

12

Mehoda nejbližších sousedů, metriky, metody sítíkové analýzy
 (k-nejbližším cenným, klasifikace sítíkové)

METRIKY

definice: Vzdáenosť mezi dvojicí bodov je funkce

$d: X \times X \rightarrow [0, +\infty)$ taková, že pro každé $x, y, z \in X$

platí: 1) $d(x, y) \geq 0$ a $d(x, y) = 0 \iff x = y$
 2) $d(x, y) = d(y, x)$ (symetrie)

3) $d(x, z) \leq d(x, y) + d(y, z)$ (pravidlo sítíkové nerovnosti)

Dvojice (X, d) se pak nazývá měřítkovým prostorom.

používání měřítek:

MINKOVSKÉHO METRIKY:

L_1 - MANHATTANSKÁ VZDÁLENOST

$$d(x, y) = \sum_{i=1}^n |x_i - y_i|$$

L_2 - EUKLEIDOVSKÁ VZDÁLENOST

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

ČEBÝSEVOVA VZDÁLENOST

$$d(x, y) = \max_i |x_i - y_i|$$

METODA NEJBLÍŽŠÍCH SOUSEDŮ (kNN)

+ nejmohutnější - klasifikace dleji jiných vlastností modelu
 - predikce je náročná

- hyperparametry kNN:

1) k (počet sousedů; $k=1$ vede k "převratu")

2) míra vzdálenosti (jakožkoliv; typicky L_1/L_2)

3) váhy (jakém mísíme přidat váhy: průměr \rightarrow vážený prům.)

KLASIFIKACE S KNN

a) nalézame k nejbližším sousedům

b) provedeme majoritní volbu (typ. váženou variaciou)

c) určíme kódové objekty

REGRESE S KNN

a) najdeme k nejbližším sousedům

b) určíme průměr (typ. vážený průměr)

c) určíme (v.) průměr

B

METODY SHLUKOVÉ ANALÝZY

HIERARCHICKÉ SHLUKOVÁNÍ

= principem 2 skluků je buď prázdná množina nebo 1 a množinu sám průniku

SHLUKOVÁNÍ OBECNĚ

VSTUPY:

- 1) množiny průniků X s vzdálostí d
- 2) množina dat $D = X$
- 3) (optional) pořadování počet skluků

VÝSTUPY:



- 1) nachlad množiny na jednotlivé skluky $C = (C_1, \dots, C_k)$, kde $C_i \subset D$ a $C_i \cap C_j = \emptyset$ pro každé $i \neq j$ a $D = \bigcup_{i=1}^k C_i$

2) (optional) dendogram

DENDOGRAM = grafická reprezentace procesu hierarchického shlukování

- slovo: vrchol = venkovní skluky
- listy = počáteční 1-průnikové skluky
- korun = finální skluk

- výška je vzdálosť skluků

- k sklukům páčíme rozdílněním dendogramu mezi k-kém a k-1. nejvyšším vrcholem

MĚŘENÍ VZDÁLENOSTI SHLUKŮ

1) metoda nejbližšího souseda

$$D(A, B) = \min_{x \in A, y \in B} d(x, y)$$



2) metoda nejvzdálenějšího souseda

$$D(A, B) = \max_{x \in A, y \in B} d(x, y)$$



3) průměrná vzdálosť (average linkage)

$$D(A, B) = \frac{1}{|A||B|} \cdot \sum_{x \in A, y \in B} d(x, y)$$



AGLOMERATIVNÍ ALGORITMUS HIERARCHICKÉHO SHLUKOVÁNÍ

... je to v podstatě cesta v dendogramu odola nahoru

1) na každou každý bod jako samostatný skluk

2) while not nachavované kritérium (např. k)

- a) najdeme 2 skluky, co jsou k sobě nejblíže
- b) spojíme je do jednoho

K-NEJBЛИŽSICH CENTER (K-MEANS) (naad)

- převádí shrnkování na optimalizační úlohu
 - ↳ tzn. je potřeba definovat účelovou funkci (každak minimalizují)
- účelová funkce pro k-means → Euklidovská vzdálenost
$$G(C) = \sum_{i=1}^k \sum_{x \in C_i} \|x - \bar{x}_i\|^2$$
 → geometrický základ i-kho skluka

POSTUP:

- 1) bod x přesuneme do skluku, kde $\|x - \mu_i\|$ je nejmenší
- 2) ⇒ dojde ke snížení součtu kvadratických vzdáleností

$$\sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2 \leq \sum_{i=1}^k \sum_{x \in C_i} \|x - \bar{x}_i\|^2$$

- 3) tento postup opakujieme; poškravíme, jakmile je hodnota účelové funkce dostatečně malá

ALGORITMUS:

init.: počáteční (např. náhodné) rozdělení k shukových bodů μ_1, \dots, μ_k

- 1) rozdělíme body do skluků $C_i = \{x \in D \mid i = \arg \min_j \|x - \mu_j\|\}$
- 2) počítáme body μ_1, \dots, μ_k jako geometrické středy těchto skluků: $\mu_i \leftarrow \frac{1}{|C_i|} \sum_{x \in C_i} x$

stop: poškravíme, když je hodnota účelové funkce dostatečně malá