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.

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.

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

RPZ Test February 8, 2024

Name:

1. (10p) Bayes.
   Consider the problem of classification to two classes. The classes are denoted 1 and 2. The feature space (measurement) is one-dimensional, $x \in \mathbb{R}$. There is the following setting:

   - prior probabilities are $p_1 = \frac{1}{1+e}$, $p_2 = \frac{e}{1+e}$,
   - conditional probabilities are

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

   where $\sigma_1 = \sigma_2 = 1$, $m_1 = -2$, and $m_2 = 2$.

   Consider the following cost matrix $W(k, d)$:

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

   where k is the class index as above and d is decision.

   (a) (3p) Define the Bayes risk $R(q)$.

   (b) (7p) Find the optimal strategy $q(x)$, which (under the settings as described) minimizes the Bayes risk $R(q)$.
   Write down the strategy in a clear, mathematically concise form.

2. (14p) Parameter estimation.
   Probability density $p(x)$ on an interval $x \in [0, \infty]$ is defined as:

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

   where $c > 0$ and $\alpha > 0$ are parameters.

   (a) (3p) Write down the constraints on a general function $f(x)$ which are necessary for the function to be a probability density. Prove that the above $p(x)$ is consistent with these constraints.

   Let us have the training set $\mathcal{X} = \{x_1, x_2, ..., x_K\}$.

   (b) (2p) Write down the term for likelihood for these data.

   (c) (6p) Assume that parameter c is given (thus, known) and that $c \leq x_k$ for all $k = 1, 2, ..., K$. Derive the formula for computing parameter α using the Maximum Likelihood (ML) method.

   (d) (3p) Let the setting is as in the previous point, but in addition, parameter c is also unknown. Explain how to find also c using Maximum Likelihood.

3. (10p) K-means.

   (a) (2p) Give the description of the K-means algorithm.

   (b) (4p) Prove that K-means takes only a finite number of steps to converge to a local minimum.

   (c) (4p) Assume that K-means++ is used for K-means initialization. The picture below shows four points along with their coordinates. Our aim is to select two centroids from these points. Assume that point A has been already selected and it serves as the 1st centroid. Describe which of the remaining three points will be selected as the 2nd centroid with the highest probability. Compute the probability of selecting that point as the 2nd centroid.

4. (6p) Adaboost.
   Discuss advantages and disadvantages of Adaboost (both for training and for classification).