Statistická analýza dat

Jméno: _____

Podpis: _____

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

Pokyny k vypracování: doba řešení je 150min, jasně zodpovězte pokud možno všechny otázky ze zadání, pracujte s pojmy používanými v předmětu, můžete používat kalkulátory.

Statistické minimum. *(10 b)* Zodpovězte otázky níže. V každé z podotázek je právě jedna odpověď správná.

(a) *(2 b)* Pokud je p-hodnota dané testovací statistiky 0.03, pak můžeme:

i) předpokládat, že extrémní hodnota testovací statistiky je způsobena náhodou,

ii) dovodit, že za předpokladu platnosti nulové hypotézy existuje 3% pravděpodobnost získání extrémnější testové statistiky než jsme pozorovali,

iii) předpokládat, že nulová hypotéza platí s pravděpodobností 0.03,

iv) přijmout nulovou hypotézu na hladině významnosti 0.05.

(b) *(2 b)* Pokud chcete testovat hypotézu o střední hodnotě z populace s šikmým (nesymetrickým) rozdělením, měli byste:

i) použít stratifikovaný vzorek,

ii) použít velkou hladinu významnosti α,

iii) mít všechny vzorky předem kategorizované jako úspěšné nebo neúspěšné,

iv) získat vzorek s velikostí větší než 30.

(c) *(2 b)* Předpoklad nezávislosti při statistickém modelování znamená, že:

i) nezávisle proměnné nejsou korelované,

ii) chyby modelu spolu vzájemně nesouvisí,

iii) nulová hypotéza by neměla být zamítnuta,

iv) residua modelu nejsou nezávislá.

(d) *(2 b)* Víte, že výšky žen jsou normálně rozděleny se střední hodnotou 160cm. Která z následujících pravděpodností je nejvyšší? Pravděpodobnost náhodného výběru:

i) jedné ženy s výškou mezi 155 a 165 centimetry,

ii) 15ti žen s průměrnou výškou mezi 155 a 165 centimetry,

iii) 100 žen s průměrnou výškou mezi 155 a 165 centimetry.

iv) Tři výše uvedené jevy mají stejnou pravděpodobnost.

(e) *(2 b)* Chcete odhadnout podíl českých občanů, kteří podporují očkování proti Covidu. Jak velký vzorek (velikost značena n) by byl potřeba k zajištění 95% pravděpodobnosti, že skutečný podíl populace p nebude dále než 3 procentní body od podílu zjištěného ze vzorku? (tipy: počet příznivců očkování v našem vzorku bude sledovat binomické rozdělení se střední hodnotou np a směrodatnou odchylkou $\sqrt{np(1-p)}$, měli byste pracovat s konzervativním odhadem, že $p = 0.5$, aproximujte binomické rozdělení normálním, víte, že $z_{0,025} = 1.96$).

i) 512,

ii) 1068,

iii) 2506,

iv) 3152,

v) 6304.

Multivariátní regrese. *(10 b)* Sestavujete multivariátní lineární model. Nezávisle proměnných je velký počet, vyšší než je počet trénovacích vzorků, které máte k dispozici. Formálně tedy platí: $y \in R$, $\mathbf{x} \in R^p$, $T = \{\langle \mathbf{x_i}, y_i \rangle\}_{i=1}^m$, $p > m$, předpokládáme $y = \mathbf{x}^T\beta + \epsilon$.

(a) *(2 b)* Odhadněte a komentujte výsledek přímé aplikace metody nejmenších čtverců na množinu T.

(b) *(2 b)* Popište co nejpodrobněji aplikaci metody výběru podmnožiny příznaků (subset selection) v souvislosti s tvorbou optimálního lineárního regresního modelu.

(c) *(2 b)* Vysvětlete jaký problém ve srovnání s přístupem ad a) metoda ad b) řeší a zároveň vysvětlete, jaké úskalí naopak vzniká.

(d) *(2 b)* Popište přístup k selekci příznaků a regresi založený na dopředné a zpětné krokové selekci příznaků (forward and backward stepwise regression). Na základě jakého kritéria modely srovnáváme? Odhadněte složitost tohoto způsobu selekce příznaků.

(e) *(2 b)* Který z dvojice přístupů ad d) byste upřednostnili v této konkrétní úloze a proč? Srovnejte také oba přístupy s metodou subset selection.

Logistická regrese. *(10 b)* Diskutujte logistickou regresi.

(a) *(2 b)* Popište za jakých podmínek je využití logistické regrese vhodné (definujte úlohu včetně jejích variant daných různými typy proměnných).

(b) *(3 b)* Zapište definiční vztah logistické regrese. Vysvětlete význam proměnných. Srovnejte s lineární regresí (je logistická regrese lineární či nelineární, kdy selhává jedna metoda a kdy naopak druhá).

(c) *(2 b)* Jaká je interpretace koeficientů v logistickém modelu (srovnejte s lineární regresí, kde koeficient vyjadřuje průměrnou změnu výstupu při jednotkové změně dané závisle proměnné a zafixování hodnot ostatních závisle proměnných)?

(d) *(3 b)* Na konkrétním příkladu logistické regrese ilustrujte pojem matoucí proměnná (confounding variable).

Robustní statistika. *(10 b)* Předpokládejte, že sestavujete multivariátní lineární model $\hat{y} = \mathbf{x}^T \hat{\beta}$. Vaším cílem je odhadnout vektor parametrů modelu β z trénovacích dat.

(a) *(2 b)* Jak tento odhad provedete pokud předpokládáte, že skutečný vztah mezi závisle proměnnou Y a vektorem nezávisle proměnných $\mathbf{X}$ lze popsat generativním modelem $Y = \mathbf{X}^T \beta + \epsilon$, kde ϵ je gaussovský šum $N(0, 1)$? Metodu pojmenujte a zapište její kritérium.

(b) *(2 b)* Zdůvodněte, proč je daná metoda v dané situaci optimální.

(c) *(2 b)* Jaké metody byste použili, pokud vztah zůstane nezměněn, ale šum ϵ bude směsí $\alpha N(0, 1) + (1 - \alpha)N(\gamma, 1)$, kde α je blízké 1 a γ je neznámé a konečné?

(d) *(2 b)* Dojde vzhledem k předchozí situaci ke změně volby metody v případě, že ϵ je laplaceovský šum $Laplace(0, 1)$. Pokud ano, popište k jaké a proč.

(e) *(2 b)* Proč nejsou metody popsané v bodě c) a d) vhodné i v situaci ad a)?

Shlukování. *(10 b)* Uvažujte problém shlukování.

(a) *(1 b)* Definujte problém shlukování slovně.

(b) *(2 b)* Formálně definujte shlukování jako optimalizační problém. Zařaďte jej do správné třídy složitosti. Zdůvodněte.

(c) *(3 b)* Je spektrální shlukování příkladem shlukovacího algoritmu, který shlukování jako optimalizační problém formuluje a řeší? Vysvětlete.

(d) *(2 b)* Používá spektrální shlukování kernel funkci? V čem se její použití liší od použití v kernel k-means?

(e) *(2 b)* Popište, jak lze ve spektrálním shlukování nalézt optimální počet shluků k, pokud k není předem známo.