CS 182: Introduction to Machine Learning, Fall 2022
Homework 4

(Due on Wednesday, Dec. 7 at 11:59pm (CST))

Notice:

- Please submit your assignments via Gradescope. The entry code is G2V63D.
- Please make sure you select your answer to the corresponding question when submitting your assignments.
- Each person has a total of five days to be late without penalty for all the assignments. Each late delivery less than one day will be counted as one day.

1. [20 points] *[Clustering and Mixture Models]*

   (a) Describe the k-means clustering algorithm step-by-step; [10 points]

   (b) Given a set of 5 samples

   $$X = \begin{bmatrix} 0 & 0 & 1 & 5 & 5 \\ 2 & 0 & 0 & 0 & 2 \end{bmatrix}$$

   try the k-means clustering algorithm to cluster the samples into 2 clusters. (You should write a detailed derivation) [10 points]

2. [20 points] *[Clustering and Mixture Models]* A Gaussian mixture model (GMM) is a model of the form

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$$

with model parameters π_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$, where $\pi_k \geq 0$ and $\sum_k \pi_k = 1$.

(a) Discuss the advantages of the GMM and why it can be used for clustering; [5 points]

(b) Given a training set $\{\mathbf{x}_i \in \mathbb{R}^D\}_{i=1}^N$, try the expectation-maximization (EM) algorithm to estimate the parameters of the GMM. [15 points]

3. [20 points] *(Nonparametric Density Estimation)* Let the i.i.d. sample $\mathcal{X} = \{x_i\}_{i=1}^n$ be drawn from some unknown probability density $p(x)$ with $x \in [0, 1]$. We can obtain the histogram estimator $\hat{p}(x)$ for $p(x)$ with bin width specified as h. We define the loss of estimation error in the L^2 space:

$$L(h) = \int_0^1 \left(\left(\hat{p}(x) - p(x)\right)^2 \mathsf{d}x = \int_0^1 \hat{p}^2(x)\mathsf{d}x - 2\int_0^1 \hat{p}(x)p(x)\mathsf{d}x + \int_0^1 p^2(x)\mathsf{d}x.\right.$$

Considering the last term $\int p^2(x)\mathsf{d}x$ is uncorrelated with $\hat{p}(x)$ and replacing the integral with the average, we get

$$L'(h) = \int_0^1 \hat{p}^2(x)\mathsf{d}x - \frac{2}{n}\sum_{i=1}^n \hat{p}(x_i).$$

Please derive

(a) the expression of $\hat{p}(x)$; [6 points]

(b) the expression of $L'(h)$ based on the histogram estimator $\hat{p}(x)$; [6 points]

(c) the h that minimizes $L'(h)$. [8 points]

4. [20 points] *[Nonparametric Regression]* Given a set of n examples, $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, a linear smoother is defined as follows. For any $\mathbf{x}$, there exists a vector $\ell(\mathbf{x}) = (\ell_1(\mathbf{x}), \ldots, \ell_n(\mathbf{x}))^\top$ such that the estimated output $\hat{y}$ of $\mathbf{x}$ is $\hat{y} = \sum_{i=1}^n \ell_i(\mathbf{x}) y_i = \ell(\mathbf{x})^\top Y$ where Y is a $n \times 1$ vector, $Y_i = y_i$.

   (a) In linear regression with basis functions h, the data is assumed to be generated by $y_i = \sum_{j=1}^m w_j h_j(\mathbf{x}_i) + \epsilon_i$. The least squares estimate for the coefficient vector $\mathbf{w}$ is given by $\mathbf{w}^* = \left(H^\top H\right)^{-1} H^\top Y$, where H is a $n \times m$ matrix, $H_{ij} = h_j(\mathbf{x}_i)$. Given an input $\mathbf{x}$, please derive the estimated output $\hat{y}$ (The solution should be in matrix form, and you may use the vector $h(\mathbf{x}) = [h_1(\mathbf{x}), \ldots, h_m(\mathbf{x})]^\top$). Furthermore, is linear regression a linear smoother? [10 points]

   (b) In kernel regression, if we use the kernel $K(\mathbf{x}_i, \mathbf{x}) = \exp\left\{\frac{-\|\mathbf{x}_i - \mathbf{x}\|^2}{2\sigma^2}\right\}$, given an input $\mathbf{x}$, please derive the estimated output $\hat{y}$. Furthermore, is this kernel regression a linear smoother? [10 points]

5. [20 points] [*Coding: EM*] Complete "HW4-Coding.ipynb". After completion, you should convert your notebook to PDF, and concatenate the writing part and the coding part into one PDF which is the file to submit.