Statistická analýza dat

Jméno: _____

Podpis: _____

Cvičení	
Zkouška (písemná + ústní)	≥ 25
Celkem	≥ 50
Známka	

Pokyny k vypracování: doba řešení je 120min, jasně zodpovězte pokud možno všechny otázky ze zadání, pracujte s pojmy používanými v předmětu, můžete používat kalkulátory.

Statistické minimum. *(10 b)* Zodpovězte následující otázky:

(a) *(5 b)* Uvažujte náhodný vektor $\mathbf{X}$. Definujte kovarianční a korelační matici. Jaké mají tyto matice vlastnosti? K čemu se dají použít?

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

Analýza rozptylu. *(10 b)* Odpovězte na otázky níže.

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

na hodnocení modelů (zanořených) → porovnání deviance $D = 2(l_S - l_M)$
↳ jde u lineárních - ANOVA používá F-test
tak u GLM, kde celé menno specif. test LRT
H_0 modely fungují stejně
H_A: větší model přichází zanořený model

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

předpoklad normálního rozdělení
porovnání deviance

(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		

stupně volnosti; RSS; Fstat; p-value

$0{,}05 < 0{,}164$ (α)
↓
nezamítáme H_0

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

after a test was conducted
↳ „Mělo dostatečnou šanci zamítnout nesprávnou H_0?"

Diskriminační analýza. *(10 b)* Níže diskutujte vlastnosti lineární a kvadratické diskriminační analýzy (LDA a QDA).

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

$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$

z Bayesova teorému

$$P(Y=k \mid X=x) = \frac{P(X=x \mid Y=k)\,P(Y=k)}{P(X=x)} = \frac{\pi_k \cdot f_k(x)}{\sum_{j=1}^{k} \pi_j f_j(x)}$$

$\rightarrow Norm(\mu, \sigma_k = \sigma)$

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

lineární rozhodovací hranice — kvadratická -||-

předpokladu stejných kovariančních matic u LDA - stejné Norm

(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 (LDA

LDA (QDA bude přeučovat)

(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

QDA

(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?

klasif. přesnost QDA: lin. → QDA horší než LDA, QDA poroste; nelin. → QDA → poroste, ale od začátku lepší

(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.

$f(4\%)$

$$p_k(x) = \frac{\pi_k N(\mu, \sigma^2)}{\sum_{i=1}^{k} \pi_j N(\mu_i, \sigma^2)} = \frac{0{,}8 \cdot N(0{,}1,\ 0{,}36)}{0{,}8\, N(0{,}1,\ 0{,}36) + 0{,}2 \cdot N(0,\ 0{,}36)} =$$

80% → $\mu = 10\%$

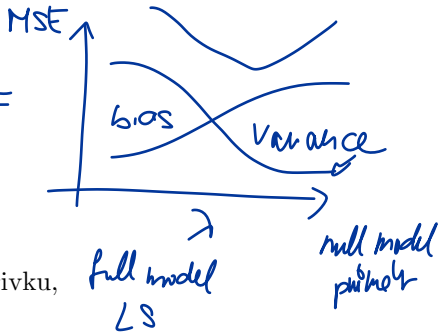
$f(x) = N(0{,}1,\ 0{,}36)$

k = Dividenda

$$\delta_k(x) = x \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) = \_\_$$

Multivariátní regrese. *(10 b)* Sestavujete multivariátní lineární model. Závisle proměnných je velký počet, hledáte model, který minimalizuje kritérium (y_i je hodnota závisle proměnné v i-tém vzorku, x_ij je hodnota j-té nezávisle proměnné v i-tém vzorku):

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



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.

$\lambda \uparrow$ ~ hladší koef k nule

(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.

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}.

3 různé metody

$MAD = med\{|x_i - med\{x_i\}|\}$

$Q = \{|x_i - x_j|, i<j\}_{q_{25}}$

$Sn = med_i\{med_j|x_i - x_j|\}$

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

$MAD = med\{|x_i - med\{x_i\}|\}$

1) odečtám medián vzorků
2) od každého vzorku odečtu
3) z to celého v abs. odečtám medián

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

(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.