Statistické minimum

(b) *(5 b)* Vysvětlete význam pojmu matoucí proměnná (confounding variable). Uveďte příklad a naznačte vliv na model.

Matoucí proměnná: Mění se současně s závislou proměnnou, je těžké zjistit kauzalitu. Např. při testování klávesnic má vliv předchozí zkušenost s nimi
Dummy variable: máme kategorickou proměnnou, se kterou potřebujeme nějak pracovat. Každé kategorii přidělíme číslo, které ale nemá žádný číselný význam, nemá smysl je porovnávat.
Všechny druhy dat:

- numerická data:
    - diskrétní data s velkým množstvím dat
    - spojitá data
- kategorická data: jsou rozdělena do tříd
    - ordinal data: mají pořadí e.g. dosažené vzdělání
    - intervalové data: mají pořadí + stejný interval v každé třídě

Analýza rozptylu

(a) *(2 b)* K čemu se používá parametrická jednostupňová analýza rozptylu (parametric one-way ANOVA)? Formulujte její nulovou a alternativní hypotézu.

H0: Všechny průměry jsou shodné
H1: Alespoň mezi dvěma průměry jednotlivých skupin je rozdíl.

Používá se, když chceme zjistit, jestli je nějaký rozdíl mezi testovanými skupinami, a skupiny jsou víc než 2. Pro 2 skupiny můžeme použít t-test.

(b) *(3 b)* Jaké má tato metoda předpoklady? Jak je budete testovat? Co se stane, pokud splněny nejsou?

Předpokládáme, že

- třídy mají normální rozdělení. Příslušnost k rozdělení testujeme chi^2 testem
- všechny třídy mají stejný rozptyl. pro porovnání rozptylů použijeme Welchův test.
- vzorky jsou nezávislé, pokud nejsou, můžeme dělat opakovanou ANOVU
- **kdyz nejsou splneny predpoklady:** záleží

    když ty data nejsou normálně rozdělený, ale skupinky maj pořád stejný rozptyly, tak ten test asymptoticky pořád funguje

    meaning že dokud máš dost dat, tak se v zásadě nic neděje

    protože tam zafunguje nějaká centrální limitní věta, a všechno bude asymptoticky normální

    když neplatí shodnost rozptylů těch skupin, tak je to trochu horší

    pokud máš aspoň vyvážený počty dat v jednotlivejch skupinách, tak se empiricky ukazuje že to pořád celkem funguje (ale už bez nějakejch matematickejch garancí jako byly u porušení normality)

    pokud máš různý rozptyly a různý četnosti skupin, případně chceš bejt opatrnej, tak existuje zobecnění anovy, říká se tomu Welchova formulace myslim, a ta pak funguje v pohodě

(c) *(3 b)* Podrobně popište výstupní tabulku ANOVA testu na konci posloupnosti příkazů níže.

```
F<-unlist(mapply(rep,times=c(8,9,10),x=c(1,2,3)))
O<-F+rnorm(n=27,mean=0,sd=2)
summary(aov(O ~ as.factor(F)))
            Df Sum Sq Mean Sq F value Pr(>F)
F            2  11.41   5.707   1.953  0.164
Residuals   24  70.14   2.923
```

oznacim N ... pocet dat, K ... pocet skupin

vezmu to odspoda, radek Residuals ma `DF = N-K`, `Sum Sq = residualni soucet ctvercu`, `Mean Sq = Sum Sq / DF`

residualni soucet ctvercu je soucet druhejch mocnin rozdilu jednotlivejch bodu od prumery jejich skupiny

v radku F je `DF = K-1`, `Sum Sq = soucet ctvercu skupin`, `Mean Sq = Sum Sq / DF`, `F value = (Mean Sq F)/(Mean Sq Residuals)`, `Pr(>F) = 1 - CDF_F(F value)`

soucet ctvercu skupin je suma druhejch mocnin rozdilu mezi prumerama jednotlivejch skupin

CDF_F je distribucni funkce F rozdeleni s DF_F a DF_Residuals stupni volnosti

df = stupne volnost = n-1
sum.sq = prvni radek = sst
= druhy radek = sse
mean sq = prvni radek = sst/df, druhy radek = sse/df
F statistics = msst/msse
p value - to je snad jasny kurva (porovnavam F statistiku s tabulkovou hodnotou)

prvni radek = rozptyl mezi skupinama (variance co si nevysvetlil)
druhy radek = rozptyl ve skupinach (to, co sem vysvetlil)

.
(d) *(2 b)* K čemu slouží následný post-hoc test? Na jakém principu je založen?

anova pouze rekne, ze tam je nejakej rozdil mezi group means
post hoc test rekne, jaky konkretni skupiny za to muzou
pouziva se k tomu tukey's honest signif. difference test
= t test ktery zkouma family-wise error rate, porovnava vsechny pary group means
= rozpozna vsechny, kterey jsou vetsi nez expected standard error
post hoc test by se mel pouzit vzdy, kdyz je vysledek statisticky signifikantni (jinak to nema smysl)

Diskriminační analýza

(a) *(2 b)* Z jaké myšlenky obě metody vycházejí? Napište definiční vztah.

- snazi se vyjadrit zavislou promennou jako linearni kombinaci jeji featur
- narozdil od anovy, ktera uvazuje kategorickou nezavislou a spojitou zavislou, pouziva LDA spojitou nezavislou a kategorickou zavislou (proste opak)
- predpoklad: normalne rozdelene nezavisle promenne
    - random sampling

- stejne variance mezi skupinama
- relativne robustni proti malym porusenim predpokladu
- uzitecna kdyz mam maly pocet samplu nebo classes well separated
- chce velke rozdily meanu ale male variance ve skupinach
- pouziva se pro k>3, pro k=2 je to fisherova diskr. analyza

(b) *(2 b)* Jaký je základní rozdíl mezi LDA a QDA? Z čeho plyne?

MIMO: pca vs lda = pca - nejvetsi rozptyl mezi daty, lda = nejmensi rozptyl uvnitr kazde tridy

lda = linearni, qda = nelinearni
u qda nemusim predpokladat stejne kovariance mezi tridy
separujici povrch bude kvadratickej (a ne treba primka jako u lda)

(c) *(1 b)* Předpokládejte, že řešíte problém s lineární bayesovskou rozhodovací hranicí. Která z metod dosáhne vyšší přesnosti nad trénovacími daty? Která nad testovacími? Proč?

- QDA bude mít vyšší přesnost na trénovacích datech

čistě proto že ta rozhodovací hranice je wiggly-wiggly, a umožní mi to se víc overfitnout
sice je hezký že bayesovskej klasifikátor je lineární, ale na ty konkrétní trénovací data se stejně přeučim líp

- bude myslim LDA lepší na testovacích

  právě protože je to optimální rozhodnutí, a nebude to přeučený na nelineární vzory v trénovacích datech, který byly způsobený náhodou

(d) *(1 b)* Předpokládejte, že řešíte problém s nelineární bayesovskou rozhodovací hranicí. Která z metod dosáhne vyšší přesnosti nad trénovacími daty? Která nad testovacími? Proč?

- QDA bude mít vyšší přesnost na trénovacích datech

čistě proto že ta rozhodovací hranice je wiggly-wiggly, a umožní mi to se víc overfitnout
sice je hezký že bayesovskej klasifikátor je lineární, ale na ty konkrétní trénovací data se stejně přeučim líp

- na testovacich qda protoze ta rozhodovaci hranice neni linearni

(e) *(1 b)* Uvažujte obecnou klasifikační úlohu. S rostoucím počtem trénovacích příkladů relativní testovací klasifikační přesnost QDA vzhledem k LDA poroste, bude klesat nebo se nebude měnit? Proč?

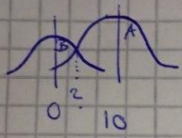
bylo by divný kdyby přesnost QDA klesala vůči LDA, že jo
protože je to koneckonců zobecnění LDA
nejsem si úplně jistej jestli se nebude měnit a nebo poroste
kdybych uvažoval idealizovanej případ, počet dat jdoucí k nekonečnu, tak data v trénovací množině budou velmi podobný těm v testovací
už jsi toho při tréninku viděl tolik, že tě při testu data nepřekvapěj, řekneš si "jo, to sem viděl - bylo hůř"
pak bych řek, že bude QDA lepší oproti LDA
protože tim že trénovací a testovací data jsou si tak podobný, tak nehrozí overfit, protože to na čem se naučíš je tak podobný tomu na čem testuješ
a LDA tam nakreslí rovnou čáru jak retard, QDA ji aspoň ohne
tudíž by to mělo klasifikovat líp případnou nelinearitu (byť pořád ne nijak zázračně)
kdyby byla skutečně potřeba lineární rozhodovací hranice, tak by se LDA chovalo líp, protože ta kvadratická by nakonec uhnula a klasifikovala blbě (z kvadratický rovnou neuděláš)

ale píšou "obecná klasifikační úloha", takže bych nepředpokládal specifickej případ lineární rozhodovací hranice

.

(f) *(3 b)* Máte určit, zda na akcii firmy s loňským ročním výnosem 4% bude vyplacena dividenda. Z burzovní analýzy velkého počtu firem víte, že firem, které vyplácí dividendu, je 80% a jejich průměrný roční výnos je 10%. Firmy bez dividendy mají průměrný výnos 0%. Rozdělení výnosů v obou skupinách je normální s rozptylem $\hat{\sigma}^2 = 0.36$. Budete aplikovat LDA, nebo QDA? Nemusíte důsledně počítat pravděpodobnost, stačí přesně zapsat.

$\sqrt{6{,}28 \cdot 0.36}$



$$0.8 f_A = 0.2 f_B \qquad f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$4 f_A = f_B$$

$$4e^{-\frac{(x-10)^2}{2\sigma^2}} = e^{-\frac{(x)^2}{2\sigma^2}} \quad /\ln$$

$$\ln(4) + (-1)\cdot\frac{(x-10)^2}{2\sigma^2} = \frac{-x^2}{2\sigma^2} \quad /\cdot 2\sigma^2$$

$$2\sigma^2 \cdot \ln(4) - [x^2 - 20x + 100] = -x^2$$

$$0.998 - x^2 + 20x - 100 = -x^2 \quad /+x^2$$

$$0.998 + 20x - 100 = 0 \quad /+100,\ -0.998$$

$$20x = 100 - 0.998$$

$$20x = 99.002$$

$$x = 4.95$$

Rozhodovací hranice je 4.95. Pokud výnos > 4.95, rozhodne LDA, že dividenda bude vyplacena.

Multivariátní regrese

$$\sum_{i=1}^{m}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\hat{\beta}_j^2$$

Parametr λ nejprve nastavíte na 0, poté jej postupně zvyšujete. S nárůstem λ

(a) *(2 b)* trénovací reziduální součet čtverců (residual sum of squares, RSS)

i) zpočátku poroste, od jisté doby ale začne klesat a vytvoří invertovanou U křivku,
ii) zpočátku bude klesat, od jisté doby ale začne růst a vytvoří U křivku,
iii) bude stále růst,
iv) bude stále klesat,
v) zůstane konstantní.

(b) *(2 b)* testovací RSS

i) zpočátku poroste, od jisté doby ale začne klesat a vytvoří invertovanou U křivku,
ii) zpočátku bude klesat, od jisté doby ale začne růst a vytvoří U křivku,
iii) bude stále růst,
iv) bude stále klesat,
v) zůstane konstantní.

(c) *(2 b)* variance

i) zpočátku poroste, od jisté doby ale začne klesat a vytvoří invertovanou U křivku,
ii) zpočátku bude klesat, od jisté doby ale začne růst a vytvoří U křivku,
iii) bude stále růst,
iv) bude stále klesat,
v) zůstane konstantní.

(d) *(2 b)* zaujetí (bias)

i) zpočátku poroste, od jisté doby ale začne klesat a vytvoří invertovanou U křivku,
ii) zpočátku bude klesat, od jisté doby ale začne růst a vytvoří U křivku,
iii) bude stále růst,
iv) bude stále klesat,
v) zůstane konstantní.

(e) *(2 b)* neredukovatelná chyba (irreducible error ϵ)

i) zpočátku poroste, od jisté doby ale začne klesat a vytvoří invertovanou U křivku,
ii) zpočátku bude klesat, od jisté doby ale začne růst a vytvoří U křivku,
iii) bude stále růst,
iv) bude stále klesat,
v) zůstane konstantní.

lambda parametrem reguluju/penalizju, s lambda = 0 si ten model dela co chce (nijak ho nepenalizuju, tudiz se prefitovava), predpokladejme ze lambda=10 je pro nas idealni
budu uvadet na prikladech lambda=0, 10, inf

-
a) cim vic pri trenovani penalizuju, tim vic roste RSS -> bude stale rust, tedy iii)
b) lambda:
   i) = 0: na trenovacich datech jsem se uspesne prefitoval,tudiz na test. datech mam velkou chybu
   ii) = 10: jsem “idealne”, nebo aspon dobre naucenej -> furt tam nejaka chyba bude, ale rozhodne mensi nez pri overfitu
   iii) = inf: pokud jde lambda k nekonecnu, tak beta koeficienty se blizi k 0, a s tema si moc neskrtnu -> velka chyba

   vysledek tudiz U krivka -> ii)

c) lambda = 0 -> prefitovavam se, variance velka.. lambda = inf -> beta je 0, variance mala, tudi klesa - iv)

d) bias = jak moc se stredni hodnota odhadu bet lisi od skutecny hodnoty, zadouci vlastnost (ne tak jako malej rozptyl ale)

- pri lambda = inf mam beta = 0, tudiz bias je velky,

- pri lambda = 0 se prefitovavam, a bias je nulovej (moje odhady jsou tak dobry, ze jejich stredni hodnota se rovna realny hodnote) -> bude stale rust

e) kdyby klesala, tak bysme ji mohli zregulovat -> klesat nebude, ale nemuzeme ani rict, ze bude rust - tady to nejde moc obecne rict, nejspravnejsi je asi v) konstantni

Robustní statistika

Robustní statistika. *(10 b)* Odhadněte dvěma různými metodami robustně rozptýlenost (scale) ze vzorku {-1.84,1.18,0.0499,-0.751,-0.00707,-2.05,-1.47,-0.0520,-0.991,-0.945}.

(a) *(2 b)* Metoda 1 (popis a aplikace na vzorek):

Median absolute deviation

- formula: $\mathrm{MAD} = \mathrm{med}\{|x_i - \mathrm{med}\{x_i\}|\}$
- breakdown point 50%
- For Normal distribution
  - ARE=0.37
  - $\hat{\sigma} = 1.4826 \cdot \mathrm{MAD}$

(b) *(2 b)* Metoda 2 (popis a aplikace na vzorek):

Sample standard deviation

- (unbiassed) formula: $\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- (biassed) formula: $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- breakdown point 0
- ARE=1 — optimal for Normal distribution

.

(c) *(2 b)* Dejte tyto odhady do vztahu s obvyklým odhadem standardní odchylky.

(d) *(2 b)* Popište kritéria, jež jsou určující pro kvalitu robustního odhadu rozptýlenosti.

.

(e) *(2 b)* Diskutujte výhody a nevýhody vámi zvolených metod podle kritérií popsaných v předchozím bodě.