RPZ Test 8. února 2024

Jméno:

1. (10b) Bayesovské rozhodování.
   Uvažujme problém bayesovské klasifikace do dvou tříd označených 1 a 2. Příznaky (měření) $x \in \mathbb{R}$ jsou jednodimenzionální a víme, že:

   - apriorní pravděpodobnosti jsou $p_1 = \frac{1}{1+e}$, $p_2 = \frac{e}{1+e}$,
   - podmíněné pravděpodobnosti jsou

   $$p(x|1) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2\sigma_1^2}(x-m_1)^2}, \quad p(x|2) = \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{1}{2\sigma_2^2}(x-m_2)^2}, \tag{1}$$

   kde $\sigma_1 = \sigma_2 = 1$, $m_1 = -2$ a $m_2 = 2$.

   Uvažte ztrátovou funkci $W(k, d)$ definovanou tabulkou:

	$d = 1$	$d = 2$
$k = 1$	0	e^5
$k = 2$	e^2	0

   kde d je rozhodnutí (decision).

   (a) (3b) Definujte bayesovské riziko $R(q)$.

   (b) (7b) Najděte strategii $q(x)$, která pro výše uvedené zadání všech veličin minimalizuje bayesovské riziko $R(q)$. Strategie $q(x)$ pro každé x rozhodne o klasifikaci do jedné ze tříd $\{1, 2\}$.
   Zapište výslednou strategii v přehledné matematické formě.

2. (14b) Odhad parametrů.
   Hustota pravděpodobnosti $p(x)$ na intervalu $x \in [0, \infty]$ je pomocí parametrů $c > 0$ a $\alpha > 0$ definována takto:

   $$\begin{aligned} p(x) &= \frac{\alpha c^\alpha}{x^{\alpha+1}} \quad \text{pro } x \in [c, \infty] \,, \\ p(x) &= 0 \quad \text{pro } x \in [0, c) \,, \end{aligned}$$

   (a) (3b) Napište, jaké podmínky musí obecná funkce $f(x)$ splňovat, aby byla hustotou pravděpodobnosti, a ověřte, že $p(x)$ je splňuje.

   Mějme trénovací data $\mathcal{X} = \{x_1, x_2, ..., x_K\}$.

   (b) (2b) Pro tato trénovací data napište vzorec pro věrohodnost (likelihood).

   (c) (6b) Předpokládejme, že parametr c je daný (tedy je znám) a platí, že $c \leq x_k$ pro všechna $k = 1, 2, ..., K$. Odvoďte vztah pro výpočet parametru α pro tato data metodou maximální věrohodnosti (ML).

   (d) (3b) Nechť navíc není znám ani parametr c. Napište a zdůvodněte, jak tento parametr pomocí maximalizace věrohodnosti najdeme.

(a) $R(q) = \int_x \sum_{k \in H} p(x,k) \cdot W(k, q(x))$

$$p(x|1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x+2)^2}$$

$$p(x|2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-2)^2}$$

$$p(x,1) = p(x|1)\,p(1) = \left(\frac{1}{\sqrt{2\pi}} e^{-1/2(x+2)^2}\right) \cdot \frac{1}{1+e}$$

$$p(x,2) = p(x|2)\,p(2) = \left(\frac{1}{\sqrt{2\pi}} e^{-1/2(x-2)^2}\right) \cdot \frac{e}{1+e}$$

$p(x,1) \cdot W(1,2)$ → e^5

$p(x,2) \cdot W(2,1)$ → e^2

$$\frac{1}{\sqrt{2\pi}} e^{-1/2(x+2)^2} \cdot \frac{e^5}{1+e} < \frac{1}{\sqrt{2\pi}} e^{-1/2(x-2)^2} \cdot \frac{e^3}{1+e}$$

$$e^{5-1/2(x+2)^2} < e^{3-1/2(x-2)^2}$$

$$5 - 1/2(x+2)^2 < 3 - 1/2(x-2)^2$$

$$5 - 1/2(x^2 + 4x + 4) < 3 - 1/2(x^2 - 4x + 4)$$

$$5 - 1/2x^2 - 2x - 2 < 3 - 1/2x^2 + 2x - 2$$

$$2 < 4x$$

$$x > 1/2$$

$$q(x) \begin{cases} d = 1 & x > 1/2 \\ d = 2 & \text{else} \end{cases}$$

(2) Jestli má být $f(x)$ hustotou pst, musí integrovat do 1 a být nezáporná. $f(x)$ je nezáporná na int $(0, c)$

$$\int_{-\infty}^{\infty} p(x)\,dx = \int_{c}^{\infty} \alpha c^{\alpha} x^{-(\alpha+1)}\,dx =$$

$-\alpha - 1 + 1 = -\alpha$

$$= \left[\frac{\alpha c^{\alpha} x^{-\alpha}}{-\alpha}\right]_{c}^{\infty} = \left[\frac{-c^{\alpha}}{x^{\alpha}}\right]_{c}^{\infty} = \frac{-c^{\alpha}}{\infty} - \frac{-c^{\alpha}}{c^{\alpha}} =$$

$$= 1 \quad \square$$

(c) $L(\alpha, c) = \prod_{x \in X} p(x) = \prod_{x \in X} \frac{\alpha c^\alpha}{x^{\alpha+1}} = \alpha^k c^{k\alpha} \prod_{x \in X} x^{-\alpha-1}$

$$\ell(\alpha, c) = k \log \alpha + k \alpha \log c - (\alpha+1) \sum_{x \in X} \log x$$

$$\frac{\partial \ell}{\partial \alpha} = \frac{k}{\alpha} + k \log c - \sum_{i=0}^{k} \log x_i = 0 \quad / \cdot \alpha$$

$$k + \alpha k \log c - \alpha \sum \ldots = 0$$

$$\alpha = \frac{-k}{k \log c - \sum_{i=0}^{k} \log x_i}$$

(d) $\frac{\partial \ell}{\partial \alpha} = \frac{k}{\alpha} + k \log c - \sum_{i=0}^{k} \log x_i = 0$

$$\frac{\partial \ell}{\partial c} = k \log \alpha + \frac{k}{c} - \sum_{i=0}^{k} \log x_i = 0$$

$$\frac{k}{\alpha} + k \log C = \sum_{x \in X} \log x$$

$$k \log \alpha + \frac{k}{C} = \sum_{x \in X} \log x \quad)))$$

$$\frac{k}{\alpha} + k \log C = k \log \alpha + \frac{k}{C} \quad / : k$$

$$\frac{1}{\alpha} + \log C = \log \alpha + \frac{1}{C}$$

$$\frac{1}{\alpha} - \log \alpha = \frac{1}{C} - \log C$$

3. (10b) K-means.

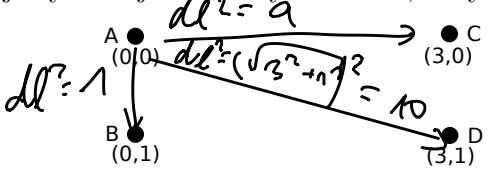
   (a) (2b) Vysvětlete algoritmus K-means.

   (b) (4b) Dokažte, že K-means po konečném počtu kroků dokonverguje do lokálního minima.

   (c) (4b) Uvažme inicializaci K-means pomocí K-means++. Na obrázku níže jsou zakresleny čtyři body v rovině spolu se svými souřadnicemi. Z těchto bodů vybíráme celkem dva centroidy. Předpokládejme, že bod A už byl vybrán jako první centroid. Napište, který bod bude nejpravděpodobněji vybrán jako druhý centroid, a vyčíslete tuto pravděpodobnost.

$dl = \min_{c \in C} \|x_l - c\|$

$p(l) = \frac{dl^2}{\sum_{\text{všech } l} dl^2}$



$p(B) = \frac{1}{20}$ $p(C) = \frac{9}{20}$

$p(D) = \frac{10}{20} = 1/2$

D je nejpravděpodobnější

4. (6b) Adaboost.
   Popište výhody a nevýhody metody Adaboost (jak pro učení, tak pro klasifikaci).

(3) (a) algoritmus k-means rozděluje data na k tříd. V prvním kroku se náhodně zvolí zvolí centroidy (typicky se vybere jeden z data pointů), poté se data přiřadí k nejbližšímu (pokud jsou dva nejbližší, bere se ten s nižším indexem) centroidu. Nové centroidy se spočítají jako geometrický medián bodů k nim přiřazených. Tento krok se opakuje do té doby dokud se rozřazení nezmění (nedojde v updatu při změně)

(b) (i) v každém kroku k-means se chyba přiřazení zmenšuje

(ii) existuje pouze konečně mnoho rozřazení data pointů na shluky

(4) výhody: jednoduchá implementace
dobře generalizuje
lineární klasifikátor

?!?

nevýhody: overfittuje příznaky