

# Statistická analýza dat

Jméno: \_\_\_\_\_

Podpis: \_\_\_\_\_

Cvičení	
Zkouška (písemná + ústní)	$\geq 25$
<b>Celkem</b>	$\geq 50$
<b>Známka</b>	

**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.

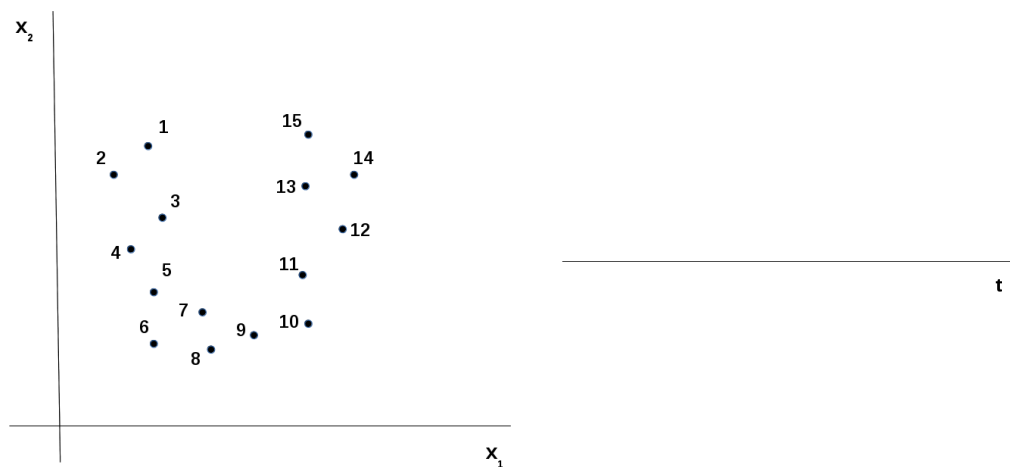
(a) (2 b) Definujte problémy redukce dimenzionality formálně (vstup, výstup, předpoklady, kritéria řešení).

(b) (2 b) Definujte a vysvětlete pojem multidimenzionální škálování.

(c) (2 b) Definujte pojem geodetická vzdálenost (geodesic distance). Popište možnosti jejího využití při redukcí dimenze.

(d) (2 b) Napište pseudokód metody založené na multidimenzionálním škálování a geodesické vzdálenosti. Pojmenujte tuto metodu.

(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. Závisle 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 запиšte kritériální funkce, které tyto dvě metody minimalizují.

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

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

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

(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.

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

**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?

(b) (1 b) Myšlenku robustní regrese demonstруйте graficky (postačí příklad jedné závislé a jedné nezávislé 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.

(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 srovnajte 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.

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

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

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

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

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

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

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

(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.