

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 kalkulatory.

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

- (a) (2 b) Call centrum přijímá v průměru 180 nezávislých hovorů za hodinu, 24 hodin denně. Počet přijatých hovorů během kterékoli minuty je příkladem náhodné veličiny s rozdělením:
- i) normálním,
 - ii) binomickým,
 - iii) uniformním,
 - iv) Poissonovým,
 - v) žádným z výše uvedených.
- (b) (2 b) Uvažujeme dvě náhodné události s nenulovou pravděpodobností výskytu. Víme-li, že události nemohou nastat současně, tak tyto dvě události:
- i) musí být nezávislé,
 - ii) mohou být nezávislé,
 - iii) nemohou být nezávislé,
 - iv) neplatí nic z výše uvedeného.
- (c) (2 b) Předpokládejme, že rychlosti vozidel na konkrétním místě rychlostní silnice mají normální rozdělení s průměrem rovným 100 km/h a směrodatnou odchylkou rovnou 8 km/h. Jaké je z-skóre pro rychlost 94 km/h?
- i) -0.75,
 - ii) 0.75,
 - iii) -6,
 - iv) 6,
 - v) žádné z výše uvedených.
- (d) (2 b) Krabice s 8 kuličkami obsahuje 4 červené, 2 zelené a 2 modré kuličky. Pokud vyberete dvě kuličky bez opakování, jaká je pravděpodobnost, že vyberete červenou kuličku následovanou kuličkou zelenou?
- i) 0.0625,
 - ii) 0.1250,
 - iii) 0.1429,
 - iv) 0.6,
 - v) žádná z výše uvedených.
- (e) (2 b) 500 náhodných vzorků o velikosti $n=900$ bylo vzorkováno z velké populace, ve které je 10% leváků. Pro každý vzorek zjišťujeme podíl leváků, následně vykreslíme histogram těchto 500 podílů. Který z intervalů obsáhne přibližně 68% hodnot v histogramu?
- i) 0.1 ± 0.01 ,
 - ii) 0.1 ± 0.0134 ,
 - iii) 0.1 ± 0.0167 ,
 - iv) 0.1 ± 0.02 ,
 - v) žádný z výše uvedených.

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 variantě 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 β 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 soustředte na předzpracování příznaků diskutované výše. Jak jej přesně provedete v souvislosti s dělením na train/test/validate? Vysvětlete.

Diskriminační analýza. (10 b) Řešíte binární klasifikační úlohu, ve které je počet pozitivních příkladů 50 krát nižší než je počet negativních příkladů. Použili jste lineární diskriminační analýzu (LDA) a vytvořili jste klasifikátor. Vaším úkolem je vyhodnotit, jak dobře váš klasifikátor odděluje třídy.

(a) (1 b) Definujte matici záměn (confusion matrix) sloužící ke znázornění výkonu klasifikátoru. Pojmenujte její jednotlivé položky a vysvětlete jejich význam.

(b) (1 b) Jak z uvedené matice vypočtete klasifikační přesnost? V čem je úskalí použití klasifikační přesnosti v naší úloze?

(c) (2 b) Sestavili jste konkrétní matici záměn, váš klasifikátor neklasifikuje žádný z 1,000 testovacích příkladů do pozitivní třídy. Jak k tomu mohlo dojít? Jak svůj model upravíte, aby byl počet pozitivních predikcí nenulový? Jaký to bude mít vliv na klasifikační přesnost?

(d) (1 b) Co uděláte pokud pro vás nesprávné označení pozitivního příkladu za negativní nese dvojnásobnou ztrátu oproti nesprávnému označení negativního příkladu za pozitivní?

(e) (2 b) Na základě dříve zavedené matice záměn definujte ROC křivku.

(f) (1 b) Je ROC křivka použitelná v naší úloze? Vysvětlete proč.

- (g) (2 b) Vysvětlete pojem plocha pod ROC křivkou (AUC). Co definuje? K čemu lze tuto plochu využít? Nakreslete obrázek s několika různými křivkami a vysvětlete jaké klasifikátory charakterizují. Nezapomeňte popsat osy.

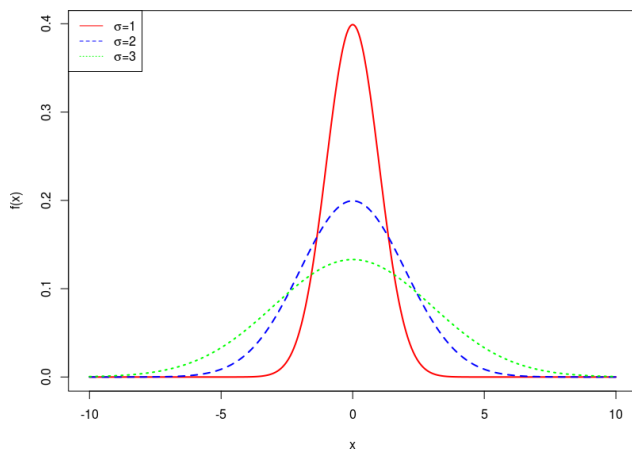
Redukce dimenzionality. (10 b) Uvažujte metodu kernel PCA.

- (a) (2 b) Co je to kernel funkce? Jaké má argumenty a co je jejím výstupem? Jak ji interpretujeme v kernel PCA?
- (b) (1 b) Co je to kernel matice? Jaké má rozměry? Kterou matici stejných rozměrů nahrazuje ve srovnání s PCA?
- (c) (2 b) Uveďte příklad dvou známých kernel funkcí. Definujte je matematicky. Odvoďte/odhadněte dimenzi transformovaného prostoru po aplikaci těchto funkcí pokud je dimenze vstupního prostoru 2.
- (d) (2 b) Srovnejte kernel PCA s metodou PCA z pohledu funkce. Kdy je kernel PCA vhodná? Jaké má oproti PCA nevýhody?
- (e) (2 b) Srovnejte kernel PCA s metodou PCA z pohledu výpočetní složitosti.
- (f) (1 b) Srovnejte kernel PCA s metodou Locally Linear Embedding (LLE). Je mezi nimi nějaký vztah?

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 o polovinu méně než zdravých. O statistických vlastnostech krevního testu i rozdělení osob do skupin chcete rozhodnout pomocí EM shlukování.

- (a) (1 b) Proveďte inicializaci EM shlukovacího algoritmu nastavením počátečního přiřazení ke shlukům. Pracujte s čísly ze zadání.

- (b) (3 b) Proveďte první krok EM algoritmu, vzhledem k inicializaci začnete M krokem a výstupem 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, půjde o E krok a 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).