# Statistická analýza dat

**Jméno:** _____

**Podpis:** _____

| | |
|---|---|
| Cvičení | |
| Zkouška (písemná + ústní) | $\geq 25$ |
| **Celkem** | $\geq 50$ |
| **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) *(4 b)* Co je to distribuční funkce? Co je to kvantil a kvantilová funkce? Co je to funkce pravděpodobnostní hustoty? Definujte pojmy formálně a uveďte jaký je mezi nimi vztah.

(b) *(4 b)* Formulujte centrální limitní větu. Kde se dá využít?

(c) *(2 b)* Definujte pojem střední hodnota a aritmetický průměr. Vysvětlete rozdíl mezi nimi.

**Redukce dimenzionality.** *(10 b)* Uvažujte problém redukce dimenzionality.

(a) *(2 b)* Definujte probléme redukce dimenzionality formálně (vstup, výstup, předpoklady, kritéria řešení).

vstup → data v prostoru $\mathcal{X}$ → dimenze $D$ ($\mathbb{R}^D$) → intrinsic dim = počet proměnných pro uspokojivou reprezentaci dat
$\{x_i\}_{i=1}^m$ → chci osy $D=d$
výstup → prostor $\mathcal{Y}$ s dim $L$ ($L<D$), zobrazení $F: \mathcal{X} \to \mathcal{Y}$ a $f: \mathcal{Y} \to \mathcal{M} \subset \mathcal{X}$
předpoklady → organizace bodů - leží přibližně na manifoldu

(b) *(2 b)* Definujte a vysvětlete pojem multidimenzionální škálování. Jde o ~~lineární~~ nebo nelineární metodu redukce dimenze?

MDS → zachovávám vzdálenosti - body blízko sebe v obou prostorech
→ min stress funkce → $\text{stress}(T, f) = \frac{\sum_{i,j}^{m} (f(\delta_{ij}) - d_{ij})^2}{\sum_{i,j}^{m} d_{ij}^2}$
prostor $\mathcal{Y}$; prostor $\mathcal{X}$; transformační funkce

(c) *(2 b)* Definujte pojem geodetická vzdálenost (geodesic distance). Jak ji lze vypočítat z dat a jaká jsou rizika tohoto výpočtu?

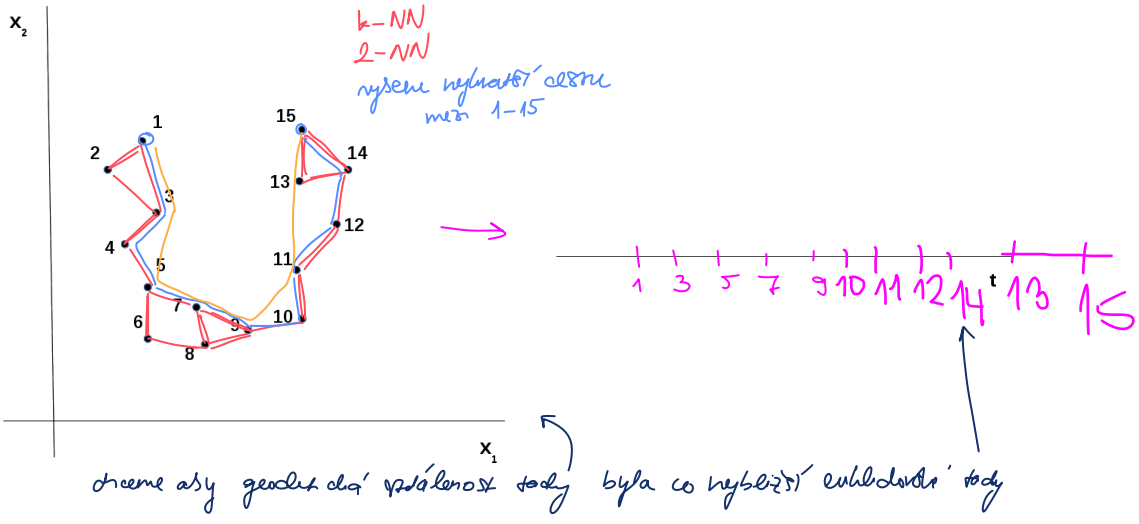
min vzdálenost cesty v grafu, která propojuje 2 body a ta samotná cesta je součástí manifoldu
výpočet z dat → Floyd's / Dijkstra's algoritmus
rizika → časově náročné, velká hustá matice výsledků

(d) *(2 b)* Napište pseudokód metody založené na multidimenzionálním škálování a geodetické vzdálenosti. Pojmenujte tuto metodu. → ISOMAP

1) urči sousedy (k-NN)
2) sestrojit graf (spojení bodu s jeho sousedy)
3) nejkratší cesta (geodesic dist.)
4) low-dim. embedding → aplikace MDS - transformace dat do $\mathcal{Y}$
↳ co nejlépe zachová nejkratší cesty v grafu

ISOMAP

(e) *(2 b)* Na obrázcích níže naznačte funkci metody popsané výše (tj. graficky naznačte způsob mapování bodů z prostoru vyšší dimenze do prostoru dimenze nižší). Může výstup vypadat i jinak než jste zakreslili?



**Multivariátní regrese.** *(10 b)* Sestavujete multivariátní lineární model. Nezávisle proměnných je velký počet, jejich relevance je odlišná, některé z těchto proměnných jsou zcela irelevantní.

(a) *(2 b)* Pojmenujte 2 základní metody, pomocí kterých lze dosáhnout smrštění (shrinkage) výsledného modelu a zapište kriteriální funkce, které tyto dvě metody minimalizují.

ridge (hřeben.) $\rightarrow$ $f = \sum_{i=1}^{m} \left(y_i - \hat{\beta}_0 - \left(\sum_{j=1}^{p} \hat{\beta}_j x_{i,j}\right)^2 + \lambda \sum_{j=1}^{p} (\hat{\beta}_j)^2\right)$

lasso $\rightarrow$ $f = \sum_{i=1}^{m} \left(y_i - \hat{\beta}_0 \sum_{j=1}^{p} \hat{\beta}_j x_{i,j}\right)^2 + \lambda \sum_{j=1}^{p} |\hat{\beta}_j|$

(b) *(2 b)* Vysvětlete, v čem se výstup výše uvedených metod bude lišit. Zdůvodněte.

ridge $\rightarrow$ hladí všechny k nule, efekt redukce rozptylu

lasso $\rightarrow$ některé nuluje úplně $\rightarrow$ selekce příznaků

(c) *(1 b)* Vyžaduje některá z výše uvedených metod předzpracování dat? Pokud ano, jaké a proč?

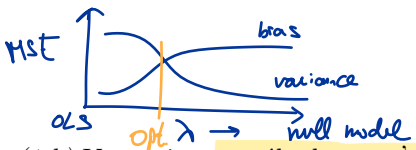
ano, obě protože narozdíl od LS není invariantní vůči škálování $\rightarrow$ nutné vydělit rozptylem (standardizace) $\tilde{x}_{i,j} = \frac{x_{i,j}}{\sqrt{\frac{1}{n}\sum (x_{i,j} - \bar{x}_j)^2}}$

(d) *(2 b)* Má některá z výše uvedených metod parametr, který je třeba nastavit? Jak se toto nastavení provede? Co parametr určuje? Popište podrobně.

$\lambda$ – reguluje regularizaci
$\lambda = 0 \rightarrow$ LS (full model)
$\lambda = \infty \rightarrow$ průměr (null model)
nastavení pomocí CV $\rightarrow$ rozdělení dat do $n$ foldů
v $\forall$ kroku $(n-1)$ trénuj
vystřídám vždy 1 test fold
udělám CV pro $\forall \alpha \rightarrow$ vyberu to $\lambda$ s nejmenší chybou / 1 $\sigma$ od min $\lambda$

(e) *(2 b)* Vysvětlete, proč mohou smrštěné modely dosáhnout nižší testovací chyby než referenční plný model vytvořený metodou nejmenších čtverců. Vysvětlení podpořte kompromisem mezi zaujetím (bias) a rozptylem (variance) obou typů modelů (plný vs smrštěný). Oba pojmy potřebné k vysvětlení definujte.

kvůli bias-variance trade off
↳ zaujetí ↳ rozptyl (flexibilita)
pro OLS bude vždy vysoká složka chyby variance (vždy přeučování $\rightarrow$ vždy přefitovaná)
u zvyšování síly regularizace nastane min (optimum) pro osu chyby $\rightarrow$ vylepší regularizace a lépe než u OLS




(f) *(1 b)* V čem jsou nevýhody smršťování oproti klasické aplikaci nejmenších čtverců?

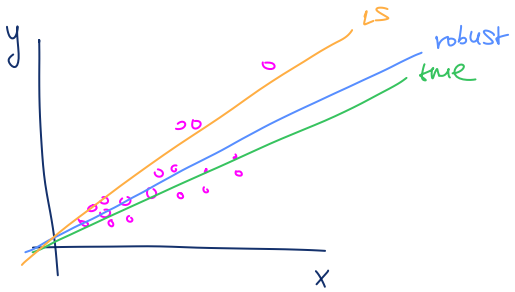
nastavování parametru $\lambda$
když bude moc velké $\lambda \rightarrow$ model bude underfitovat

**Robustní statistika.** *(10 b)* Diskutujte pojem robustní regrese. Odpovídejte na otázky níže.

(a) *(2 b)* Co to je robustní regrese a za jakých podmínek je její využití vhodné?

- regrese, která není tolik citlivá na outlery
- používá metody M-estimation jako ztrátovou funkci
- výhodné použití při outlierech v datech
- když máme jiný než Gaussovský šum ?

(b) *(1 b)* Myšlenku robustní regrese demonstrujte graficky (postačí příklad jedné závisle a jedné nezávisle proměnné, bodový graf a srovnání výstupu robustní a klasické regrese).




(c) *(2 b)* Jak lze regresní úlohu přeformulovat, aby šlo o robustní regresi? Popište alespoň 2 možnosti formulace. Kde jsou možné problémy?

nahrazení square loss funkce → Huber / Hampel

$\hat{\beta} = \arg\min_{\beta} \sum \rho(x^T\beta, y_i)$

Hampel není konvexní opt problém

To use them you need to select parameters (scale or others)
I Robust losses might have unfavourable efficiency.
I Hampel loss is not convex — difficult to optimize

(d) *(4 b)* Máte k dispozici dva párové vzorky:

$s_1 = \{293, 311, 331, 295, 337, 328, 291, 306, 323, 316\}$,

$s_2 = \{298, 322, 321, 321, 343, 331, 289, 316, 329, 322\}$.

Statisticky srovnejte oba vzorky na základě vhodné míry polohy (estimation of location, central tendency). Využijte jednu parametrickou a jednu neparametrickou, tj. robustní metodu. Pro obě metody formulujte jasný závěr. ↳ t-test Wilcoxon-signed rank test

Můžete využít těchto formulí: $T = \frac{\bar{d}-D_0}{s_d/\sqrt{n}}$, $W = \sum_{i=1}^{n}(sgn(x_i - y_i)R_i)$.

Příslušné tabulkové hodnoty: $t_{0.95,9} = 1.883$, $t_{0.975,9} = 2.262$, $t_{0.99,9} = 2.281$, $t_{0.995,9} = 3.250$;

$w_{0.95,10} = 40$, $w_{0.99,10} = 51$.

df

(e) *(1 b)* Srovnejte výhody a nevýhody obou přístupů pro danou dvojici vzorků.

(+) W → neorlivní ji outliery

(+)T → more powerfull

(−)T → pouze pro normální rozdělení

## WILCOXON TEST

kdy zamítám H0?
H0: no difference
HA: is a difference

$s_1 = \{293, 311, 331, 295, 337, 328, 291, 306, 323, 316\},$

$s_2 = \{298, 322, 321, 321, 343, 331, 289, 316, 329, 322\}.$

$|y_i - x_i| = \{5, 11, -10, 26, 6, 3, -2, 10, 6, 8\}$

$R = \quad 3 \quad 9 \quad 7 \quad 10 \quad 4 \quad 2 \quad 1 \quad 8 \quad 5 \quad 6$

$\quad\quad (-) \qquad\qquad (-)$

$\Sigma_+ =$

$\Sigma_- =$

$W = \Sigma R_i \cdot \text{sign} = 39$ při 37 → nelze zamítnout $H_0$ ↳ jsou stejné

37 < w_critical --> co to znamená?

## T-TEST

$\bar{d} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = 8{,}5 \longrightarrow T = \frac{8{,}5 - D_0 = 0}{\frac{7{,}58}{\sqrt{n}}} = 3{,}541667$

$n = 10$

$s_d = \sqrt{\frac{\Sigma(|y_i - x_i| - \bar{d})^2}{n}} = \sqrt{\frac{(-24)^2}{10}} = 7{,}58$

Příslušné tabulkové hodnoty: $t_{0.95,9} = 1.883$, $t_{0.975,9} = 2.262$, $t_{0.99,9} = 2.281$, $t_{0.995,9} = 3.250$;

$w_{0.95,10} = 40$, $w_{0.99,10} = 51$.

df

jako $\alpha$ 2

moje $t = 3{,}54 >$ než ∀ kritické $t_{\alpha,9}$ → $H_0$ zamítám s $\alpha = 0{,}01 = 1\%$

↳ jsou různé populace $s_1$ a $s_2$

| s1 | s2 | rozdil | Wilcoxon1 | Wilcoxon2 | T-test | | rozdil-mean |
|---|---|---|---|---|---|---|---|
| 293 | 298 | 5 | 3 | 3 | mean | 8,5 | -3,5 |
| 311 | 322 | 11 | 9 | 9 | std_excel | 6,835366 | 2,5 |
| 331 | 321 | -10 | -7 | -8 | std_my | 7,589466 | -18,5 |
| 295 | 321 | 26 | 10 | 10 | | | 17,5 |
| 337 | 343 | 6 | 4 | 4 | T stat | **3,541667** | -2,5 |
| 328 | 331 | 3 | 2 | 2 | | | -5,5 |
| 291 | 289 | -2 | -1 | -1 | | | -10,5 |
| 306 | 316 | 10 | 8 | 7 | | | 1,5 |
| 323 | 329 | 6 | 5 | 5 | | | -2,5 |
| 316 | 322 | 6 | 6 | 6 | | | -2,5 |
| | | **W stat** | **39** | **37** | | | -24 |

**Power analysis.** *(10 b)* Odpovězte na otázky níže.

| | zamítnu | nezamítnu |
|---|---|---|
| | $\neg H_0$ | $H_0$ |
| $H_0$ | I = $\alpha$ | ✓ |
| $\neg H_0$ | ✓ | II = $\beta$ |

správně zamítnutí $1-\beta$

(a) *(2 b)* Vysvětlete pojem síla statistického testu.

pravděpodobnost správného zamítnutí

$P(\text{reject } H_0 \mid H_A \text{ is true}) = 1-\beta$

↳ II. druh chyby

$P(\text{not reject } H_0 \mid H_A \text{ true}) = \beta$

(b) *(3 b)* Na čem síla testu závisí a jak (uveďte tři faktory, pro každý faktor popište typ vztahu)?

→ velikost vzorku → přímá úměra

→ hladina významnosti $\alpha$ → přímá

→ effect size d (velikost pozorovaného rozdílu) → čím větší rozdíl d - menší síla testu

(c) *(5 b)* Kolik účastníků testu je třeba pozvat na testování, pokud chcete mít 90% šanci vidět problémy, které postihují 30% všech uživatelů? Napište a popište rovnici pro výpočet velikosti vzorku pro objevování problémů. Proveďte výpočet.

$$n = \frac{\ln(1-X)}{\ln(1-Y)}$$

X % problémů

Y % uživatelů, kteří objevují

k detekci X % problémů, které objevují Y % uživatelů

X % chance of detecting

$$n = \frac{\ln(1-0{,}9)}{\ln(1-0{,}3)} = \frac{-2{,}303}{-0{,}36} = 6{,}4 \rightarrow 7 \text{ účastníků}$$