test 1 2024

1) Multivariátní regrese. (10 b) Máte k dispozici sadu měření pro 1,000 lidí. U každého člověka bylo změřeno velké množství proměnných typu výška, váha, BMI, krevní tlak, atd. Chcete zjistit, které z těchto proměnných mají vliv na účinnost léku, který byl pozorovaným lidem podán v různém množství. Množství podaného léku a jeho účinnost jsou v datech také zaznamenány. Chcete také co nejpřesněji předpovídat účinnost léku u budoucích pacientů. Použijete lineární regresi ve varianté LASSO, volba metody je tedy součástí zadání.
(a) (1 b) Zapište optimalizační kritérium metody LASSO. Pojmenujte jeho proměnné.
(b) (2 b) Vysvětlete, proč je metoda LASSO vhodná pro výběr relevantních proměnných. Proč překonává například výběr všech signifikantních 8 koeficientů v běžné lineární regresi? A co to znamená, že jej překoná?
(c) (1 b) Vysvětlete, proč metoda LASSO v této úloze obvykle překoná běžnou lineární regresi z pohledu přesnosti predikce účinnosti léku u budoucích pacientů.
(d) (2 b) Metoda LASSO vyžaduje v této úloze předzpracování příznaků. Vysvětlete proč. Zapište toto předzpracování matematicky.

(e) (2 b) K ověření výkonu metody LASSO a také ke stanovení jejích parametrů použijete train/test/validate rozdělení dat. Popište co nejpodrobněji, k čemu jednotlivé množiny použijete a proč postup nelze zjednodušit
(f) (2 b) Nyní se soustřeďte na předzpracování příznaků diskutované výše. Jak jej přesně provedete v souvislosti s dělením na train/test/validate? Vysvětlete.
2) Statistické minimum. (10 b) Zodpovězte následující otázky:
(a) (3 b) Formálně definujte binomické rozdělení. Vysvětlete jeho parametry. Uveďte příklad jeho použití. Jakou pravděpodobnostní distribuci byste použili k jeho aproximaci, pokud je počet pokusů velký?

(b) (4 b) Mějme sáček, který obsahuje 3 míčky. Každý míček je buď červený nebo modrý. Počet modrých míčků, popsaný parametrem K, tedy může být 0, 1, 2 nebo 3. Mohli jste si náhodně s opakováním vybrat ze sáčku 4 míčky, dostali jste 3 modré míčky a 1 červený. Odhadněte hodnotu K pomocí metody maximální věrohodnosti (postup je nejdůležitější, záleží také na samotné hodnotě K, ale najděte ji tak formálně, jak jen můžete).
(c) (3 b) Jiřinu svědí kůže. Existuje test na alergii na kočky, ale tento test není vždy správný. U lidí, kteří opravdu alergii mají, test říká "Ano" v 80% případů. U lidí, kteří alergii nemají, test říká "Ano" v 10% případů. Pokud má alergii 1% populace a Jiřinin test říká "Ano", jaká je pravděpodobnost, že Jiřina skutečně má alergii?
3) Redukce dimenzionality. (10 b) Uvažujte metodu kernel PCA. (a) (2 b) Vysvětlete základní myšlenku kernel metod.

(b) (1 b) Uveďte příklad dvou známých kernel funkcí. Definujte je matematicky.	
(c) (2 b) Slovně (obrázky) popište princip metody kernel PCA	
(d) (2 b) Definujte metodu matematicky.	

(e) (1 b) Srovnejte s metodou PCA z pohledu funkce a výpočetní složitosti.
(f) (2 b) Srovnejte s metodou Isomap (princip, výhody a nevýhody, kdy byste použili každou z metod).
4)

Zobecněné lineární modely. (10 b) Zodpovězte otázky o GLM níže. V každé z podotázek je právě jedna odpověď správná.

- (a) (2 b) Uvažujme lineární regresní model: $\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$, $y_i \sim \mathcal{N}(\mu_i, \sigma)$. Lze jej interpretovat jako GLM s link funkcí q:
 - i) $g(\mu_i) = 0$,
 - (ii) $g(\mu_i) = \mu_i$,
 - iii) $g(\mu_i) = \log \mu_i$,
 - iv) $g(\mu_i) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}(\frac{x_i-\mu_i}{\sigma})^2}$
- (b) (2 b) Uvažujme R data frame d. Ten obsahuje nezávisle proměnné X_1,\ldots,X_{100} a závisle proměnnou Y. d má 1,000 řádků. Volání $lm(Y\sim.,d)$ a $glm(Y\sim.,d,family=gaussian)$:
 - 🗓 definují stejnou úlohu, získáme identický odhad parametrů v identickém čase,
 - (ii) definují stejnou úlohu, použijeme ale odlišné optimalizátory (lm rychlejší a výhodnější),
 - iii) definují stejnou úlohu, lm ale pracuje s přibližným kritériem nejmenších čtverců (lm rychlejší, ale méně přesný odhad parametrů modelu),
 - iv) definují jiné úlohy, glm postihne i nelineární vztahy mezi proměnnými, výsledky volání budou odlišné.
- (c) (2 b) Binární logistická regrese (závisle proměnná nabývá hodnot0a 1):
 - (j) je variantou GLM, má lineární prediktor, logit link funkci a Bernoulliho pravděpodobnostní model,
 - ii) není variantou GLM, nejde o lineární regresi ani její zobecnění,
 - iii) není variantou GLM, nelze pro ni nalézt kanonickou link funkci,
 - iv) není variantou GLM, závisle proměnná je diskrétní a nemá exponenciální rozdělení.
- (d) (2 b) Pro modelování count proměnných jako je například denní počet selhání stroje je obecné vhodná Poissonova regrese. Je výhodná zejména pokud:
 - (j) jsou predikované počty nízké (pro vysoké počty bychom ji mohli nahradit lineární regresí),
 - ii) jsou predikované počty vysoké (pro nízké počty bychom ji mohli nahradit lineární regresí),
 - iii) rozptyl v počtech se nemění s jejich střední hodnotou (jedna z vlastností Poissonova rozdělení),
 - v počtech roste exponenciálně s jejich střední hodnotou (jedna z vlastností Poissonova rozdělení).
- (e) $(2\ b)$ GLM obvykle hodnotíme pomocí deviance. O tomto kritériu **nelze** prohlásit:
 - i) počítá odchylku log-věrohodnosti modelu od modelu saturovaného, tj. nejpřesnějšího možného,
 - ii) je použitelné pro statistické porovnání/testování zanořených GLM,
- V
- (iii) zohledňuje počet parametrů, které model má,
- iv) čím menší je deviance modelu, tím věrněji se model přizpůsobuje datům použitým pro hodnocení.

5) Shlukování. (10 b)

Máte k dispozici výsledek krevního testu u šesti osob. Výsledkem testu je reálné číslo, zde konkrétně jsou výsledky: 10, 12, 13, 17, 18, 21. Mezi osobami jsou zdraví i nemocní lidé, nevíte ale, kdo je kdo. Víte pouze, že výsledek testu má v populaci zdravých i nemocných normální rozdělení. Víte také, že nemocní lidé mají vyšší hodnoty testu a je jich stejně jako zdravých. O statistických vlastnostech krevního testu i rozdělení osob do skupin chcete rozhodnout pomocí EM shlukování.

(a) (1 b) Proveďte inicializaci EM shlukovacího algoritmu. Pracujte s čísly ze zadání.

(b) (3 b) Proveďte první krok EM algoritmu (v závislosti na způsobu inicializace začněte buď E krokem výstupem bude pravděpodobnostní rozdělení osob do skupin, nebo M krokem budou konkrétní parametry modelu). K orientačnímu odhadu pravděpodobností můžete využít standardní normální rozdělení na obrázku níže. Výpočet nemusí být detailní pro všech 6 vzorků.
(c) (3 b) Proveďte druhý krok EM algoritmu (v závislosti na předchozím kroku budete pokračovat bud M krokem - výstupem budou konkrétní parametry modelu, nebo E krokem -výstupem bude pravděpodobnostní rozdělení osob do skupin).
(d) (3 b) Graficky a slovně popište očekávaný výsledek EM algoritmu po doběhnutí (jak bude vypadat model, jak budou objekty rozděleny do shluků).