PDV 20.12.2024

piatok 20. decembra 2024 11:10

- skúškové otázky: najbližšie ako
- ANKETA

Ak projekt vyriešite skôr, napíšte mi, prosím mail

Minule: Testy

perm. test: H_0: rozdelenie X = rozdelenie Y (vzhľadom na stat. $\check{S}$)

H_1: $\neq$

napr. $\check{S}_* = \bar{x} - \bar{y}$ X: muži, Y: ženy

- $x_1, \dots, x_m, y_1, \dots, y_n$

  preusporiadame značky M/Ž $\Rightarrow \check{S} = \dots$ $\rightarrow$ # preusporiadanie

- H_0 zam. ak $\check{S}_*$ je v 5% najextrémnejších

Pr.: 3 muži, 2 ženy

váhy 87,95,72 63,79

H_1: muži sú ťažší ako ženy

$\check{S}_* = \bar{x} - \bar{y} = 13{,}7$

87, 95, 72, 63, 79

$m_1\ m_2\ m_3\ \check{z}_1\ \check{z}_2 \rightarrow \check{S} = \dots$

$m_1\ m_2\ \check{z}_1\ \check{z}_2\ m_3 \rightarrow \check{S} = \dots$

$\vdots$

$\downarrow$

5! S-ich

p-val = % extrémnejších $\check{S}$

- častokrát: iba k náhodným permutáciám: aproximácia

Big Data

- také, ktoré nevieme spracovať na 1 počítači
  $\rightarrow$ Twitter, FB, Google...
- rýchlosť – statické: fixné
  - "historické": pomaly pribúdajúce – denné výnosy
  - streamovacie: real-time pribúdajúce
    - nemáme vždy kapacitu uložiť
    - $\rightarrow$ snímanie kamier, Twitter...

  } Big data

- kvalita – rozhoduje sa za účelom (chceme získať účinok lieku)

- výsledok plošného zberu: všetko, čo súvisí so skúmaným } Big data

→ nemusia byť reprezentatívne...

Hlavný problém: nemôžeme použiť klasické postupy (parametre)

Riešenia: I. vyhnúť sa práci s veľkými dátami:

→ 1.) Vzorkovanie (sampling, subsampling) - iba časť dát

→ 2.) Sumarizácia (sufficient statistics) - pracovať iba s tým, čo potrebujem

II. Špec. postupy pre Big data

1.) Vzorkovanie:

- modely typicky majú rozumný # parametrov → môže stačiť menej dát

a) Náhodný výber: reproducibilita?

: môže byť ťažké získať viac náh. vzoriek, ktoré sa neprekrývajú (train, val, test)

Nenáhodne:

b) Prvých k záznamov → môže byť veľmi nereprezentatívne

c) "každé L-té": 1, 1001, 2001, 3001, ...

: také $i \equiv 1 \pmod{1000}$

alebo $\equiv 34$ ...

: ľahko neprekrývajúce sa výbery: $i \equiv 1 \pmod{1000}$ → 1.)

$i \equiv 2 \pmod{1000}$ → 2.)

: môže byť zlé, ak dáta majú špec. štruktúru:

napr. zdravý-chorý-zdravý-chorý...

Všeob.: pozor na reprezentatívnosť

- nechceme stratiť relevantných outlierov

↳ "malé triedy" (napr. nezbankrotované firmy)

→ stratifikované vzorkovanie:

- najprv rozdelím na mužov a ženy → potom z každej skupiny: výber

→ váhované vzorkovanie: rôzne pp. rôznym záznamom

(napr. vysoká pp. pre objekty zo vzácnej triedy)

2.) Sumarizácia

- uložím si len to, čo mi treba pre model
  → sufficient statistics (postačujúce štat.)
- napr. kontingenčná tabuľka namiesto veľa záznamov

II. Keď to inak nejde: špecializované postupy

- relačné databázy (SQL)
- "shamelessly" paralelizovateľné výpočty
  - ak sa dá rozložiť na menšie výpočty, ktoré sú spustiteľné samostatne
  - každý výpočet na 1 počítači / jadre ...
- paralelné výpočty
  - každá jednotka svoje, ale musí sa to priebežne spájať / aktualizovať
  - treba minimalizovať množstvo dát, ktoré si počítače medzi sebou posielajú

→ ∃ "jazyky": frameworky - MapReduce, Hadoop

- treba používať overené nástroje, neimplementovať vlastné riešenia
  → pri veľa dátach sa niečo pokaziť, frameworky to majú ošetrené