Tsinghua-Berkeley Shenzhen Institute
Learning From Data
Fall 2022

Writing Assignment 4

Issued: Sunday 27th November, 2022 **Due:** Sunday 11th December, 2022

POLICIES

- **Acknowledgments:** We expect you to make an honest effort to solve the problems individually. As we sometimes reuse problem set questions from previous years, covered by papers and web pages, we expect the students **NOT** to copy, refer to, or look at the solutions in preparing their answers (relating to an unauthorized material is considered a violation of the honor principle). Similarly, we expect you to not google directly for answers (though you are free to google for knowledge about the topic). If you do happen to use other material, it must be acknowledged here, with a citation on the submitted solution.

- **Required homework submission format:** You can submit homework either as one single PDF document or as handwritten papers. Written homework needs to be provided during the class on the due date, and a PDF document needs to be submitted through Tsinghua's Web Learning (`http://learn.tsinghua.edu.cn/`) before the end of the due date.

- **Collaborators:** In a separate section (before your answers), list the names of all people you collaborated with and for which question(s). If you did the HW entirely on your own, **PLEASE STATE THIS**. Each student must understand, write, and hand in answers of their own.

POLICIES

- **Acknowledgments:** We expect you to make an honest effort to solve the problems individually. As we sometimes reuse problem set questions from previous years, covered by papers and web pages, we expect the students **NOT** to copy, refer to, or look at the solutions in preparing their answers (relating to an unauthorized material is considered a violation of the honor principle). Similarly, we expect you to not google directly for answers (though you are free to google for knowledge about the topic). If you do happen to use other material, it must be acknowledged here, with a citation on the submitted solution.

- **Required homework submission format:** You can submit homework either as one single PDF document or as handwritten papers. Written homework needs to be provided during the class on the due date, and a PDF document needs to be submitted through Tsinghua's Web Learning (`http://learn.tsinghua.edu.cn/`) before the end of the due date.

- **Collaborators:** In a separate section (before your answers), list the names of all people you collaborated with and for which question(s). If you did the HW entirely on your own, **PLEASE STATE THIS**. Each student must understand, write, and hand in answers of their own.

4.1. (2 points) Let $\mathcal{H}^2_{\text{rec}}$ be the class of axis aligned rectangles in $\mathbb{R}^2$, namely, $\mathcal{H}^2_{\text{rec}} = \{h_{(a_1,a_2,b_1,b_2)} : a_i < b_i, i = 1,2\}$ where

$$h_{(a_1,a_2,b_1,b_2)}(\boldsymbol{x}) = \begin{cases} 1, & a_1 < x_1 < b_1, a_2 < x_2 < b_2 \\ 0, & \text{otherwise} \end{cases}.$$

Please show that $\text{VCdim}(\mathcal{H}^2_{\text{rec}}) = 4$.

4.2. (2 points) (Mean Square Error) We mentioned Bias-Variance Tradeoff in class. Given the noiseless model $y = h(x)$, we define the MSE of $\hat{h}$, an estimator of X as $\text{MSE}(\hat{h}) \triangleq \mathbb{E}[(\hat{h}(x)-y)^2]$. The variance of $\hat{h}$ is defined as $\text{Var}(\hat{h}) \triangleq \mathbb{E}[(\hat{h}(x)-\mathbb{E}[\hat{h}(x)])^2]$ and the bias is defined as $\text{Bias}(\hat{X}) \triangleq \mathbb{E}[\hat{h}(x)] - h(x)$.

(a) (1 point) Please prove that

$$\text{MSE}(\hat{h}(x)) = \text{Var}(\hat{h}(x)) + (\text{Bias}(\hat{h}(x)))^2.$$

(b) (1 point) Our data are added with an independent Gaussian noise say, $y = h(x) + \epsilon$, where $\mathbb{E}[\epsilon] = 0$ and $\mathbb{E}[\epsilon^2] = \sigma^2$ and the estimator is $\hat{h}(x)$. We define the empirical MSE as $\mathbb{E}[(\hat{h}(x) - h(x) - \epsilon)^2]$. Please prove that

$$\mathbb{E}[(\hat{h}(x) - h(x) - \epsilon)^2] = \text{MSE}(\hat{h}) + \sigma^2.$$

The equation tells us that the empirical error is a good estimation of the true error. Thus, we can minimize the empirical error in order to properly minimize the true error.

4.3. (2 points) Important inequalities in Learning Theory.

(a) (1 point) (Markov's Inequality) Let X be a non-negative random variable, then for every positive constant a, please show that

$$P(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

(b) (1 point) (Chebyshev's inequality)For random variable X, if its expected value $\mathbb{E}(X)$ and variance $Var(X)$ are both finite, for every positive constant a, please show that

$$P(|X - \mathbb{E}(X)| \geq a) \leq \frac{Var(X)}{a^2}.$$

4.4. (4 points) (K-means) Given input data $\mathcal{X} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$, $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$, the k-means clustering partitions the input into k sets $C_1, \ldots, C_k$ to minimize the within-cluster sum of squares:

$$\arg\min_{C} \sum_{j=1}^{k} \sum_{\boldsymbol{x} \in C_j} \|\boldsymbol{x} - \boldsymbol{\mu}_j\|^2,$$

where $\boldsymbol{\mu}_j$ is the center of the j-th cluster:

$$\boldsymbol{\mu}_j \overset{\text{def}}{=} \frac{1}{|C_j|} \sum_{\boldsymbol{x} \in C_j} \boldsymbol{x}, \quad j = 1, \ldots, k.$$

(a) (2 points) Show that the k-means clustering problem is equivalent to minimizing the pairwise squared deviation between points in the same cluster:

$$\sum_{j=1}^{k} \frac{1}{2|C_j|} \sum_{\boldsymbol{x}, \boldsymbol{x}' \in C_j} \|\boldsymbol{x} - \boldsymbol{x}'\|^2.$$

(b) (2 points) Show that the k-means clustering problem is equivalent to maximizing the between-cluster sum of squares:

$$\sum_{i=1}^{k} \sum_{j=1}^{k} |C_i||C_j|\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2.$$