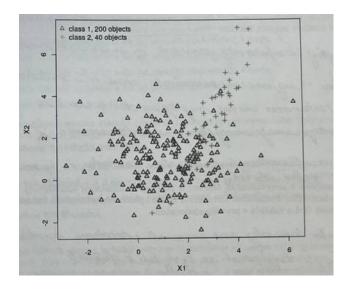
test 2 2024

1) Statistické minimum. Zodpovězte následující otázky.		
(a) (5 b) Testujete hypotézu pomocí t-testu. Na základě zjištěné hodnoty t-statistiky jste došli k p-hodnotě 0.009. Pracujete s hladinou významnosti a=0.01. Zhodnoťte následující tvrzení (pravda/nepravda +zdůvodnění).		
(i) (1 b) Test je dostatečně silným důkazem neplatnosti nulové hypotézy, nulovou hypotézu zamítneme.		
(ii) (1 b) Za předpokladu platnosti nulové hypotézy můžeme s pravděpodobností 0.009 pozorovat t- statistiku jako v našem konkrétním testu nebo ještě extrémnější.		
(iii) (1 b) Alternativní hypotéza platí s pravděpodobností 0.009.		
(iv) (1 b) Nulová hypotéza platí s pravděpodobností 0.009.		
(v) (1 b) Alternativní hypotéza je pravděpodobnější než nulová.		
(b) (5 b) Procházíte zkouškovým testem, který obsahuje 5 otázek. Pro každou otázku můžete zaškrtnout právě jednu správnou odpověď z 5 předvyplněných odpovědí. Víte, že pouze jedna z těchto 5 odpovědí je skutečně správná. Do testu vstupujete s nulovou znalostí a odpovědi budete volit náhodně. Jaká je pravděpodobnost, že projdete úspěšně testem, pokud je třeba		

2) Klasifikace. (10 b) Máte k dispozici uměle generovaná trénovací data uvedená v obrázku níže.

správně zodpovědět nejméně 3 otázky?

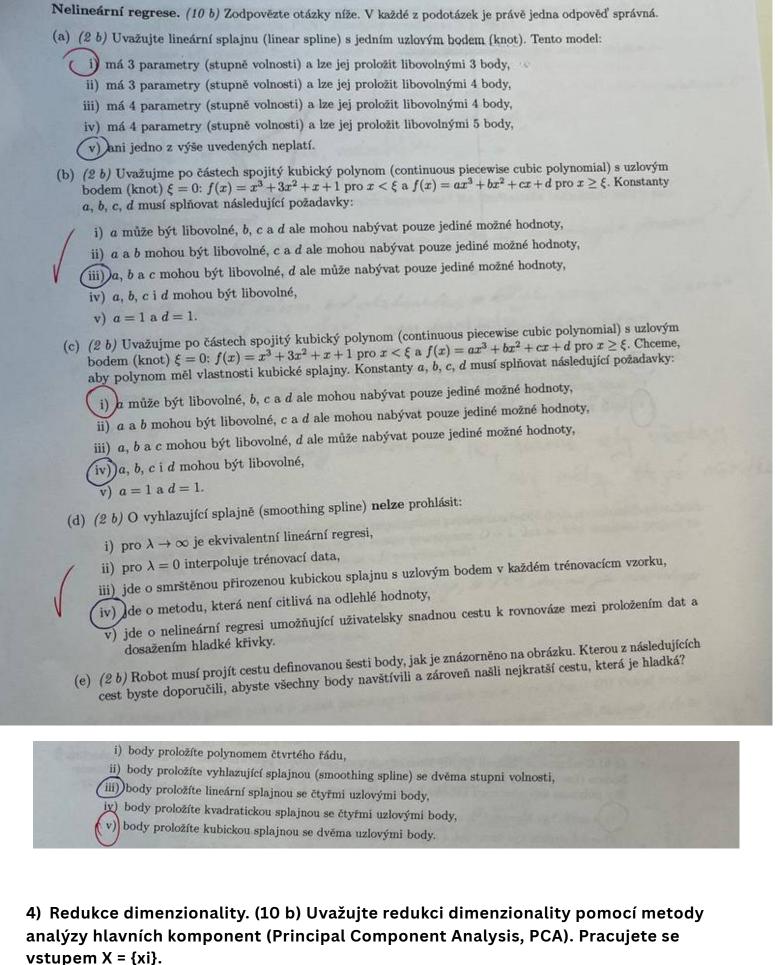
(a) (1 b) Navrhněte, jak by mohl vypadat generátor uvedených dat. Určete typ generátoru (rozdělení) a definujte jeho parametry (pojmenujte je a odhadněte jejich hodnoty).



(b) (2 b) Trénovací data použijete pro vytvoření klasifikátorů. Budete pracovat s lineární diskriminační analýzou (LDA), kvadratickou diskriminační analýzou (QDA) a logistickou regresí (LR). V obrázcích níže postupně vyznačte očekávané rozhodovací hranice všech tří metod. Rozhodnutí zdůvodněte.

(c) (2 b) Vysvětlete, která z uvedených metod bude dosahovat nejvyšší klasifikační přesnosti nad trénovacími daty a nad neviděnými testovacími daty ze stejného generátoru. Obě rozhodnutí zdůvodněte. Za jakých okolností dosáhne některá z metod 100% trénovací přesnosti?

(d) (2 b) Pracujeme s nevyváženým třídami a pravděpodobnostními klasifikátory. Nahraďte klasifikační přesnost vhodnějším hodnotícím kritériem. Popište způsob výpočtu tohoto kritéria a vysvětlete, jak řeší nevyváženost tříd a nakládá s pravděpodobnostmi.
(e) (2 b) Předpokládejme, že od této chvíle nepracujeme s triviální ztrátovou funkcí (všechny chyby hodnoceny identickou penaltou). Pokud ted' označíme instanci ze třídy 2 za instanci třídy 1, naše ztráta je 2x vyšší než pro záměnu opačnou. Popište, jak upravíte LDA, QDA a LR v této situaci.
(f) (1 b) Dojde v předchozím bodě ke změně rozhodovacích hranic? Pokud ano, do obrázků výše doplňte jinou barvou/typem čáry a popište, že jde o variantu se změněnou ztrátovou funkcí.
3)



(a) (2 b) Vysvětlete, kdy použijete PCA a v jakých situacích byste upřednostnili jiné metody redukce dimenze. Vysvětlete proč a jaké.
(b) (2 b) Aplikovali jste PCA. Kolik existuje hlavních komponent? Na čem založíte rozhodnutí kolik z nich použijete k redukci dimenze?
(c) (2 b) Před použitím PCA je doporučována standardizace proměnných (transformace, po níž mají všechny proměnné nulovou střední hodnotu a jednotkovou standardní odchylku). Vysvětlete, na čem se toto doporučení zakládá. Současně uveď te situaci, kdy standardizace není nutná/vhodná.
(d) (1b) Pracujete s doménou, ve které se zdá být smysluplnou proměnnou rozdíl dvou proměnných původních.Rozhodnete se tedy původní data doplnit o novou proměnnou D+ 1. Jak se toto doplnění projeví na výstupu PCA (srovnejte s původním během s D proměnnými)?
(e) (1 b) Lze PCA použít pokud je počet proměnných větší než je počet vzorků (m < D)? Pokud ano, jak?

(f) (2 b) Vysvětlete, co mají metody LDA a PCA společnéh obrázkem.	o a v čem se liší. Můžete vysvětlit
5) Robust statistics (viz san solved, Problem 8.)	