

Předzpracování dat (osnova přednášky)

Tomáš Sieger

20. března 2014

Osnova

- kroky
 - (výběr relevantních dat)
 - agregace dat
 - transformace dat
 - ošetření odlehlých pozorování (*outlier detection*)
 - ošetření chybějících hodnot
 - výběr příznaků a výpočet druhotných příznaků (*feature selection and extraction*)
- spíše heuristiky, ne hluboká teorie

Výběr relevantních dat

- ne všechno je žádoucí zpracovávat
- nutná diskuze se zadavatelem
- často nutný vhled do kontextu dat (jak, proč, kdy, na čem byla data získána)
- typy dat a na co se hodí ^[1]
 - popisná analýza (data: sčítání lidu (*census*))
 - průzkumná analýza (korelace vs. kauzalita)
 - deduktivní analýza (*inferential*) (data: výběr z populace)
 - prediktivní analýza (neplést s kauzální analýzou!)
 - analýza kauzálních vztahů (data: randomizovaná studie)

Agregace dat

- cíl: získat jedinou tabulku použitelnou pro algoritmy dobývání dat (DD)
- vstup typicky: mnoho různě klíčovaných tabulek

Transformace dat

- cíl: “rozumné” rozdělení dat, stabilizace rozptylu
 - usnadní vizualizaci
 - může usnadnit interpretaci
 - vyjde vstříc metodám DD (př. výpočet histogramu, regrese)
- spojité veličiny
 - př.: $\log(x)$, \sqrt{x} , Boxova-Coxova transformace
 - stabilizace rozptylu: $\arcsin(\sqrt{p})$ pro pravděpodobnost p , Fisherova $\frac{1}{2} \ln \frac{1+\rho}{1-\rho}$ [2] pro korelační koeficient ρ
 - diskretizace (ekvidistatní, ekvifrekvenční)
 - normalizace (cíl: učinit různé datové výběry srovnatelné; učinit příspěvky jednotlivých příznaků srovnatelné)
- diskrétní veličiny
 - slučování vedlejších málo zastoupených kategorií

Ošetření odlehlých pozorování

- cíl: zbavit se odlehlých pozorování, která by negativně ovlivnily analýzu
- problém: co je odlehlé pozorování?
 - př. hmotnost tloušťka v klinických datech, plat v dotazníkovém šetření, živelná událost v pojišťovnictví
 - často nutná diskuze se zadavatelem, vzhled do kontextu
- vždy postupovat opatrně, pozor na ztrátu cenných dat a vychýlení odhadů!
- může pomoci průzkumná analýza (je odlehlá jen hodnota jednoho příznaku, nebo více příznaků?)
- **NE**: aplikovat slepě automatické metody odstraňování odlehlých pozorování

Ošetření chybějících pozorování

- chybějící pozorování [3]
 - z principu (př. počet těhotenství u mužů)
 - skutečně chybějící
- co dělat?
 - zahodit pozorování s některým chybějícím příznakem ?!
 - zahodit příznaky s některou chybějící hodnotou ?!
 - nahradit chybějící hodnotu indikátorem chybějící hodnoty ?!
 - nahradit chybějící hodnotu nějakým číslem ?!
 - průměr, medián, modus, nejbližší soused, predikce na základě nějakého modelu
- POZOR na vychýlení odhadů!
- co dělat, když potřebuji nahrazení číslem je jediná možnost?
 - mnohonásobné nahrazení (*multiple imputation, sensitivity analysis*)
- longitudinální data: *Missing Completely at Random (MCAR)*, *Missing at Random (MAR)*, *Missing Not at Random (MNAR)*

Výběr příznaků a výpočet druhotných příznaků

Feature Selection & Extraction

- proč:
 - vyřazení příznaků, které nesmějí být použity (př. jinak zakódovaná odezva)
 - snížení počtu příznaků, aby $\# \text{ příznaků} < \# \text{ instancí}$
 - jinak hrozí numerická nestabilita, nekonvergence, přeučení (*overfitting*)
 - př.: klasifikace 2 jedinců podle 10ti příznaků
 - př.: klasifikace 200 tkáňových vzorků s 100.000 geny do dvou tříd
 - fakt: výkon klasifikátoru s narůstajícím počtem příznaků nejprve strmě narůstá, pak klesá
- metody:
 - výběr příznaků (*feature selection*)
 - výpočet druhotných příznaků (*feature extraction*)

Výběr příznaků

Feature Selection

- redukovat dim. příznakového prostoru výběrem pouze některých příznaků
 - ručně - na základě znalosti povahy problému, zkušenosti
 - automaticky
 - někdy přirozenou součástí algoritmu DD (př. rozhodovací stromy, náhodný les, *boosting*, regularizace: hřebenová regrese (L2), Lasso (L1), *elastic net*)
 - nebo zautomatizované hledání vhodné podmnožiny příznaků “hrubou silou”
- heuristika v učení s učitelem: hledat mezi příznaky, které mají vztah k odezvě
 - pomocí korelace, vzájemné informace, analýzy kontingenčních tabulek (χ^2 test, logistická/Poissonská regrese), ...
 - pozor, můžeme tak vynechat informativní příznak!
 - př.: 2 jasně oddělené populace ve 3D nemusejí být po rotaci viditelně oddělené v 1D projekcích
- *boosting*: použít více klasifikátorů, každý používá pouze

Výpočet druhotných příznaků

Feature Extraction

- redukovat dim. příznakového prostoru výpočtem malého množství nových příznaků
- obecně těžké: $f, \mathbb{X}_i \in \mathbb{R}^p \rightarrow f(\mathbb{X}_i) \in \mathbb{R}^{p_0}, p_0 < p$
- př.: součet příznaků, rozdíl příznaků, ..., (*Body Mass Index (BMI)*)
- často: hlavní komponenty (*Principal Component Analysis (PCA)*)
 - problém: “míchání” jednotek, ztráta interpretace
- užitečný trik: vyrovnání příznakového prostoru (*feature space straightening*)

Literatura

- [1] Jeff Leek: **Data Analysis**,
<https://www.coursera.org/course/dataanalysis>
- [2] Jiří Anděl: **Statistické metody**, Matfyzpress, Praha
- [3] **Missing Data GUI**, *<http://cran.r-project.org/web/packages/MissingDataGUI/index.html>*
- [4] Cosma Rohilla Shalizi: **Advanced Data Analysis from an Elementary Point of View**,
<http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/>
- [5] **R**: A language and environment for statistical computing.
<http://www.R-project.org/>