Mnohorozměrná statistika

1 Vícerozměrná (mnohorozměrná) data

- na každém objektu měříme několik veličin, obecně závislých
- příklady:
  - výška, věk, hmotnost, rasa, pohlaví, ... člověka,
  - různé makroekonomické charakteristiky států EU,
  - charakteristiky podnebí (teplota, znečištění vzduchu, množství srážek, povětrnost,...) různých míst ČR
  - ...
- počet objektů může být menší než počet měřených veličin

Mnohorozměrná statistika
- statistické metody pro analýzu mnohorozměrných dat
- spíše průzkumné metody (explorativní charakter)
  - hledání zákonitostí, struktury,
  - zkoumání vztahů mezi objekty a proměnnými,
  - pochopení souvislostí,
  - působí méně exaktním dojmem,
- větší zacílení na jedince než v klasické statistice.

Hlavní cíle vícerozměrné statistiky:
- redukce vícerozměrnosti,
- vizualizace dat, identifikace odlehlých hodnot (netypických jedinců),
- hledání struktury v datech (podobnosti objektů, vztahy mezi proměnnými),
- klasifikační úlohy.

Základní metody:
1. Analýza hlavních komponent (PCA)
2. Faktorová analýza (FA)
3. Diskriminační analýza (DA)
4. Shluková analýza (CA)
5. Další: Kanonické korelace (CC), Korespondenční analýza, Mnohorozměrné škálování, Metoda hlavních koordinát, Klasifikační stromy, atd.

Poznámky
- Většina metod „ideální“ pro normální rozdělení.
- Často standardizace dat.

2 Popisná statistika a vizualizace dat

- 2D grafy, matice grafů
- mnohorozměrná vizualizace (Chernoffovy tváře)
- korelační matice (tabulka korelačních koeficientů)
- kovariance, kovarianční matice

3 Analýza hlavních komponent (PCA)

- snaha snížit variabilitu do méně dimenzí,
- idea: jsou-li dvě proměnné silně korelované, můžeme je do jisté míry nahradit jednou novou (společnou) proměnnou,
- chceme, aby ztráta informace byla minimální.

K čemu slouží
- lepší pochopení vztahů mezi proměnnými,
- lepší pochopení vztahů mezi jedinci (vizualizace vícerozměrných dat ve 2D obrázcích),
- hledání odlehlých hodnot,
- následné využití v dalších statistických metodách (regrese).

Postup
1. odhad (výpočet) hlavních komponent z korelační (kovarianční) matice,
2. volba počtu hlavních komponent (různá kritéria),
   - komponenty s rozptylem více než 1,
   - sutinnový graf,
   - požadavek na vysvětlenou variabilitu (např. 70 % apod.),
3. interpretace hlavních komponent: korelace s původními proměnnými (loadings), grafické znázornění,
4. znázornění dat vzhledem k novým souřadnicím (hlavním komponentám) — tzv. biploty.

4 Faktorová analýza (FA)

- aplikace zejména v psychologii, sociálních vědách atd.,
- snaha vysvětlit variabilitu v proměnných pomocí menšího počtu latentních (nepozorovatelných) faktorů,
- rozšiřuje interpretační schopnosti PCA,
  - rotace faktorů tak, aby vzniklé veličiny byly hodně korelované s některými málo proměnnými a velmi málo korelované s ostatními proměnnými,
  - interpretace závisí na výsledku rotace – nutná znalost dané problematiky,
  - různé metody rotace (varimax, promax,...),
- počet faktorů je třeba specifikovat na začátku analýzy.

5 Shluková analýza (CA)

- snaha rozdělit objekty do skupin, jejichž prvky jsou si v určitém smyslu podobné,
- shluk = skupina objektů, které jsou si podobné a jsou odlišné od objektů patřících do jiných shluků,
- vhodná standardizace dat,
- interpretace výsledných shluků.

Metody shlukové analýzy

- Hierarchické metody:
  - počet shluků nemusí být na začátku stanoven
  - na začátku je každý objekt jeden shluk, potom postupně v každém kroku spojíme dva nejbližší shluky dokud nejsou všechny objekty v jednom velkém shluku.
- Nehierarchické metody:
  - na začátku musíme stanovit požadovaný počet shluků,
  - analýza rozdělí objekty do shluků „optimálně“.

Hierarchické metody

- založené na vzdálenosti (vzdálenější objekty méně podobné),
- vzdálenost dvou objektů (Eukleidovská, Manhattan, ...)
- vzdálenost dvou shluků (princip nejbližšího, nejvzdálenějšího souseda, Wardova metoda, metoda centroidů),
- různé metody mohou dávat různé výsledky,
- grafická prezentace: dendrogram.

Metoda k-means

- na začátku specifikujeme počet shluků,
- minimalizuje součet čtverců od center shluků,
- iterační postup,
- výsledek může značně záviset na počáteční volbě center.

6 Diskriminační analýza (DA)

- v datech k skupin, na každém objektu měřeno několik znaků,
- chceme najít pravidlo založené na měřených znacích, pomocí něhož by bylo možné rozlišit jednotlivé skupiny,
- příklady: klasifikace rostlin, credit scoring, antropologie, ...,
- tzv. učení s učitelem (na rozdíl od shlukové analýzy),
- metoda předpokládá normalitu, ale není moc citlivá na porušení.

Postup

1. krok: vytvoření diskriminačního pravidla:
   - vytvoření pravidla pro rozlišení skupin na základě trénovacích dat,
   - příslušnost objektů do skupin je známá,
   - je možné nastavit tzv. apriorní pravděpodobnosti jednotlivých skupin,
   - ověření „kvality“ diskriminace.
2. krok: klasifikace nových objektů:
   - zařazení nového jedince do některé ze skupin,
   - jeho skutečnou příslušnost neznáme.

7 Kanonické korelace (CC)

- zkoumá vztah mezi dvěma skupinami proměnných,
- příklad: vztah známek na vysvědčení a sérií IQ testů, vztah environmentálních a druhových proměnných, ...,
- kanonické proměnné = lineární kombinace původních, které maximalizují vzájemné korelace,
- kanonické korelace = příslušné korelace kanonických proměnných