Statistická analýza dat

disclaimer: žádná záruka správnosti -
nekontrolováno z žádným správným řešení

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.

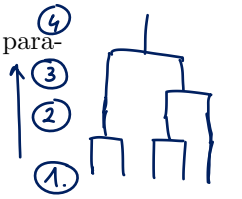
Statistické minimum. *(10 b)* Zodpovězte následující otázky:

(a) *(6 b)* Definujte věrohodnostní funkci (likelihood). K čemu se metoda maximální věrohodnosti používá? Jak se dá věrohodnosti využít při testování statistických hypotéz?

(b) *(4 b)* Vysvětlete rozdíl mezi bodovým a intervalovým statistickým odhadem parametrů.

Hierarchické shlukování. *(10 b)* Níže diskutujte vlastnosti hierarchického shlukování.

(a) *(3 b)* Formálně definujte hierarchické aglomerativní shlukování (vstupy, výstup, algoritmus a jeho parametry).



bottom-up

Vstup → data $\{x_i\}_{i=1}^{m} \in X$

výstup → taxonomie (dendrogram) → až jako vše nám dává rozdělení do shluků

bottom-up · ∀ objekt ~ vlastní cluster → iterativně shlukuje podobné clustery do méně a méně clusterů

parametry · nepotřebuje k, pro vše potřebuje účinnou, kde centroid, vzdálenostní funkci pro určování podobnosti clusterů (single linkage, complete, avg .)

(b) *(2 b)* Co to je taxonomie a dendrogram? Důsledně a detailně interpretujte jejich význam v hierarchickém shlukování (proč je taxonomie informativnější než prostý rozklad na shluky, kde lze v dendrogramu nalézt podobnost mezi objekty a jejich shluky, apod.).

taxonomie = struktura → informativnější než pouhé dělení → může říkat, kde dělit

dendrogram = binární strom

→ na osc y

(c) *(3 b)* Odhadněte složitost hierarchického aglomerativního shlukování. Důležitý je výsledek, ale i postup, kterým odhad odůvodníte.

(bottom-up)

→ musíme projít všechny vzdálenosti mezi body (v prvním kroku clustry)

→ spočítat center shluků/mých

→ spočítat vzdálenosti mezi clustery

init vzdálenosti → $O(m^2 n)$ | $O(n) \to d(x_i, x_j)$

(m-1) iterací

↳ $O(m)$ hledání nejmenší vzdálenosti

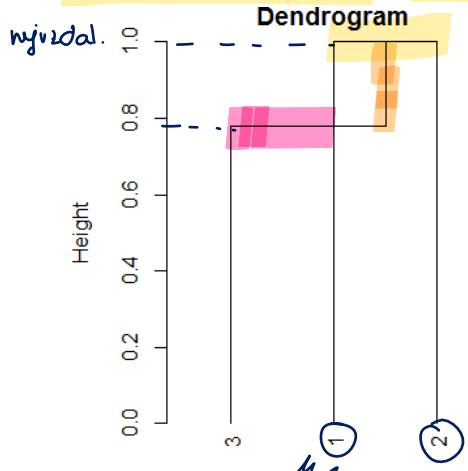
$O(m^2 n)$ → složitost výpočtu vzdálenosti

↓ počet objektů

- Problém s rozsáhlými daty, neboť dolní odhad pro složitost aglomerativního shlukování je $\Omega(m^3)$, kde m je mohutnost shlukované množiny (počet jejích prvků)
- **Nebezpečí** uvíznutí v lokálním optimu

1.Krok	m^2
2.Krok	$(m-1)^2$
3.Krok	$(m-2)^2$
4.Krok	$(m-3)^2$
m-1.Krok	$(m-(m-1))^2$

(d) *(2 b)* Na dendrogramu níže je zobrazena tzv. inverze, tj. situace, kdy podobnost v průběhu aglomerativního shlukování neklesá ale roste. Může tato situace nastat? Za jakých podmínek?



nejvzdál.

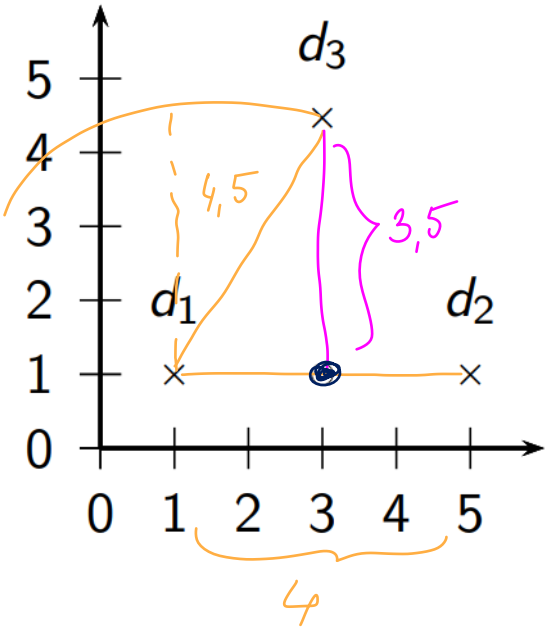
u centroidů

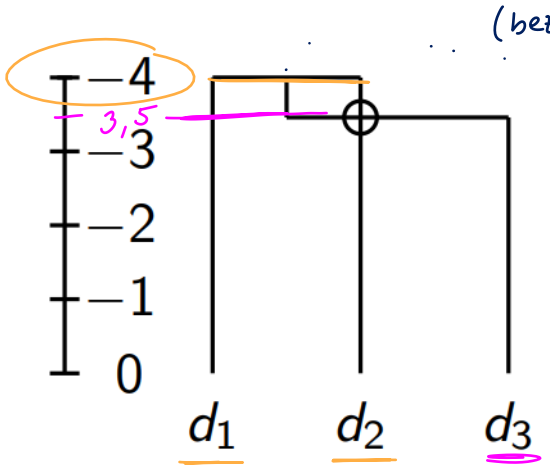
u single-linkage ne

u complete linkage taky ne

když mezi středem 1, 2 a bodem 3 → je (stejná nebo) menší vzdál než byla mezi body 1 a 2

- Below: Similarity of the first merger $(d_1 \cup d_2)$ is -4.0, similarity of second merger $((d_1 \cup d_2) \cup d_3)$ is ≈ -3.5.





než bylo max d_1 a d_2
ale k d_3 je dobliž
2, k centroidu zbývá d_3
– vypočítám centroid
1) d_1 a d_2 jsou si nejblíž

Multivariátní regrese. *(10 b)* Jste strojní zámečník a snažíte se zjistit, jak souvisí chyba obrábění hřídele s nastavením parametrů obráběcího stroje. Sestavili jste multivariátní lineární model. Model vyjadřuje vztah mezi výrobní chybou (rozdíl mezi cílovým ideálním průměrem hřídele a skutečným průměrem hřídele, ProdError) a nastevením deseti různých spojitých parametrů stroje (P1-P10). Níže je uveden výstup, který jste obdrželi:

```
summary(lm(ProdError ~ P1+P2+P3+P4+P5+P6+P7+P8+P9+P10))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2638     0.2556   1.032    0.329
P1            0.2471     0.2564   0.964    0.360
P2           -0.6112     0.1979  -3.089    0.013 *
P3            0.2728     0.2341   1.165    0.274
P4            0.1093     0.2061   0.530    0.609
P5           -0.3165     0.4674  -0.677    0.515
P6           -0.4419     0.2660  -1.661    0.131
P7            0.1244     0.3152   0.395    0.702
P8            0.2452     0.2657   0.923    0.380
P9            0.1287     0.3093   0.416    0.687
P10           0.3544     0.2956   1.199    0.261
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7549 on 9 degrees of freedom
Multiple R-squared:  0.7237,        Adjusted R-squared:  0.4168
F-statistic: 2.358 on 10 and 9 DF,  p-value: 0.1062
```

(a) *(2 b)* Rozhodněte, zda je alespoň jeden z parametrů stroje (nezávisle proměnných) užitečný pro odhad výrobní chyby. Jinými slovy, formálně rozhodněte, zda lze zamítnout $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{10} = 0$. Správně odůvodněte.

(pro malý počet p mi napovídají p-values
ale pro velký počet p) → F-stat chci aby byla co největší
ta je malá za vysoké p-value → takže nezamítáme H_0

(b) *(2 b)* Dá se z dat o modelu usoudit z kolika vzorků různých hřídelů byl model sestaven? Pokud ano, kolik jich bylo?

ze stejné velkosti
obecně pro lin reg → $df = m - p - 1 = 10$ ↑
$m - 10 - 1 = 9$
$m = 9 + 10 + 1$
$m = 20$
df: ↓ vzorky, ↑ predikátory

reality \ decision (ne)zamítám	H_0	$\neg H_0$ (zamítám)
H_0	✓	I. druh
$\neg H_0$	II. druh	✓ power of test

(c) *(2 b)* Jakým způsobem zohledníme tento počet při hodnocení užitečnosti modelu? Jak se změní chyba I druhu a chyba II druhu s rostoucím počtem vzorků pokud zachováme konstantní hladinu významnosti α?

F-statistika bere v potaz tento počet → normalizace stupni volnosti
chyba I druhu = když H_0 platí, ale já zamítám → α-hladina významnosti (5%)
chyba II druhu = H_0 neplatí, já nezamítám H_0 → ↳ nezávisí na vzorcích
↳ nesprávně zamítnutí
power testu ~ pst správného zamítnutí
↳ bude růst s ↑n
chyba II bude klesat s ↑n

(d) *(4 b)* Podrobně popište způsob, jak byste svůj model ověřili nad vzorky, které máte aktuálně k dispozici. Můžete tvořit další pomocné modely. Popište metodu validace, definujte chybovou funkci a určete, s čím budete vypočtenou chybu srovnávat.

CV

Cross-validace
chybová funkce → $AR^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$
- srovnání chyb mezi sebou – AR^2 je komparativní
pro různý výběr
prediktorů → step-wise selection

RMSE

ověřit model nad vzorky - jestli částečně dává smysl nebo?
CV dat abych mohl spočítat RMSE
průměrný RMSE se liší od nullovým modelem (průměr)
R^2, ale zase máme málo vzorků

Logistická regrese. *(10 b)* Máte data o loňské skupině studentů kurzu SAN. U každého studenta známe počet hodin, po který se připravoval na zkoušku (*hours*), jeho studijní průměr z posledního ročníku bakalářského studia (*avg*) a údaj o tom, zda přišel z programu OI či nikoli (*OI*). Cílovou binární veličinou Y je to, zda daný student získal známku A. Naučíte logistický model a získáte koeficienty $\beta_0 = -1$, $\beta_{hours} = 0.05$, $\beta_{avg} = -1$ a $\beta_{OI} = 1$.

(a) *(3 b)* Jaká je interpretace koeficientů v logistickém modelu (srovnejte s lineární regresí, kde koeficient vyjadřuje průměrnou změnu výstupu při jednotkové změně dané nezávisle proměnné a zafixování hodnot ostatních nezávisle proměnných)? Vysvětlete postupně pro β_0, β_{hours} a β_{OI}.

β_0 je nějaký posun celého modelu (+1 hodina)

$\beta_{hours} = 0{,}05 \rightarrow$ jednotkový posun v X posune Y faktorem $\exp(\beta_{hours})$

↳ naměst šance → $\exp(0{,}05)$ vzrůst o 5%

$\beta_{avg} = -1 \rightarrow \exp(-1) = 0{,}36 \leftarrow$ -||-

$\beta_{OI} = 1 \rightarrow \exp(\beta_{OI}) = \exp(1) = 2{,}72$

↳ šance je o 172% vyšší

(b) *(2 b)* Vypočítejte, jak bude podle vašeho modelu klasifikován student, který přišel z OI, připravoval se 20h a jeho *avg* bylo 2?

$$\text{logit}(p) = -1 + 0{,}05 \cdot \overset{20}{\text{hours}} + (-1)\,\overset{2}{\text{avg}} + 1 \cdot \overset{1}{OI} = z = -1$$

$$\log\left(\frac{p}{1-p}\right) \rightarrow \frac{1}{1+e^{-z}} = \frac{1}{1+e^{+1}} = \frac{1}{1+2{,}72} = 0{,}26$$

$p(Y=1)$ ↓ získat A

(c) *(2 b)* Jak dlouho by se výše uvedený student musel připravovat na zkoušku, aby měl právě 50% pravděpodobnost, že dostane známku A?

$$\frac{1}{1+e^{-z}} = \frac{1}{2} \qquad \left[\ln_e x = y \;\; e^y = x\right]$$

↳ $e^{-z} = 1$

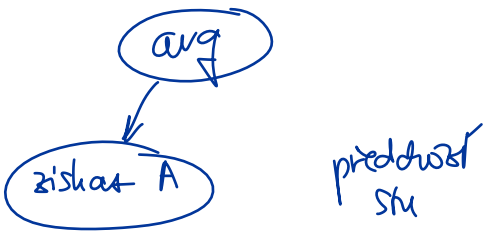
$\ln_e(1) = 0$

$z = 0$

$$-2 + \text{hours} \cdot 0{,}05 = 0$$

$$\text{hours} = \frac{2}{0{,}05} = 40 \text{ hodin}$$

(d) *(3 b)* Na uvedeném modelu/příkladu ilustrujte pojem matoucí proměnná (confounding variable). Definujte pojem, ukažte jednoduchý vztah mezi proměnnými. Model můžete libovolně upravit.



předchozí stu

Robustní statistika. *(10 b)* Odhadněte třemi různými metodami robustně polohu (location) 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: q%-trimmed; q = 10%

{-1.84, 1.18, 0.0499, -0.751, -0.00707, -2.05, -1.47, -0.0520, -0.991, -0.945}.

n = 10 vzorků
→ seřadím a škrtnu krajní vzorek
→ spočítám průměr

(b) *(2 b)* Metoda 2: q%-Windsorized

{-1.84, 1.18, 0.0499, -0.751, -0.00707, -2.05, -1.47, -0.0520, -0.991, -0.945}.

- krajní dva nahradím za bordery
- spočítám průměr

(c) *(2 b)* Metoda 3: median

{-1.84, 1.18, 0.0499, -0.751, -0.00707, -2.05, -1.47, -0.0520, -0.991, -0.945}.

- najdu prostřední hodnoty → průměr mezi

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

breakpoint = bod zlomu → kolik odlehlých hodnot model vydrží než ho jimi rozsíjou

ARE = asymptotic relative efficiency
↳ $ARE(\hat{\eta}_1, \hat{\eta}_2, P) = \frac{V_2}{V_1}$ → poměr variancí estimátorů
o kolik vzorků víc bude potřeba víc aby měli stejnou efficiency

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

nejlepší by byl Hodges-Lehmann, ale jdou nespočítám na papír
medián → 50% breakpoint
(+) ARE vyjde pro q-trimmed a q-Windsorized