test 3 2024

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

(a) (4 b) Vysvětlete základní pojmy statistického testování hypotéz: nulová hypotéza, alternativní hypotéza, jednostranný a dvojstranný test. Demonstrujte na příkladu.

(b) (3b) Ukažte, že maximálně věrohodným odhadem střední hodnoty normálního rozdělení s pravděpodobnostní ze vzorku (x1,...,xm} je aritmetický průměr.

$$f(x;\mu,\sigma^2) = rac{1}{\sigma\sqrt{2\pi}}e^{-rac{1}{2}\left(rac{x-\mu}{\sigma}
ight)^2}.$$

(c) (3 b) Máte k dispozici velmi velkou množinu měření výšek dospělých mužů v České republice, tužku, papír a klobouk. Víte, že střední výška českého muže je 178 cm. Vaším úkolem je vygenerovat vzorek o velikosti 10, který bude mít t-rozdělení s 5 stupni volnosti. Jak to uděláte? Jaké bude mít rozdělení parametry? Proceduru popište co nejpřesněji

2) Multivariátní regrese. (10 b) Odpovězte na otázky níže.
(a) (4 b) Chcete statisticky srovnat přesnost dvou lineárně regresních modelů s odlišným počtem parametrů přímo z trénovacích dat (tj. modely nemůžete testovat na testovacích/validačních datech, nemůžete použít ani křížovou validaci). Jak to uděláte? Popište podrobně, definujte nutné pojmy.
(b) (2 b) Jak rozhodnete, zda lepší z modelů výše, resp. alespoň jeden z těchto modelů, je užitečný? Formulujte jako statistickou hypotézu a naznačte její test.
(c) (2 b) Vysvětlete význam pojmu matoucí proměnná (confounding variable). Uveďte příklad a naznačte její možný vliv na multivariátní model.
(d) (2b) Vysvětlete zda a jak lze v regresním modelu použít kvalitativní nezávisle proměnnou (např. obor studia={informatika, elektrotechnika, biomedicína}). Jaký je význam koeficientů, které se k této proměnné váží?

3) 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 Bo=-1, Bhours = 0.05, Baug=-1 a Bol=1.
(a) (26) Vypočítejte, jak bude podle vašeho modelu klasifikován student, který přišel z OI, připravoval se 20h a jeho avg bylo 2?
(b) (2b) Jak dlouho by se výše uvedený student musel připravovat na zkoušku, aby měl právě 50% pravděpodobnost, že dostane známku A?
(c) (2 b) Vysvětlete význam koeficientu Bol pomocí poměru šancí (odds ratio). Definujte pojem a dejte do souvislosti s Bol
(d) (2 b) Zaveďte novou proměnnou, která bude pro stávající proměnné matoucí (confounding variable). Vysvětlete, proč je právě tato proměnná matoucí.)

(e) (2 b) Vysvětlete, zda je logistická regrese citlivější nebo méně citlivá na odlehlé hodnoty než metoda LDA
4) Detekce anomálií. (10 b) Odpovězte na otázky níže.
(a) (2 b) Jak je definována anomálie? Uveďte dvě definice a vysvětlete, jak ovlivňují proces detekce anomálií.
(b) (1 b) Dejte do vztahu pojem anomálie a odlehlá hodnota (outlier).
(c) (1 b) Co je obecně největší problém u detekce anomálií?

(d) (2 b) V čem je výhoda a nevýhoda detekčních přístupů založených na k nejbližších sousedech? Srovnejte s parametrickými metodami založenými na modelech.
(e) (2 b) Jaké metody jsou navržené pro detekci anomálií ve vyskokých dimenzích? Stručně popište.
5)

Shlukování. $(10\ b)$ Zodpovězte otázky níže. V každé z podotázek je právě jedna odpověď správná.

- (a) (2 b) Která z následujících optimalizačních formulací shlukování je správná (Q je strategie, $\mathcal{X} = \{x_1, \dots, x_m\}$ $Q: \mathcal{X} \to D$ objektů, D je množina rozhodnutí, W je ztrátová funkce)? Hledáme optimální strategii
 - $\operatorname{argmin}_{q} \sum_{i=1}^{m} W(x_{i}, q(x_{i})),$
 - $\lim_{n\to\infty} \prod_{i=1}^m W(x_i, q(x_i)),$
 - (iii) $\underset{a}{\operatorname{argmax}} \sum_{i=1}^{m} W(x_i, q(x_i)),$
 - $\lim_{q} \operatorname{argmax} \prod_{i=1}^{m} W(x_i, q(x_i)).$
- (b) (2 b) Algoritmus k-středů (k-means) nemá následující vlastnost:
 - i) zaručeně konverguje, vždy vrátí řešení,
 - ii) je hladový, garantuje pouze nalezení suboptimálního rozkladu,
 - (iii) pro jakoukoli inicializaci vrátí stejný rozklad,
 - iv) rozklad na shluky je vždy disjunktní.
- (c) (2 b) Za soft (pravděpodobnostní) shlukování lze označit algoritmus:
 - i) k-středů,
 - (ii) EM GMM shlukování (optimalizace parametrů směsi normálních rozdělení EM algoritmem),
 - iii) aglomerativní hierarchické shlukování,
 - iv) DBSCAN založený na hustotě objektů v příznakovém prostoru.
- (d) O spektrálním shlukování lze prohlásit:
 - i) je exaktním polynomiálním řešením problému minimálního řezu grafu vyjadřujícího podobnost mezi shlukovanými objekty,
 - je přibližným řešením problému minimálního vyváženého řezu grafu vyjadřujícího podobnost mezi shlukovanými objekty,
 - iii) využívá algoritmy lineární algebry, původní příznakový prostor transformuje pouze lineárně,
 - (iv)) díky efektivní spektrální analýze grafů je výpočetně méně náročné než algoritmus k-středů.
- (e) Z vlastních čísel nenormalizovaného Laplaciánu grafu podobnosti použitého ve spektrálním shlukování nelze:
 - i) přímo určit počet uzlů grafu/shlukovaných objektů,
 - ii) přímo určit počet souvislých komponent grafu (souvislý graf má jedno nulové vlastní číslo, počet nulových vlastních čísel se rovná počtu souvislých komponent grafu),
 - (iii) dentifikovat odlehlé objekty (outliery),
 - iv) odhadnout přírozený počet shluků v množině objektů, kterou graf podobnosti popisuje.