Metodika analýzy dat: Od základů po aplikace metod strojového učení

Statistické testy

Jakub Steinbach, Jan Vrba

Ústav počítačové a řídicí techniky
VŠCHT

2.10.2024

Obsah slajdů I

1. Úvod do statistického testování

2. Typy statistických testů
   - Testy založené na porovnání průměru mezi kategoriemi
   - Testy závislosti a asociace
   - Testy rozdělení dat

Úvod do statistického testování

Úvod do statistického testování

Proč testovat? Mám pocit, že existuje souvislost mezi dvěma proměnnými (například velikostí příjmu a výskytem depresí), tak chci ověřit, jestli je tato souvislost podpořená daty samotnými.

Postup při statistickém testování

1. Formulujeme nulovou hypotézu H_0 o neexistenci rozdílu (například velikost příjmu nemá vliv na výskyt deprese).
2. Volíme hladinu významnosti α, jež stanovuje pravděpodobnost výskytu falešně pozitivního výsledku (tedy, že přijmeme alternativní hypotézu, i když ve skutečnosti platí H_0). Tato hladina přitom reprezentuje pravděpodobnost, s jakou tato situace může nastat. Obvykle se tato hladina volí $\alpha = 0.05 = 5\%$, případně nižší.
3. Formulujeme alternativní hypotézu H_1, jež může být jednostranná (například čím je vyšší velikost příjmu, tím je nižší výskyt deprese), či oboustranná (například velikost příjmu má statisticky významný vliv na výskyt deprese).
4. Zvolíme vhodnou testovou statistiku, jež bude sloužit jako kritérium pro rozhodnutí o H_0. U testové statistiky předpokládáme znalost jejího rozdělení, a tudíž i pravděpodobnost jejího výskytu při platnosti H_0.

Postup při statistickém testování

5. Na základě pozorovaných veličin spočítáme hodnotu testové statistiky a její pravděpodobnost výskytu.
6. Pokud je pravděpodobnost výskytu menší než je hladina významnosti, zamítáme H_0 a přijímáme H_1. V opačném případě nelze rozhodnout o platnosti ani jedné z hypotéz.

Typy statistických testů

Studentův t-test test

Tento test se používá v případech, kdy chceme porovnat průměrné hodnoty dvou souborů dat. Předpokladem je normální rozdělení dat a rozptyly dat v obou souborech jsou podobné. Pro případ odlišných rozptylů lze využít Welchův t-test.
Rozlišují se tři aplikace testů:

- **jednovýběrový test** - pro případy, kdy porovnáváme průměrnou hodnotu souboru dat proti hypotetické nebo očekávané hodnotě
- **dvouvýběrový test** - pro případy, kdy porovnáváme průměrnou hodnotu dvou nezávislých souborů dat
- **párový test** - pro případy, kdy porovnáváme dva související nebo závislé soubory dat (například před a po nějaké změně)

ANOVA

ANOVA (analýza rozptylů) se používá v podobných případech jako t-test, tedy pro porovnání průměrných hodnot, nicméně mezi více skupinami dat. Předpokladem je normální rozdělení dat a rozptyly dat v obou souborech jsou podobné. ANOVA vede pouze k rozhodnutí, zdali existuje statisticky významný rozdíl mezi kategoriemi, nespecifikuje však, mezi kterými (k tomu je třeba využít některý z post-hoc testů). V aplikaci se používají dva typy testů:

- **jednofaktorová ANOVA** - podobná t-testu, umožňuje testovat vliv jedné závislé proměnné na jedné nezávislé pro více kategorií
- **dvoufaktorová ANOVA** - umožňuje testovat vliv dvou nezávislých proměnných na jednu závislou proměnnou

Korelace a korelační koeficient

Korelace popisuje sílu a směr vztahu mezi dvěma proměnnými. Rozlišujeme tři "směry"korelace:

- **pozitivní korelace** - se zvyšující se hodnotou jedné proměnné se zvyšuje i hodnota druhé a opačně, se snižující hodnotou jedné proměnné se snižuje i hodnota druhé proměnné
- **negativní korelace** - se zvyšující se hodnotou jedné proměnné se snižuje hodnota druhé a opačně, se snižující hodnotou jedné proměnné se zvyšuje hodnota druhé proměnné
- **nulová korelace** - neexistuje žádný vztah mezi oběma proměnnými

Limitace korelačního koeficientu je to, že obě proměnné musí vyjadřovat míru a musí být seřaditelné podle této míry. typicky se může jednat o číselné proměnné, kde číslo vyjadřuje kvantitu, ale i o ordinální proměnné, kde hodnoty lze seřadit podle nějakého vztahu (například nejmenší, velmi malý, malý, střední, velký, velmi velký, největší).

Pearsonův korelační koeficient

Pearsonův korelační koeficient se používá pro vyjádření **lineární** korelace mezi dvěma **číselnými** kvantitativními proměnnými. Korelační koeficient r dosahuje hodnot $< -1, 1 >$, s tím, že:

- $r = -1$ značí dokonalou zápornou lineární závislost (body leží na klesající přímce)
- $r = 0$ značí nulovou korelaci (body leží náhodně rozprostřené v rovině)
- $r = 1$ značí dokonalou kladnou linární závislost (body leží na rostoucí přímce)

Spearmanův korelační koeficient

Spearmanův korelační koeficient se používá pro vyjádření korelace řazení dvou ordinálních či číselných proměnných. Je tedy rozšířením Pearsonova korelačního koeficientu, jelikož se neomezuje na lineární závislost. Podobně u zde ρ dosahuje hodnot $< -1, 1 >$, s tím, že:

- $\rho = -1$ značí dokonalou zápornou závislost (body leží na klesající křivce, již lze popsat monotónní funkcí)
- $\rho = 0$ značí nulovou korelaci (body leží náhodně rozprostřené v rovině)
- $\rho = 1$ značí dokonalou zápornou závislost (body leží na rostoucí křivce, již lze popsat monotónní funkcí)

Point-biseriální korelace

V případech, kdy jedna z proměnných je binární (dosahuje pouze dvou hodnot), lze spočítat tzv. point-biseriální korelaci, jež je určena upraveným Pearsonovým korelačním koeficientem. Alternativně lze použít vztahy pro Pearsonův a Spearmanův korelační koeficient. Výsledek nám potom naznačí, jak se liší hodnoty pro číselnou či ordinální proměnnou pro každou z kategorií bninární proměnné.

$\mathcal{X}^2$-test

$\mathcal{X}^2$-test se používá k určení asociace mezi dvěma kategorickými proměnnými. Používá se především ve dvou variantách:

- $\mathcal{X}^2$-test nezávislosti - pro ověření, zdali jsou obě kategorie na sobě závislé
- $\mathcal{X}^2$-test dobré shody - pro ověření, zdali výskyt v jednotlivých kategoriích odpovídá hypotetickému nebo očekávanému výskytu

Pro použití testu je třeba splnit následující podmínky:

1. proměnné jsou kategorické, v případě číselných se musí hodnoty sloučit do skupin
2. jednotlivá pozorování jsou nezávislá (každý prvek má přesně jednu hodnotu v každé kategorii)
3. počet pozorování pro každou kombinaci hodnot jednotlivých kategorií je minimálně 5

Alternativy $\mathcal{X}^2$-testu

Jako alternativa pro malé datové soubory lze využít následující testy, jež neodhadují rozhodovací kritérium z pravděpodobnostního rozdělení, ale stanovují jej přesně:

- pro malé soubory dat, jež nesplňují podmínku minimálního výskytu dat v kategoriích lze aplikovat Fisherův exaktní test
- pro případ dvou binárních kategorií je nejpřesnější Boschlooův test

Oba testy lze využít jako jednovýběrové i jako dvouvýběrové.

Kolmogorov-Smirnovův test

Tento test se používá pro určení, zdali soubor dat sleduje určené rozdělení pravděpodobnosti, případně zdali dva datové soubory náleží stejnému rozdělení pravděpodobnosti.
Rozlišují se dvě aplikace testů:

- **jednovzorkový test** - pro případy, kdy porovnáváme pravděpodobnostní rozdělení souboru dat proti rozdělení hypotetickému nebo očekávanému
- **dvouvzorkový test** - pro případy, kdy porovnáváme pravděpodobnostní rozdělení dvou datových souborů

Předpokladem pro použití testu je nezávislost jednotlivých pozorování dat a data by měla být číselná anebo náležet do číselného intervalu.
Předpokladem je také větší množství vzorků v datovém souboru (přibližně alespoň 30), jinak může zkreslovat výsledek.
V případě malého souboru dat lze využít Shapiro-Wilkův test, ale pouze za předpokladu normálního rozdělení.