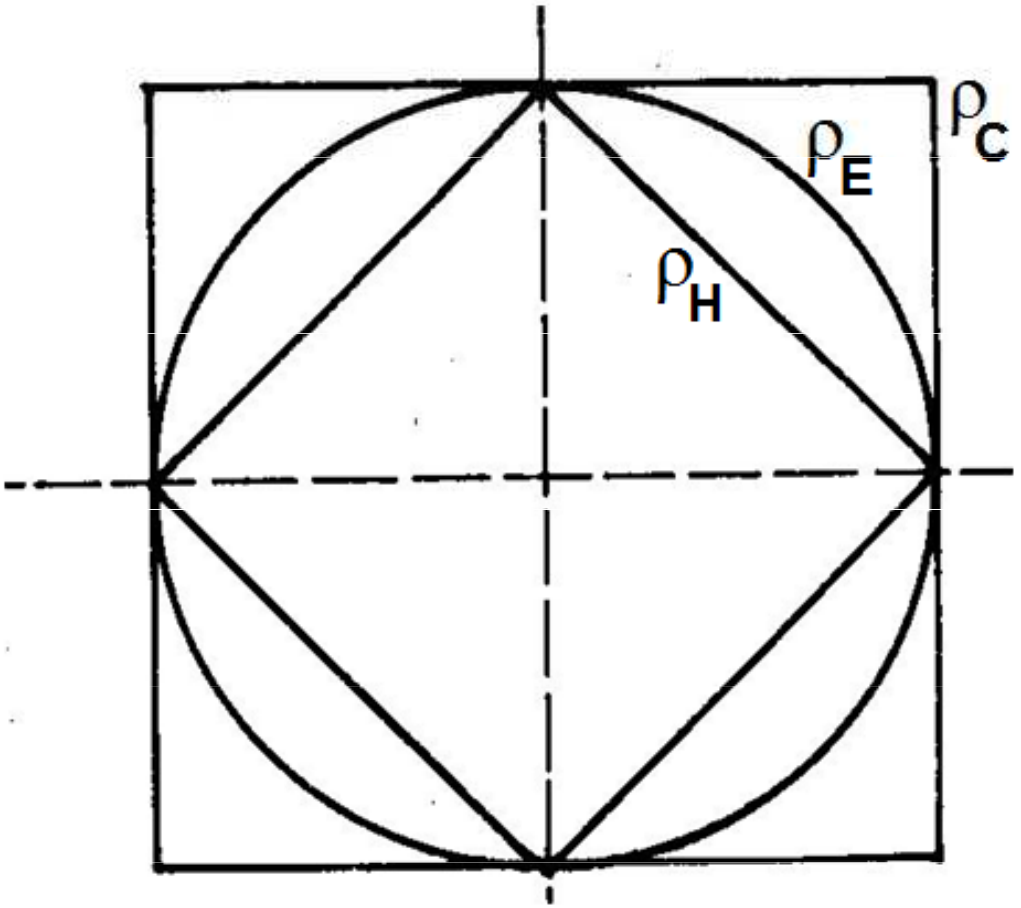
$$\rho_C(\mathbf{x}_1, \mathbf{x}_2) = \lim_{m \to \infty} \rho_M(\mathbf{x}_1, \mathbf{x}_2)$$

ČEBYŠEVOVA METRIKA

$$\rho_C(\mathbf{x}_1, \mathbf{x}_2) = \max_{\forall i} \left\{ \left| x_{1i} - x_{2i} \right| \right\}$$

- ☑ používá se ve výpočetně kriticky náročných případech, kdy je pracnost výpočtu dle euklidovsky orientovaných metrik nepřijatelná;
- ☑ geometrickým místem bodů s toutéž Čebyševovou vzdáleností od daného bodu je hyperkrychle (čtverec ve dvourozměrném prostoru)

SROVNÁNÍ GEOMETRICKÝCH MÍST



ČEBYŠEVOVA METRIKA

- pokud je třeba použít „euklidovskou“ metriku, ale s nižší výpočetní pracností, používá se v první řadě Hammingova nebo Čebyševova metrika;
- lepším přiblížením je kombinace obou metrik

$$\rho_A(\mathbf{x}_1, \mathbf{x}_2) = \max(2\rho_H / 3; \rho_C)$$

(ve dvourozměrném prostoru tvoří geometrické místo bodů o téže vzdálenosti osmiúhelník)

NEVÝHODY DOSUD UVEDENÝCH METRIK

- je nesmyslné vytvářet součet rozdílů veličin s různým fyzikálním rozměrem;
- při začlenění korelovaných veličin se zvyšuje jejich vliv na výslednou hodnotu;

NEVÝHODY DOSUD UVEDENÝCH METRIK

- je nesmyslné vytvářet součet rozdílů veličin s různým fyzikálním rozměrem;
- při začlenění korelovaných veličin se zvyšuje jejich vliv na výslednou hodnotu;

Jak si poradit ?

- transformace proměnných – vztažením k nějakému vyrovnávacímu faktoru – střední hodnotě, směrodatné odchylce, normě $\|\mathbf{x}\| = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}$, rozpětí $\Delta_i = \max_j x_{ij} - \min_j x_{ij}$, resp. standardizací

$$u_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \quad , i = 1, \ldots, n; j = 1, \ldots, K.$$

NEVÝHODY DOSUD UVEDENÝCH METRIK

- je nesmyslné vytvářet součet rozdílů veličin s různým fyzikálním rozměrem;
- při začlenění korelovaných veličin se zvyšuje jejich vliv na výslednou hodnotu;

Jak si poradit ?

- váhováním; např. Minkovského váhovaná metrika

$$\rho_{WM}(\mathbf{x}_1, \mathbf{x}_2) = \left[\sum_{i=1}^{n} a_i . \left| x_{1i} - x_{2i} \right|^m \right]^{1/m}$$

transformace pomocí váhových koeficientů je maticově

$$\mathbf{u} = \mathbf{C}^T . \mathbf{x},$$

kde koeficienty transformační matice **C** jsou dány

$$c_{ii} = a_i, \text{ pro } i = 1, \ldots, n;$$

$$c_{ij} = 0, \text{ pro } i \neq j.$$

NEVÝHODY DOSUD UVEDENÝCH METRIK

- ☑ je nesmyslné vytvářet součet rozdílů veličin s různým fyzikálním rozměrem;
- ☑ při začlenění korelovaných veličin se zvyšuje jejich vliv na výslednou hodnotu;

Jak si poradit ?

S takovým vyjádřením transformace příznakových proměnných je váhovaná Euklidova metrika definována vztahem

$$\rho_{WE}(\mathbf{x}_1, \mathbf{x}_2) = \left[(\mathbf{x}_1 - \mathbf{x}_2)^T . \mathbf{C} . \mathbf{C}^T . (\mathbf{x}_1 - \mathbf{x}_2)\right]^{1/2}.$$

Pokud jsou složky transformovaného obrazu dány lineární kombinací více složek původního obrazu, není ani matice $\mathbf{C}$, ani matice $\mathbf{C}.\mathbf{C}^T$ čistě diagonální.

KVADRATICKÁ VZDÁLENOST

$$\rho_Q^2(\mathbf{x}_1,\mathbf{x}_2)=(\mathbf{x}_1-\mathbf{x}_2)^T.\mathbf{Q}.(\mathbf{x}_1-\mathbf{x}_2)$$

- vhodný výběr matice Q je inverzní matice kovariance uvnitř množiny obrazů;
- pak se to (a po odmocnění) jmenuje Mahalanobisova metrika

$$\rho_{MA}(\mathbf{x}_1,\mathbf{x}_2)=\left[(\mathbf{x}_1-\mathbf{x}_2)^T.\mathbf{K}^{-1}.(\mathbf{x}_1-\mathbf{x}_2)\right]^{1/2}$$

CANBERRSKÁ METRIKA

- relativizovaná varianta Hammingovy metriky

$$\rho_{CA}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{n} \frac{|x_{1i} - x_{2i}|}{x_{1i} + x_{2i}}$$

- je vhodná pro proměnné s nezápornými hodnotami
  - pokud jsou obě hodnoty x_{1i} a x_{2i} nulové, potom předpokládáme, že hodnota zlomku je nulová;
  - je-li jenom jedna hodnota nulová, pak je zlomek roven jedné, nezávisle na velikosti druhé hodnoty;
  - někdy se nulové hodnoty nahrazují malým kladným číslem (menším, než nejmenší naměřené hodnoty);

CANBERRSKÁ METRIKA

- relativizovaná varianta Hammingovy metriky

$$\rho_{CA}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{n} \frac{|x_{1i} - x_{2i}|}{x_{1i} + x_{2i}}$$

- canberrská metrika je velice citlivá na malé změny souřadnic, pokud se oba obrazy nacházejí v blízkosti počátku souřadnicové soustavy. Naopak je méně citlivá na změny hodnot příznaků, pokud jsou tyto hodnoty velké.

PŘÍKLAD

Jsou dány dva vektory $\mathbf{x}_1 = (0{,}001; 0{,}001)^T$ a $\mathbf{x}_2 = (0{,}01; 0{,}01)^T$. Předpokládejme, že souřadnice prvního z vektorů se změní na $\mathbf{x}'_1$ $(0{,}002; 0{,}001)^T$. Jaká je Hammingova a canberrská vzdálenost v obou případech a jaká je relativní změna vzdáleností, vyvolaná uvedenou modifikací?

PŘÍKLAD - ŘEŠENÍ

$d_H(\mathbf{x}_1,\mathbf{x}_2) = |0{,}001-0{,}01| + |0{,}001-0{,}01| = 0{,}009 + 0{,}009 = 0{,}018;$

$d_H(\mathbf{x}'_1,\mathbf{x}_2) = |0{,}002-0{,}01| + |0{,}001-0{,}01| = 0{,}008 + 0{,}009 = 0{,}017;$

$$d_{CA}(\mathbf{x}_1,\mathbf{x}_2) = \frac{|0{,}001-0{,}01|}{0{,}001+0{,}01} + \frac{|0{,}001-0{,}01|}{0{,}001+0{,}01} = \frac{0{,}009}{0{,}011} + \frac{0{,}009}{0{,}011} = 0{,}8182+0{,}8182 = 1{,}6364$$

$$d_{CA}(\mathbf{x}'_1,\mathbf{x}_2) = \frac{|0{,}002-0{,}01|}{0{,}002+0{,}01} + \frac{|0{,}001-0{,}01|}{0{,}001+0{,}01} = \frac{0{,}008}{0{,}012} + \frac{0{,}009}{0{,}011} = 0{,}6667+0{,}8182 = 1{,}4849$$

Relativní změny vzdáleností, určující citlivost té které metriky, které jsou způsobeny změnou hodnoty první souřadnice, jsou

$$\Delta d_H = \frac{|d_H(\mathbf{x}_1,\mathbf{x}_2) - d_H(\mathbf{x}'_1,\mathbf{x}_2)|}{d_H(\mathbf{x}_1,\mathbf{x}_2)} = \frac{|0{,}018-0{,}017|}{0{,}018} = \frac{0{,}001}{0{,}018} = 0{,}056;$$

$$\Delta d_{CA} = \frac{|d_{CA}(\mathbf{x}_1,\mathbf{x}_2) - d_{CA}(\mathbf{x}'_1,\mathbf{x}_2)|}{d_{CA}(\mathbf{x}_1,\mathbf{x}_2)} = \frac{|1{,}6364-1{,}4849|}{1{,}6364} = 0{,}093$$

Ze získaných výsledků je zřejmé, že relativní změna vzdáleností je v případě canberrské metriky pro toto zadání o poznání větší.

PŘÍKLAD

Nyní mějme dány vektory $\mathbf{x}_1 = (1000; 1000)^T$ a $\mathbf{x}_2 = (100; 100)^T$ a předpokládejme, že dojde ke změně první souřadnice vektoru $\mathbf{x}_1$ na $\mathbf{x}'_1 = (1002; 1000)^T$. Jaká je Hammingova a canberrská vzdálenost pro tyto vektory a jaká je relativní změna vzdáleností, vyvolaná uvedenou modifikací?

PŘÍKLAD - ŘEŠENÍ

$d_H(\mathbf{x}_1,\mathbf{x}_2)$ = |1000-100| + |1000-100| = 900 + 900 = 1800;

$d_H(\mathbf{x}'_1,\mathbf{x}_2)$ = |1002-100| + |1000-100| = 902 + 900 = 1802;

$$d_{CA}(\mathbf{x}_1,\mathbf{x}_2)=\frac{|1000-100|}{1000+100}+\frac{|1000-100|}{1000+100}=\frac{900}{1100}+\frac{900}{1100}=0{,}8182+0{,}8182=1{,}6364$$

$$d_{CA}(\mathbf{x}'_1,\mathbf{x}_2)=\frac{|1002-100|}{1102+100}+\frac{|1000-100|}{1000+100}=\frac{902}{1102}+\frac{900}{1100}=0{,}8185+0{,}8182=1{,}6367$$

Relativní změny vzdáleností způsobených změnou hodnoty první souřadnice pak v tomto případě jsou

$$\Delta d_H=\frac{|d_H(\mathbf{x}_1,\mathbf{x}_2)-d_H(\mathbf{x}'_1,\mathbf{x}_2)|}{d_H(\mathbf{x}_1,\mathbf{x}_2)}=\frac{|1800-1802|}{1800}=\frac{2}{1800}=0{,}0011;$$

$$\Delta d_{CA}=\frac{|d_{CA}(\mathbf{x}_1,\mathbf{x}_2)-d_{CA}(\mathbf{x}'_1,\mathbf{x}_2)|}{d_{CA}(\mathbf{x}_1,\mathbf{x}_2)}=\frac{|1{,}6364-1{,}6367|}{1{,}6364}=\frac{0{,}0003}{1{,}6364}=0{,}00018$$

Jak je zřejmé, citlivost canberrské metriky je v tomto případě řádově menší.

NELINEÁRNÍ VZDÁLENOST

$$\rho_N(\mathbf{x}_1,\mathbf{x}_2)=\begin{cases}0 & \text{když } \rho_E(\mathbf{x}_1,\mathbf{x}_2)<D\\ H & \text{když } \rho_E(\mathbf{x}_1,\mathbf{x}_2)\geq D\end{cases}$$

kde D je prahová hodnota a H je nějaká konstanta.

Uvádí se, že dobrý výběr hodnot H a D by měl splňovat vztah

$$H=\frac{\Gamma(n/2)}{D^n\sqrt{\pi^n}}$$

když D splňuje nestrannost a konzistenční podmínku Parzenova odhadu, především $D^nN\to\infty$ a $D\to 0$, když $N\to\infty$ (N je počet obrazů v množině)

METRIKY PRO URČENÍ PODOBNOSTI MEZI DVĚMA OBRAZY S KVANTITATIVNÍMI PŘÍZNAKY

SKALÁRNÍ SOUČIN

$$\sigma_{ss}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^T . \mathbf{x}_2 = \sum_{i=1}^{n} x_{1i} x_{2i}$$

Ve většině případů je skalární součin jako míra podobnosti použit pro vektory $\mathbf{x}_1$ a $\mathbf{x}_2$ o stejné délce, např. a. V těch případech jsou horní, resp. dolní mez skalárního součinu a^2, resp. $-a^2$ a hodnoty skalárního součinu v tom případě závisí výhradně na úhlu, který oba vektory svírají. Hodnoty a^2 nabývá, pokud oba vektory svírají nulový úhel, hodnoty $-a^2$, pokud úhel mezi nimi je 180° a nulové hodnoty, pokud jsou oba vektory na sebe kolmé. skalární součin je tedy invariantní vůči rotaci (jejich absolutní orientace není podstatná, důležitý je pouze úhel mezi nimi), nikoliv však vůči lineární transformaci (závisí na délce vektorů).

Ze skalárního součinu vektorů o délce a je možné odvodit i metriku vzdálenosti podle vztahu

$$\rho_{ss}(\mathbf{x}_1, \mathbf{x}_2) = a^2 - \sigma_{ss}(\mathbf{x}_1, \mathbf{x}_2)$$

METRIKA KOSINOVÉ PODOBNOSTI

$$\sigma_{cos}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathbf{x}_1^T . \mathbf{x}_2}{\|\mathbf{x}_1\| . \|\mathbf{x}_2\|}$$

kde $\|x_i\|$ je norma (délka) vektoru $\mathbf{x}_i$.

= skalární součin vektorů o jednotkové délce

vhodná v případě, pokud je informativní pouze relativní hodnota příznaků;

hodnoty $\sigma_{cos}(\mathbf{x}_1, \mathbf{x}_2)$ jsou rovny kosinu úhlu mezi oběma vektory.

PEARSONŮV KORELAČNÍ KOEFICIENT

$$\sigma_{PC}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathbf{x}_{d1}^T . \mathbf{x}_{d2}}{\|\mathbf{x}_{d1}\| . \|\mathbf{x}_{d2}\|}$$

kde $\mathbf{x}_{di} = (x_{i1} - \bar{x}_i, x_{i2} - \bar{x}_i, \ldots, x_{in} - \bar{x}_i)^T$, x_{ij} představují j-tou souřadnici vektoru $\mathbf{x}_i$ a je střední hodnota určená ze souřadnic vektoru $\mathbf{x}_i$ ($\bar{x}_i = \sum_{j=1}^{n} x_{ij} / n$). Vektory $\mathbf{x}_{di}$ se nazývají diferenční vektory. Podobně jako v případě kosinové podobnosti, nabývá Pearsonův korelační koeficient hodnot z intervalu $\langle -1;1 \rangle$, rozdíl vůči kosinové míře podobnosti je ten, že určuje vztah nikoliv vektorů $\mathbf{x}_1$ a $\mathbf{x}_2$, nýbrž jejich diferenčních variant.

PEARSONŮV KORELAČNÍ KOEFICIENT

I z hodnot Pearsonova korelačního koeficientu lze určit vzdálenost obou vektorů pomocí metriky

$$\rho_{PC}(\mathbf{x}_1, \mathbf{x}_2) = \frac{1 - \sigma_{PC}(\mathbf{x}_1, \mathbf{x}_2)}{2}$$

její hodnoty se, díky dělení dvěma, vyskytují v intervalu $\langle 0;1 \rangle$.

Tato metrika se používá např. při analýze dat genové exprese.

TANIMOTOVA METRIKA PODOBNOSTI

$$\sigma_T(\mathbf{x}_1,\mathbf{x}_2)=\frac{\mathbf{x}_1^T\mathbf{x}_2}{\|\mathbf{x}_1\|^2+\|\mathbf{x}_2\|^2-\mathbf{x}_1^T\mathbf{x}_2}$$

Přičteme-li a odečteme-li ve jmenovateli výraz $\mathbf{x}_1{}^T\mathbf{x}_2$ a podělíme-li čitatele i jmenovatele zlomku toutéž hodnotou, dostaneme

$$\sigma_T(\mathbf{x}_1,\mathbf{x}_2)=\frac{1}{1+\frac{(\mathbf{x}_1-\mathbf{x}_2)^T(\mathbf{x}_1-\mathbf{x}_2)}{\mathbf{x}_1^T\mathbf{x}_2}}$$

Tanimotova podobnost vektorů $\mathbf{x}_1$ a $\mathbf{x}_2$ je nepřímo úměrná kvadrátu Euklidovy vzdálenosti vektorů $\mathbf{x}_1$ a $\mathbf{x}_2$ vztažené k jejich skalárnímu součinu. Pokud skalární součin považujeme za míru korelace obou vektorů, můžeme formulovat výše uvedený vztah tak, že $\sigma_T(\mathbf{x}_1, \mathbf{x}_2)$ je nepřímo úměrná kvadrátu Euklidovy vzdálenosti podělené velikostí jejich korelace, což znamená, že je korelaci, jako míře podobnosti přímo úměrná.

„BEZEJMENNÁ“ METRIKA PODOBNOSTI

$$\sigma_C(\mathbf{x}_1,\mathbf{x}_2)=1-\frac{\rho_E(\mathbf{x}_1,\mathbf{x}_2)}{\|\mathbf{x}_1\|+\|\mathbf{x}_2\|}$$

Vzdálenost podle metriky je rovna jedné,

když $\mathbf{x}_1 = \mathbf{x}_2$

a svého minima (tj. $\sigma_C(\mathbf{x}_1,\mathbf{x}_2)= -1$) nabývá,

když $\mathbf{x}_1 = -\mathbf{x}_2$.

Příprava nových učebních materiálů
oboru Matematická biologie

je podporována projektem ESF

č. CZ.1.07/2.2.00/28.0043

„INTERDISCIPLINÁRNÍ ROZVOJ STUDIJNÍHO OBORU MATEMATICKÁ BIOLOGIE“

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ