RPZ Test January 19, 2021

Name:

1. (6 points) Parzen Windows and Neyman-Pearson.
   There are two classes denoted 1 and 2. Let the training sets for the two classes be $X_1 = \{1, -1, 1, 2, 0\}$ and $X_2 = \{-1, 0, 1, 3\}$, respectively.

   (a) (3 points) Use the training data X_1 to estimate the conditional density $p(x \mid 1)$ using the Parzen window method. Do the same for X_2 to estimate $p(x \mid 2)$. Employ kernel function $K(x, y) = k(x - y)$, with $k(z)$ defined as:

   $$\begin{aligned} k(z) &= 1/h \quad \text{for } |z| \leq h/2 \,, \\ k(z) &= 0 \quad\;\; \text{for } |z| > h/2 \,, \end{aligned}$$

   with $h = 2$. Plot the estimated densities. Find values of $p(x \mid 1)$ and $p(x \mid 2)$ for $x = 0.5$.

   (b) (3 points) Consider class 2 as dangerous. Construct the Neyman Pearson strategy for the following setting: overlooked danger $< 40\%$. For this strategy, what are the errors ϵ_1 and ϵ_2 for class 1 and 2, respectively?

2. (8 points) Adaboost.

   (a) (3 points) Describe the Adaboost learning algorithm in a way which would enable a fellow student to implement the method.

   (b) (1 point) Describe time complexity of the training.

   (c) (1 point) Discuss time complexity of classification.

   (d) (1 point) Suggest possible speed-ups for training/classification stage.

   (e) (2 points) Highlight advantages of Adaboost with respect to SVM, Logistic regression, the neareast neighbour and neural network learning and classification methods.

3. (3 points) Maximum Likelihood Parameter Estimation. The density of a multivariate Normal distribution is given by

   $$p(\mathbf{x}) = (2\pi)^{-\frac{d}{2}} \det(\mathbf{C})^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})},$$

   where $\mathbf{x}, \boldsymbol{\mu} \in \mathbb{R}^d$, $\mathbf{C} \in \mathbb{R}^{d \times d}$, $\mathbf{C} \succeq 0$. You are given N i.i.d. observations $(x_i)_{i=1}^N$. The covariance matrix $\mathbf{C}$ is known. Find the maximum likelihood estimate of the mean $\boldsymbol{\mu}$.

4. (3 points) Classifier Selection.
   You are given a training data set and asked to find a simple classifier. Of course, the classifier should perform as well as possible. To begin with, you decide to use the K-nearest neighbor (K-NN) classifier. Explain how to choose K.

RPZ Test 19. ledna 2021

Jméno:

1. (6b) Parzenova okna a Neyman-Pearson.
   Mějme dvě třídy označené 1 a 2. Trénovací množina pro třídu 1 je $X_1 = \{1, -1, 1, 2, 0\}$, pro třídu 2 je $X_2 = \{-1, 0, 1, 3\}$.

   (a) (3b) Z dat X_1 odhadněte podmíněnou hustotu $p(x \mid 1)$ metodou Parzenových oken. Analogicky odhadněte hustotu $p(x \mid 2)$ pro druhou třídu. Použijte jádrovou funkci $K(x, y) = k(x - y)$, kde $k(z)$ je definovaná takto:

   $$\begin{aligned} k(z) &= 1/h \quad \text{pro } |z| \leq h/2 \,, \\ k(z) &= 0 \qquad \text{pro } |z| > h/2 \,, \end{aligned}$$

   kde $h = 2$. Odhadnuté hustoty graficky znázorněte. Spočítejte $p(x \mid 1)$ a $p(x \mid 2)$ pro $x = 0.5$.

   (b) (3b) Třída 2 představuje nebezpečí. Zkonstruujte Neyman-Pearsonovu strategii takovou, aby přehlédnuté nebezpečí bylo $< 40\%$. Pro nalezenou strategii spočítejte chyby klasifikace ϵ_1, ϵ_2 pro obě třídy 1 a 2.

2. (8b) Adaboost.

   (a) (3b) Vysvětlete metodu Adaboost tak, aby ji podle Vašeho popisu kterýkoli Váš kolega dokázal naimplementovat.

   (b) (1b) Popište časovou náročnost učení.

   (c) (1b) Diskutujte časovou náročnost klasifikace.

   (d) (1b) Napište, jestli a jak je možné chod algoritmu zrychlit při učení/klasifikaci.

   (e) (2b) Popište výhody Adaboostu ve srovnání s SVM, logistickou regresí, nejbližším sousedem a neuronovými sítěmi (zvažte jak fázi učení, tak i fázi klasifikace.)

3. (3b) Maximálně věrohodný odhad parametrů. Normální rozdělení je dáno jako

   $$p(\mathbf{x}) = (2\pi)^{-\frac{d}{2}} \det(\mathbf{C})^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})},$$

   kde $\mathbf{x}, \boldsymbol{\mu} \in \mathbb{R}^d$, $\mathbf{C} \in \mathbb{R}^{d \times d}$, $\mathbf{C} \succeq 0$. Mějme N nezávislých pozorování $(x_i)_{i=1}^N$ z tohoto rozdělení. Kovarianční matice $\mathbf{C}$ je známá. Najděte maximálně věrohodný odhad parametru $\boldsymbol{\mu}$.

4. (3b) Výběr klasifikátoru.
   Máte k dispozici trénovací data a máte za úkol implementovat jednoduchý klasifikátor. Chcete, aby klasifikátor fungoval co nejlépe. Rozhodl/a jste se, že začnete s metodou K-nejbližšího souseda (K-NN). Vysvětlete, jak vyberete K.