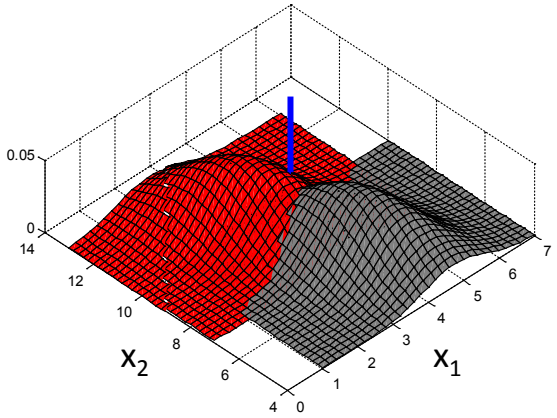
Analýza a klasifikace dat – přednáška 4

RNDr. Eva Janoušová

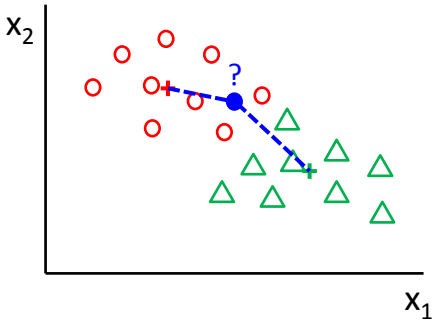
Podzim 2014

Typy klasifikátorů – podle principu klasifikace

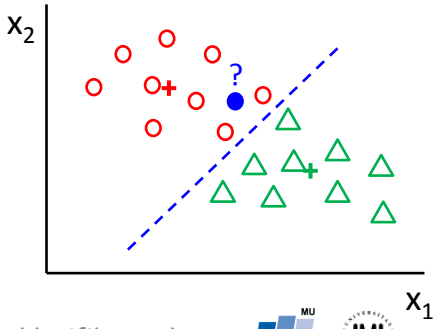
- **klasifikace pomocí diskriminačních funkcí:**
  - diskriminační funkce určují míru příslušnosti k dané klasifikační třídě
  - pro danou třídu má daná diskriminační funkce nejvyšší hodnotu



- **klasifikace pomocí vzdálenosti od etalonů klasif. tříd:**
  - etalon = reprezentativní objekt(y) klasifikační třídy
  - počet etalonů klasif. třídy různý – od jednoho vzorku (např. centroidu) po úplný výčet všech objektů dané třídy (např. u klasif. pomocí metody průměrné vazby)



- **klasifikace pomocí hranic v obrazovém prostoru:**
  - stanovení hranic (hraničních ploch) oddělujících klasifikační třídy



Typy metrik a konkrétní příklady

MEZI DVĚMA OBJEKTY

Metriky pro určení vzdálenosti mezi 2 objekty s kvantitativními proměnnými

Euklidova m., Hammingova (manhattanská) m., Minkovského m., Čebyševova m., Mahalanobisova m., Canberrská m.

Metriky pro určení podobnosti 2 objektů s kvantitativními proměnnými

Skalární součin, m. kosinové podobnosti, Pearsonův korelační koeficient, Tanimotova m.

Metriky pro určení vzdálenosti mezi 2 objekty s kvalitativními proměnnými

Hammingova m.

Metriky pro určení podobnosti 2 objektů s kvalitativními proměnnými

Tanimotova m., Jaccardův-Tanimotův a.k., Russelův-Raovův a.k., Sokalův-Michenerův a.k., Dicův k., Rogersův-Tanimotův k., Hamanův k.

MEZI DVĚMA MNOŽINAMI OBJEKTŮ

Deterministické metriky pro určení vzdálenosti mezi 2 množinami objektů

Metoda nejbližšího souseda, *k* nejbližších sousedů, nejvzdálenějšího souseda, průměrné vazby, Wardova metoda

Metriky pro určení vzdálenosti mezi 2 množinami objektů používající jejich pravděpodobnostní charakteristiky

Chernoffova m., Bhattacharyyova m. atd.

Skalární součin

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

Většinou pro vektory $\mathbf{x}_1$ a $\mathbf{x}_2$ o stejné délce (např. a); záleží na úhlu, který svírají:

úhel 0°
$S_{ss} = a^2$

úhel 90°
$S_{ss} = 0$

úhel 180°
$S_{ss} = -a^2$

skalární součin invariantní vůči rotaci – absolutní orientace nepodstatná, důležitý pouze úhel

skalární součin není invariantní vůči lineární transformaci (tzn. závisí na délce vektorů)

odvození metriky vzdálenosti:

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

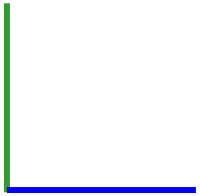
Metrika kosinové podobnosti

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

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

úhel 0°
$S_{cos} = 1$



úhel 90°
$S_{cos} = 0$

úhel 180°
$S_{cos} = -1$

Pearsonův korelační koeficient

Pearsonův korelační koeficient

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

Metrika kosinové podobnosti

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

kde $\mathbf{x}_{di} = \left(x_{i1} - \bar{x}_i, x_{i2} - \bar{x}_i, \ldots, x_{ip} - \bar{x}_i\right)^T$

$\mathbf{x}_{di}$ jsou tzv. diferenční vektory

také nabývá hodnot z intervalu ⟨-1;1⟩

odvození metriky vzdálenosti:

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

→ hodnoty se (díky dělení dvěma) vyskytují v intervalu ⟨0;1⟩

→ používá se např. při analýze dat genové exprese

Tanimotova metrika podobnosti

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

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

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

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 $S_T(\mathbf{x}_1,\mathbf{x}_2)$ je nepřímo úměrná kvadrátu Euklidovy vzdálenosti podělené velikostí jejich korelace, což znamená, že je korelaci, jako míře podobnosti přímo úměrná.

„Bezejmenná“ metrika podobnosti

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

Vzdálenost podle metriky je rovna jedné,

když $\mathbf{x}_1 = \mathbf{x}_2$

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

když $\mathbf{x}_1 = -\mathbf{x}_2$.

Typy metrik a konkrétní příklady

MEZI DVĚMA OBJEKTY

Metriky pro určení vzdálenosti mezi 2 objekty s kvantitativními proměnnými

Euklidova m., Hammingova (manhattanská) m., Minkovského m., Čebyševova m., Mahalanobisova m., Canberrská m.

Metriky pro určení podobnosti 2 objektů s kvantitativními proměnnými

Skalární součin, m. kosinové podobnosti, Pearsonův korelační koeficient, Tanimotova m.

Metriky pro určení vzdálenosti mezi 2 objekty s kvalitativními proměnnými

Hammingova m.

Metriky pro určení podobnosti 2 objektů s kvalitativními proměnnými

Tanimotova m., Jaccardův-Tanimotův a.k., Russelův-Raovův a.k., Sokalův-Michenerův a.k., Dicův k., Rogersův-Tanimotův k., Hamanův k.

MEZI DVĚMA MNOŽINAMI OBJEKTŮ

Deterministické metriky pro určení vzdálenosti mezi 2 množinami objektů

Metoda nejbližšího souseda, *k* nejbližších sousedů, nejvzdálenějšího souseda, průměrné vazby, Wardova metoda

Metriky pro určení vzdálenosti mezi 2 množinami objektů používající jejich pravděpodobnostní charakteristiky

Chernoffova m., Bhattacharyyova m. atd.

Příklad

Předpokládejme, že množina F obsahuje symboly {0, 1, 2}, tj. $k = 3$ a vektory **x** a **y** jsou:

$\mathbf{x} = (0, 1, 2, 1, 2, 1)^T$ a $\mathbf{y} = (1, 0, 2, 1, 0, 1)^T$, $p = 6$.

Spočtěte vzdálenost obou vektorů.

Kontingenční matice **A**(**x**,**y**) je:

$$\mathbf{A}(\mathbf{x},\mathbf{y}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Součet hodnot všech prvků matice **A**(**x**,**y**) je roven délce p obou vektorů, tj. v našem případě:

$$\sum_{i=0}^{2}\sum_{j=0}^{2} a_{ij} = 6$$

Hammingova metrika vzdálenosti

$$D_{HQ}(\mathbf{x},\mathbf{y}) = \sum_{i=0}^{k-1} \sum_{\substack{j=0 \\ i \neq j}}^{k-1} a_{ij}$$

- definována počtem pozic, v nichž se oba vektory liší
- tzn. je dána součtem všech prvků matice **A**, které leží mimo hlavní diagonálu.

Příklad:

$\mathbf{x} = (0, 1, 2, 1, 2, 1)^T$
$\mathbf{y} = (1, 0, 2, 1, 0, 1)^T$

↓

liší se ve 3 souřadnicích

↓

$d_{HQ}(\mathbf{x},\mathbf{y}) = 3$

$$\mathbf{A}(\mathbf{x},\mathbf{y}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

↓

3 prvky mimo diagonálu

↓

$d_{HQ}(\mathbf{x},\mathbf{y}) = 3$

Hammingova metrika vzdálenosti

- pro *k* = 2, kdy jsou hodnoty obou vektorů binární, se definiční vztah Hammingovy vzdálenosti transformuje na:

$$D_{HQB}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p}(x_i + y_i - 2x_i y_i)$$

kde třetí člen v závorce kompenzuje případ, kdy jsou hodnoty x_i i y_i rovny jedné a součet prvních členů v závorce je tím pádem roven dvěma, nicméně nastává shoda hodnot, která k celkové vzdálenosti nemůže přispět.

- protože x_i a y_i nabývají hodnot pouze 0 a 1, můžeme také psát:

$$D_{HQB}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p}(x_i^2 + y_i^2 - 2x_i y_i) = \sum_{i=1}^{p}(x_i - y_i)^2$$

- díky speciálnímu případu hodnot x_i a y_i je možná i nejjednodušší forma:

$$D_{HQB}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p}\left|x_i - y_i\right|$$

Hammingova metrika vzdálenosti – příklad 2

Určete Hammingovu vzdálenost binárních vektorů
x = $(0, 1, 1, 0, 1)^T$ a
y = $(1, 0, 0, 0, 1)^T$.

Podle definičního principu (tzn. počet pozic, ve kterých se oba vektory liší):

$$d_{HQB}(\mathbf{x},\mathbf{y}) = 3$$

Dle jiného vztahu: $d_{HQB}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p}(x_i + y_i - 2x_i y_i) =$

$$= (0+1-2\cdot 0\cdot 1) + (1+0-2\cdot 1\cdot 0) + (1+0-2\cdot 1\cdot 0) + (0+0-2\cdot 0\cdot 0) + (1+1-2\cdot 1\cdot 1) = 3$$

Dle dalšího vztahu: $d_{HQB}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p}(x_i - y_i)^2 =$

$$= (0-1)^2 + (1-0)^2 + (1-0)^2 + (0-0)^2 + (1-1)^2 = 1+1+1+0+0 = 3$$

Dle posledního vztahu: $d_{HQB}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p}|x_i - y_i| =$

$$= |0-1| + |1-0| + |1-0| + |0-0| + |1-1| =$$
$$= 1+1+1+0+0 = 3$$

Hammingova metrika vzdálenosti

V případě bipolárních vektorů, kdy jednotlivé složky vektorů nabývají hodnot +1 a -1, je Hammingova vzdálenost určena vztahem:

$$D_{HQP}(\mathbf{x},\mathbf{y}) = \frac{\left(p - \sum_{i=1}^{p} x_i y_i\right)}{2}$$

Příklad 3:

Určete Hammingovu vzdálenost bipolárních vektorů
x = $(1, 1, 1, -1, 1)^T$ a
y = $(1, -1, 1, -1, -1)^T$.

Podle definičního principu (tzn. počet pozic, ve kterých se liší): $d_{HQP}(\mathbf{x},\mathbf{y}) = 2$

Z kontingenční matice (součet prvků mimo hlavní diagonálu): $\mathbf{A}(\mathbf{x},\mathbf{y}) = \begin{bmatrix} 2 & 2 \\ 0 & 1 \end{bmatrix}$

Pomocí vztahu:

$$d_{HQP}(\mathbf{x},\mathbf{y}) = \frac{5 - \left((1\cdot 1) + (1\cdot(-1)) + (1\cdot 1) + ((-1)\cdot(-1)) + (1\cdot(-1))\right)}{2} = \frac{5 - (1-1+1+1-1)}{2} = \frac{5-1}{2} = 2$$

Typy metrik a konkrétní příklady

MEZI DVĚMA OBJEKTY

Metriky pro určení vzdálenosti mezi 2 objekty s kvantitativními proměnnými

Euklidova m., Hammingova (manhattanská) m., Minkovského m., Čebyševova m., Mahalanobisova m., Canberrská m.

Metriky pro určení podobnosti 2 objektů s kvantitativními proměnnými

Skalární součin, m. kosinové podobnosti, Pearsonův korelační koeficient, Tanimotova m.

Metriky pro určení vzdálenosti mezi 2 objekty s kvalitativními proměnnými

Hammingova m.

Metriky pro určení podobnosti 2 objektů s kvalitativními proměnnými

Tanimotova m., Jaccardův-Tanimotův a.k., Russelův-Raovův a.k., Sokalův-Michenerův a.k., Dicův k., Rogersův-Tanimotův k., Hamanův k.

MEZI DVĚMA MNOŽINAMI OBJEKTŮ

Deterministické metriky pro určení vzdálenosti mezi 2 množinami objektů

Metoda nejbližšího souseda, *k* nejbližších sousedů, nejvzdálenějšího souseda, průměrné vazby, Wardova metoda

Metriky pro určení vzdálenosti mezi 2 množinami objektů používající jejich pravděpodobnostní charakteristiky

Chernoffova m., Bhattacharyyova m. atd.

Metriky pro určení podobnosti 2 objektů s kvalitativními prom.

1. případy obecné
2. případy s dichotomickými příznaky, pro které je definována celá řady tzv. ***asociačních koeficientů***.

   (Asociační koeficienty až na výjimky nabývají hodnot z intervalu $\langle 0, 1 \rangle$, hodnoty 1 v případě shody vektorů, 0 pro případ nepodobnosti.)

Obecné metriky – Hammingova metrika podobnosti

$$S_{HQ}(\mathbf{x},\mathbf{y}) = p - D_{HQ}(\mathbf{x},\mathbf{y})$$

Příklad:

$\mathbf{x} = (0, 1, 2, 1, 2, 1)^T$
$\mathbf{y} = (1, 0, 2, 1, 0, 1)^T$

↓

liší se ve 3 souřadnicích

↓

$d_{HQ}(\mathbf{x},\mathbf{y}) = 3$

↓

shoda ve 3 souřadnicích

↓

$s_{HQ}(\mathbf{x},\mathbf{y}) = 6 - 3 = 3$

$$\mathbf{A}(\mathbf{x},\mathbf{y}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

↓

3 prvky mimo diagonálu

↓

$d_{HQ}(\mathbf{x},\mathbf{y}) = 3$

↓

součet prvků na diagonále roven 3

↓

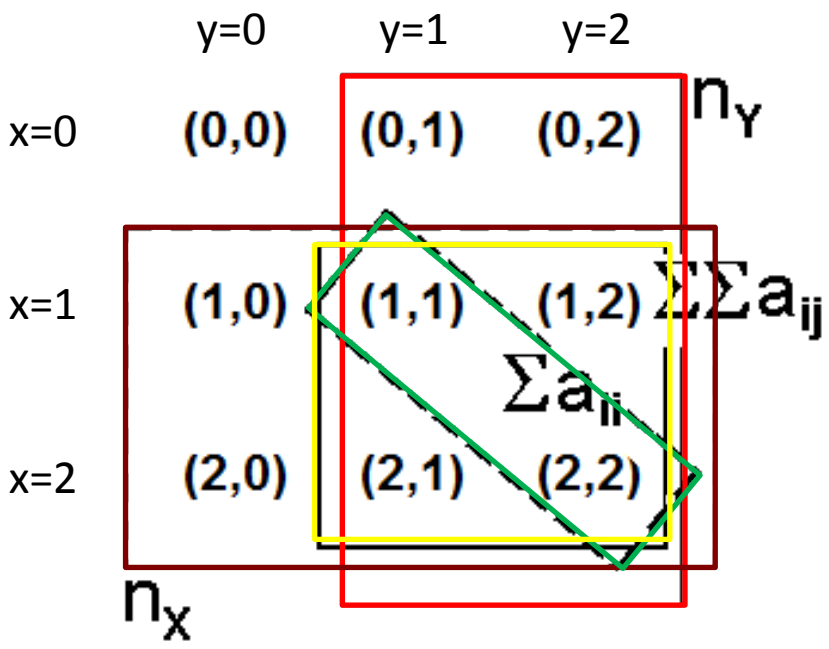
$s_{HQ}(\mathbf{x},\mathbf{y}) = 6 - 3 = 3$

Obecné metriky – Tanimotova metrika

$$S_{TQ}(\mathbf{x},\mathbf{y}) = \frac{n_{X\cap Y}}{n_X + n_Y - n_{X\cap Y}} = \frac{\sum_{i=1}^{k-1} a_{ii}}{n_x + n_y - \sum_{i=1}^{k-1}\sum_{j=1}^{k-1} a_{ij}}$$

$$n_x = \sum_{i=1}^{k-1}\sum_{j=0}^{k-1} a_{ij} \qquad n_y = \sum_{i=0}^{k-1}\sum_{j=1}^{k-1} a_{ij}$$

Pro výpočet Tanimotovy podobnosti dvou vektorů s kvalitativními příznaky jsou použity všechny páry složek srovnávaných vektorů, kromě těch, jejichž hodnoty jsou obě nulové.



Obecné metriky – Tanimotova metrika – příklad

Určete hodnoty Tanimotových podobností $s_{TQ}(\mathbf{x},\mathbf{x})$, $s_{TQ}(\mathbf{x},\mathbf{y})$ a $s_{TQ}(\mathbf{x},\mathbf{z})$, když:
$\mathbf{x} = (0, 1, 2, 1, 2, 1)^T$ a
$\mathbf{y} = (1, 0, 2, 1, 0, 1)^T$ a
$\mathbf{z} = (2, 0, 0, 0, 0, 2)^T$.
Ze zadání je množina symbolů F = {0, 1, 2}, $k = 3$, $p = 6$.

Kontingenční tabulky jsou:

$$\mathbf{A}(\mathbf{x},\mathbf{x}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix} \qquad \mathbf{A}(\mathbf{x},\mathbf{y}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix} \qquad \mathbf{A}(\mathbf{x},\mathbf{z}) = \begin{bmatrix} 0 & 0 & 1 \\ 2 & 0 & 1 \\ 2 & 0 & 0 \end{bmatrix}$$

$$s_{TQ}(\mathbf{x},\mathbf{x}) = \frac{5}{5+5-5} = 1 \qquad s_{TQ}(\mathbf{x},\mathbf{y}) = \frac{3}{5+4-3} = 0{,}5 \qquad s_{TQ}(\mathbf{x},\mathbf{z}) = \frac{0}{5+2-1} = 0$$

Další obecné metriky

- definovány pomocí různých prvků kontingenční matice **A**(**x**,**y**)

- některé z nich používají pouze počet shodných pozic v obou vektorech (ovšem s nenulovými hodnotami):

$$S_1(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{k-1} a_{ii}}{p} \qquad S_2(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{k-1} a_{ii}}{p - a_{00}}$$

- některé z nich používají i shodu s nulovými hodnotami:

$$S_3(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=0}^{k-1} a_{ii}}{p}$$

Asociační koeficienty

		$\mathbf{x}_j$	
		false/0	true/1
$\mathbf{x}_i$	false/0	D	C
	true/1	B	A

A - u obou objektů sledovaný jev nastal (obě odpovídající si proměnné mají hodnotu true, resp.1) – ***pozitivní shoda***;

B - u objektu $\mathbf{x}_i$ jev nastal (x_{ik} = true), zatímco u objektu $\mathbf{x}_j$ nikoliv (x_{jk} = false, resp.0);

C - u objektu $\mathbf{x}_i$ jev nenastal (x_{ik} = false), zatímco u objektu $\mathbf{x}_j$ ano (x_{jk} = true);

D - sledovaný jev nenastal ani u jednoho z objektů (obě odpovídající si proměnné mají hodnotu false, resp. 0) – ***negativní shoda***.

Při výpočtu podobnosti dvou objektů sledujeme, kolikrát pro všechny souřadnice obou vektorů $\mathbf{x}_j$ a $\mathbf{x}_j$ nastaly případy shody či neshody:

- A+D určuje celkový počet shod
- B+C celkový počet neshod
- A+B+C+D = *p* (tj. celk. počet souřadnic obou vektorů – tzn. počet proměnných)

Jaccardův – Tanimotův asociační koeficient

$$S_{JT}(\mathbf{x},\mathbf{y}) = \frac{A}{A+B+C}$$

		$\mathbf{x}_j$	
		false/0	true/1
$\mathbf{x}_i$	false/0	D	C
	true/1	B	A

což je díky zjednodušení i dichotomická varianta metriky podle vztahu:

$$S_{TQ}(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{k-1} a_{ii}}{n_x + n_y - \sum_{i=1}^{k-1}\sum_{i=1}^{k-1} a_{ij}}$$

Tento vztah se dominantně používá v ekologických studiích.

Další asociační koeficienty I

		x_j	
		false/0	true/1
x_i	false/0	D	C
	true/1	B	A

Russelův – Raoův asociační koeficient

$$S_{RR}(\mathbf{x},\mathbf{y}) = \frac{A}{A+B+C+D}$$

dichotomická varianta metriky:

$$S_1(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{k-1} a_{ii}}{p}$$

Sokalův – Michenerův asociační koeficient

$$S_{SM}(\mathbf{x},\mathbf{y}) = \frac{A+D}{A+B+C+D}$$

dichotomická varianta metriky:

$$S_3(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=0}^{k-1} a_{ii}}{p}$$

Další asociační koeficienty II

		$\mathbf{x}_j$	
		false/0	true/1
$\mathbf{x}_i$	false/0	D	C
	true/1	B	A

Diceův (Czekanowského) asociační koeficient

$$S_{DC}(\mathbf{x},\mathbf{y}) = \frac{2A}{2A+B+C} = \frac{2A}{(A+B)+(A+C)}$$

V případě Jaccardova a Diceova koeficientu pokud nastane úplná negativní shoda (tzn. A = B = C =0), pak často: $S_{JT}(\mathbf{x},\mathbf{y}) = S_{DC}(\mathbf{x},\mathbf{y}) = 1$.

Rogersův – Tanimotův asociační koeficient

$$S_{RT}(\mathbf{x},\mathbf{y}) = \frac{A+D}{A+D+2\cdot(B+C)} = \frac{A+D}{(B+C)+(A+B+C+D)}$$

Hamanův asociační koeficient

$$S_{HA}(\mathbf{x},\mathbf{y}) = \frac{A+D-(B+C)}{A+B+C+D}$$

nabývá na rozdíl od všech dříve uvedených koeficientů hodnot z intervalu $\langle -1, 1\rangle$. Hodnoty -1, pokud se příznaky pouze neshodují; hodnoty 0, když je počet shod a neshod v rovnováze; +1 v případě úplné shody všech příznaků

Asociační koeficienty – poznámka

		x_j	
		false/0	true/1
x_i	false/0	D	C
	true/1	B	A

Na základě četností A až D lze pro případ binárních příznaků vytvářet i zajímavé vztahy pro již dříve uvedené míry:

Hammingova metrika

$$D_H(\mathbf{x},\mathbf{y}) = B + C$$

Euklidova metrika

$$D_H(\mathbf{x},\mathbf{y}) = \sqrt{B + C}$$

Pearsonův korelační koeficient

$$S_{PC}(\mathbf{x},\mathbf{y}) = \frac{A \cdot D - B \cdot C}{\sqrt{(A+B)\cdot(C+D)\cdot(A+C)\cdot(B+D)}}$$

Výpočet vzdáleností z asociačních koeficientů

Z asociačních koeficientů, které vyjadřují míru podobnosti, lze jednoduše odvodit i míry nepodobnosti (vzdálenosti) pomocí:

$$D_X(\mathbf{x},\mathbf{y}) = 1 - S_X(\mathbf{x},\mathbf{y})$$

Typy metrik a konkrétní příklady

MEZI DVĚMA OBJEKTY

Metriky pro určení vzdálenosti mezi 2 objekty s kvantitativními proměnnými

Euklidova m., Hammingova (manhattanská) m., Minkovského m., Čebyševova m., Mahalanobisova m., Canberrská m.

Metriky pro určení podobnosti 2 objektů s kvantitativními proměnnými

Skalární součin, m. kosinové podobnosti, Pearsonův korelační koeficient, Tanimotova m.

Metriky pro určení vzdálenosti mezi 2 objekty s kvalitativními proměnnými

Hammingova m.

Metriky pro určení podobnosti 2 objektů s kvalitativními proměnnými

Tanimotova m., Jaccardův-Tanimotův a.k., Russelův-Raovův a.k., Sokalův-Michenerův a.k., Dicův k., Rogersův-Tanimotův k., Hamanův k.

MEZI DVĚMA MNOŽINAMI OBJEKTŮ

Deterministické metriky pro určení vzdálenosti mezi 2 množinami objektů

Metoda nejbližšího souseda, *k* nejbližších sousedů, nejvzdálenějšího souseda, průměrné vazby, Wardova metoda

Metriky pro určení vzdálenosti mezi 2 množinami objektů používající jejich pravděpodobnostní charakteristiky

Chernoffova m., Bhattacharyyova m. atd.

Metriky založené na pstních charakteristikách

Klasifikační třídy (množiny objektů se společnými charakteristikami) nemusí být definovány jen výčtem objektů, ale i vymezením obecnějších vlastností:

- definicí hranic oddělujících část obrazového prostoru náležející dané klasifikační třídě
- diskriminační funkcí
- **pravděpodobnostními charakteristikami výskytu obrazů v dané třídě**
- atd.

Metriky založené na pstních charakteristikách

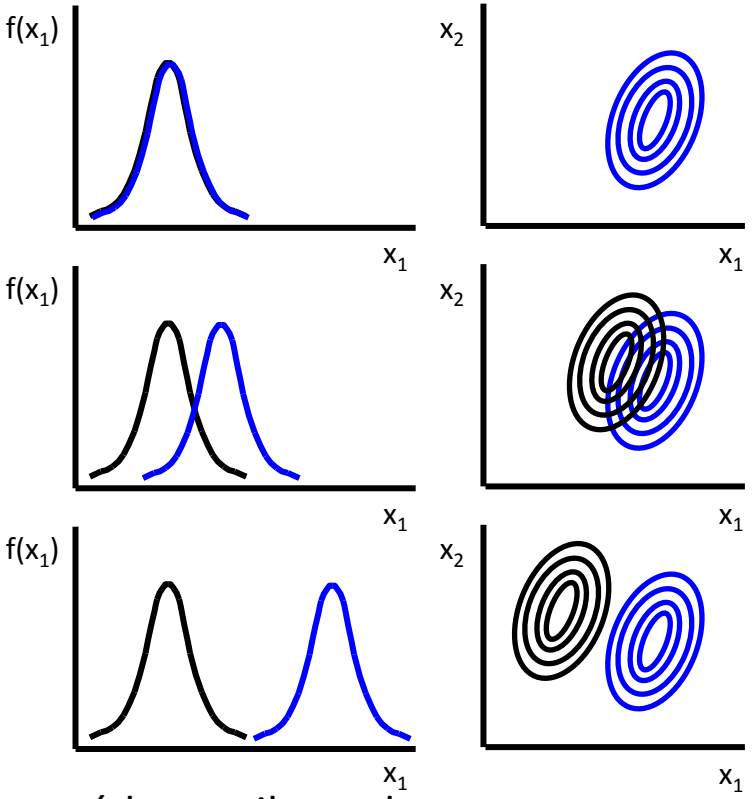
Základní myšlenkou je využití pravděpodobnosti způsobené chyby. Čím více se hustoty pravděpodobnosti výskytu obrazů **x** v jednotlivých množinách překrývají, tím je větší pravděpodobnost chyby.

Tzn. tyto metriky splňují následující vlastnosti:

1. J = 0, pokud jsou hustoty pravděpodobnosti obou množin identické, tj. když
$p(\mathbf{x}|\omega_1) = p(\mathbf{x}|\omega_2)$

2. $J \geq 0$

3. J nabývá maxima, pokud jsou obě množiny disjunktní, tj. když $\int_{-\infty}^{\infty} p(\mathbf{x}|\omega_1) \cdot p(\mathbf{x}|\omega_2) d\mathbf{x} = 0$



(Jak vidíme, není mezi vlastnostmi pravděpodobnostních metrik uvedena trojúhelníková nerovnost, jejíž splnění by se zajišťovalo velmi obtížně.)

Typy metrik a konkrétní příklady

MEZI DVĚMA OBJEKTY

Metriky pro určení vzdálenosti mezi 2 objekty s kvantitativními proměnnými

Euklidova m., Hammingova (manhattanská) m., Minkovského m., Čebyševova m., Mahalanobisova m., Canberrská m.

Metriky pro určení podobnosti 2 objektů s kvantitativními proměnnými

Skalární součin, m. kosinové podobnosti, Pearsonův korelační koeficient, Tanimotova m.

Metriky pro určení vzdálenosti mezi 2 objekty s kvalitativními proměnnými

Hammingova m.

Metriky pro určení podobnosti 2 objektů s kvalitativními proměnnými

Tanimotova m., Jaccardův-Tanimotův a.k., Russelův-Raovův a.k., Sokalův-Michenerův a.k., Dicův k., Rogersův-Tanimotův k., Hamanův k.

MEZI DVĚMA MNOŽINAMI OBJEKTŮ

Deterministické metriky pro určení vzdálenosti mezi 2 množinami objektů

Metoda nejbližšího souseda, *k* nejbližších sousedů, nejvzdálenějšího souseda, průměrné vazby, Wardova metoda

Metriky pro určení vzdálenosti mezi 2 množinami objektů používající jejich pravděpodobnostní charakteristiky

Chernoffova m., Bhattacharyyova m. atd.

Příprava nových učebních materiálů pro obor Matematická biologie

je podporována projektem OPVK

č. CZ.1.07/2.2.00/28.0043

„Interdisciplinární rozvoj studijního oboru Matematická biologie“

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ