Homework 10

Due April 24 at 11pm

Unless stated otherwise, justify any answers you give. You can work in groups, but each student must write their own solution based on their own understanding of the problem.

When uploading your homework to Gradescope you will have to select the relevant pages for each question. Please submit each problem on a separate page (i.e., 1a and 1b can be on the same page but 1 and 2 must be on different pages). We understand that this may be cumbersome but this is the best way for the grading team to grade your homework assignments and provide feedback in a timely manner. Failure to adhere to these guidelines may result in a loss of points. Note that it may take some time to select the pages for your submission. Please plan accordingly. We suggest uploading your assignment at least 30 minutes before the deadline so you will have ample time to select the correct pages for your submission. If you are using LaTeX, consider using the minted or listings packages for typesetting code.

1. (Equivalence of partition estimators and least square regression with feature maps) Let $\mathcal{A} = \{A_1, \dots, A_J\}$ be a partition of $[0,1]^d$. Consider the feature map $\Phi : [0,1]^d \to \mathbb{R}^J$ defined by

$$\Phi(x) = (\mathbf{1}\{x \in A_1\}, \dots, \mathbf{1}\{x \in A_J\}). \tag{1}$$

To be more explicit, the vector $\Phi(x)$ contains 0 in all of its entries, except in the entry j that satisfies $x \in A_j$, where it contains a 1. Let $(\mathbf{x_1}, \mathbf{y_1}), \dots, (\mathbf{x_n}, \mathbf{y_n})$ be a training sample from distribution P. We recall the definition of the partition estimator associated with $\mathcal{A}$: let I_j be the set of indexes such that $\mathbf{x_i} \in A_j$ and let $\mathbf{n_j}$ be the number of elements of I_j. **For the sake of simplicity, we assume that $\mathbf{n_j} > 0$ for every j.** The partition estimator is defined by

$$\hat{f}_{\mathcal{A}}(x) = \frac{1}{\mathbf{n_j}} \sum_{i \in I_j} \mathbf{y_i}$$

for every $x \in [0,1]^d$ such that $x \in A_j$. The goal of this problem is to show that the partition estimator is equal to the least square regression estimator obtained with the feature map Φ. Let $\tilde{\mathbf{X}}$ be the $n \times K$ matrix whose rows are given by the vectors $\Phi(\mathbf{x_i})$.

(a) Show that $\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}$ is a $K \times K$ diagonal matrix equal to

$$\begin{pmatrix} \mathbf{n_1} & & & \\ & \mathbf{n_2} & & \\ & & \ddots & \\ & & & \mathbf{n_J} \end{pmatrix}.$$

(b) Recall from the previous chapter that the optimal vector $\hat{a}$ in least-square regression is equal to $\hat{a} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \mathbf{Y}$, where $\mathbf{Y} \in \mathbb{R}^n$ is the vector with entries $\mathbf{y_i}$. The associated predictor is then given by $\hat{f}_{\mathrm{LS}}(x) = \langle \hat{a}, \Phi(x) \rangle$. Show that $\hat{f}_{\mathrm{LS}}(x) = \hat{f}_{\mathcal{A}}(x)$ for every $x \in [0,1]^d$.

2. (Neighbors in high dimension) Assume that we have access to n observations $\mathbf{x_1}, \ldots, \mathbf{x_n}$ that are uniformly sampled in the cube $[0,1]^d$. We assume that the dimension d is "large".

(a) Sample $n = 500$ uniform observations in the cube $[0,1]^d$ for $d = 2, 10, 500$ and $10,000$. Let $\mathbf{x}$ be another uniform observation in $[0,1]^d$. Show the plot of the histogram of the distances $\|\mathbf{x} - \mathbf{x_i}\|^2$ (for $i = 1, \ldots, n$) for those different values of d. Compare the standard deviation and the expectation for different values of d. What do you observe?

(b) Argue thanks to the previous question that all the squared distances $\|\mathbf{x} - \mathbf{x_i}\|^2$ are roughly equal in high dimension. Explain why in high-dimension the notion of "nearest-neighbor" becomes irrelevant.