Princípy dátovej vedy

blok prednášok „Harman“

Typy dát

Počítačové typy dát („premenných“)

binárne/logické (*binary/logical/Boolean*), celočíselné (*integer*), reálne (*floating-point/real*), znakové (*character/string*), ...

O týchto táto prednáška nebude...

Štatistické typy dát („premenných“)

- Rôzne modely, metódy spracovania a vizualizácie dát sú určené pre rôzne štatistické dátové typy.

- Štatistická typológia dát uľahčuje komunikáciu medzi dátovými vedcami.

Základné delenie: podľa „esencie informácie“ v dátach

- Numerické/kvantitatívne dáta (*numerical/quantitative data*)

--- diskrétne numerické dáta (*discrete numerical data*)

--- spojité numerické dáta (*continuous numerical data*)

- Kategorické/kvalitatívne dáta (*categorical/qualitative data*)

--- nominálne kategorické dáta (*nominal categorical data*)

--- ordinálne kategorické dáta (*ordinal categorical data*)

Poznámka: Numerické dáta sú automaticky ordinálne a kategorické dáta sú automaticky diskrétne.

Úloha: Aké sumárne štatistiky sú primerané pre vyššie uvedené typy dát (modus, medián, kvartily, priemer, výberový rozptyl)?

Menej dôležité delenia, s ktorými sa môžeme stretnúť:

Delenie podľa „kardinality“ množiny možných hodnôt dát

- Dichotomické/binárne dáta (*dichotomous/binary data*)

- Polytomické dáta (*polytomous data*)

Množina možných hodnôt dát sa niekedy nazýva nosič rozdelenia dát (*support of the data distribution*)

Delenie podľa „relevantných vzťahov“ medzi hodnotami dát

- Intervalové dáta (*interval data*)
- Pomerové dáta (*ratio data*), ...

Delenie podľa „štatistickej závislosti“ premenných

- Nezávislé dáta (*independent data*)
- Závislé dáta (*dependent data*)

Delenie podľa „(re)prezentácie“ dát

- Vektorové dáta (*vector data*)
- Tabuľkové dáta (*tabular data*)

Delenie podľa „formy“ dát

- Štrukturované dáta (*structured data*)
- Neštrukturované dáta (*unstructured data*)

Delenie podľa „zdroja“ dát

- Textové dáta (*text data*)
- Zvukové dáta (*sound data*)
- Obrazové dáta (*image data*), ...

Úloha: Diskutujme, aké prívlastky podľa vyššie uvedených delení by sme mohli priradiť nasledovným dátam:

1) Výsledky dotazníka, v ktorom sme sa pýtali 1000 ľudí na dve otázky: a) Súhlasíte so zvýšením daní? b) Súhlasíte s vystúpením Slovenska z NATO? Každý respondent pri oboch otázkach zakrúžkoval jednu z možností: "nie", "skôr nie", "mám neutrálny postoj", "skôr áno", "určite áno".

2) Časový rad hodnôt koncentrácií SO_2 (oxid siričitý), NO (oxid dusný) a CO (oxid uhoľnatý) na monitorovacej stanici pri križovatke v meste; dáta sú zaznamenané každú celú hodinu počas obdobia jeden rok.

Prevod nenumerických typov dát do numerickej formy

Niekedy sa takýto proces, resp. takéto zobrazenie nazýva vnorenie dát (*data embedding*). Inokedy sa pojmom vnorenie dát označuje príbuzná technika: zmenšenie dimenzie veľarozmerných dát (*data dimensionality reduction*).

Prevod kategorických/kvalitatívnych dát

- *dummy variables,*

- *one-hot encoding, ...*

Prevod textových dát

- *bag of words,*

- *word-to-vec*, ...

Ide o oblasť takzvaného spracovania prirodzeného jazyka (*natural language processing*, NLP)

Prevod zvukových a obrazových dát

- Diskrétna Fourierova transformácia (*discrete Fourier transform*),

- Využitie landmarkov (*landmarks*), ...

Ide o oblasť takzvaného spracovania zvuku a obrazu (*sound and image processing*)

Univerzálny princíp rôzne typy dát

Vytvorenie numerickej matice vzájomných „nepodobností“ a použitie niektorej z metód mnohorozmerného škálovania (*multidimensional scaling*).

Funkčné transformácie numerických dát

- Prevod dát na iné jednotky: lineárna transformácia

- Zmena rozdelenia dát na také rozdelenie, s ktorým sa lepšie pracuje, alebo ktoré lepšie spĺňa predpoklady použitej metódy, napr. logaritmická transformácia na šikmé dáta (*skewed data*): ak majú pôvodné dáta približne lognormálne rozdelenie (čo sa často stáva), tak zlogaritmované dáta majú približne normálne rozdelenie

- Lepšia separácia dát do zhlukov (*cluster analysis*), napr. odmocninová transformácia

- Informatívnejšia vizualizácia dát, napr. logaritmická transformácia, alebo rotácia a projekcia viacrozmerných dát