Homework Assignment 3

CSE 251A: ML - Learning Algorithms

Due: May 10th, 2022, 9:30am (Pacific Time)

Instructions: Please answer the questions below, attach your code in the document, and insert figures to create **a single PDF file**.

Grade: ____ out of 100 points

1 (20 points) SVM and kernel

1. (Linearly separability)(5 points) In lecture, we mentioned **Hard SVM** only works for **linearly separable** data points. Now you are given dataset $S = \{(\mathbf{x}_i, y_i), i = 1, .., N\}$ where each data point $(\mathbf{x}, y)$ contains a feature vector $\mathbf{x} \in \mathbb{R}^2$ and a label $y \in \{+, -\}$. Is the dataset linearly separable? If yes, please specify a decision boundary that linearly separates the data points with positive label from the data points with negative label.

   - Case 1:

label (y)	features ($\mathbf{x}$)
Positive (+)	$(0, 0)$
Negative (−)	$(1, 0), (0, 1), (-1, 0), (0, -1)$

   - Case 2:

label (y)	features ($\mathbf{x}$)
Positive (+)	$(0, 1), (1, 0)$
Negative (−)	$(-1, 0), (0, -1)$

2. (Feature map and kernel)(10 points) Sometimes, data points that are not linearly separable in one feature space might be linearly separable in another. Therefore, we need a **transformation function (feature map)** ϕ, that maps features from one space to another space. Suppose for feature vector $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$ we define **feature map** $\phi : \mathbb{R}^2 \to \mathbb{R}^3$ as: $\phi(\mathbf{x}) = \left(x_1^2, x_2^2, \sqrt{2}x_1x_2\right)$. Given the **feature map** ϕ, the corresponding **kernel** is defined as: $k(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{z})$, where $\mathbf{x}, \mathbf{z} \in \mathbb{R}^2$. Prove that $k(\mathbf{x}, \mathbf{z}) = (\mathbf{x} \cdot \mathbf{z})^2$ is the **kernel** for ϕ. In other words, prove that $\phi(\mathbf{x}) \cdot \phi(\mathbf{z}) = (\mathbf{x} \cdot \mathbf{z})^2$.

   Hint: for vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \mathbf{x} \cdot \mathbf{y} = x_1y_1 + x_2y_2 + \ldots + x_ny_n$.

3. (5 points) Recall Case 1 in part 1. Apply ϕ defined in part 2 to the 5 data points in Case 1 and compute the new features in $\mathbb{R}^3$. Are they now linearly separable in $\mathbb{R}^3$? (Optional: specify a valid decision boundary in $\mathbb{R}^3$ if linearly separable)

2 (10 points) K-means

Suppose you are given 4 data points: $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4 \in \mathbb{R}^2$, where $\mathbf{x}_1 = (1, 0), \mathbf{x}_2 = (1, 1)$, $\mathbf{x}_1 = (-1, 0), \mathbf{x}_1 = (-1, -1)$. Assume we know these points can be put into 2 clusters ($k = 2$). Suppose we start with **centroids** $\mathbf{c}_1 = (2, 0), \mathbf{c}_2 = (0, -1)$. Calculate where the new centroids would be after 1 iteration of **K-means**. Do the new centroids locations make sense?

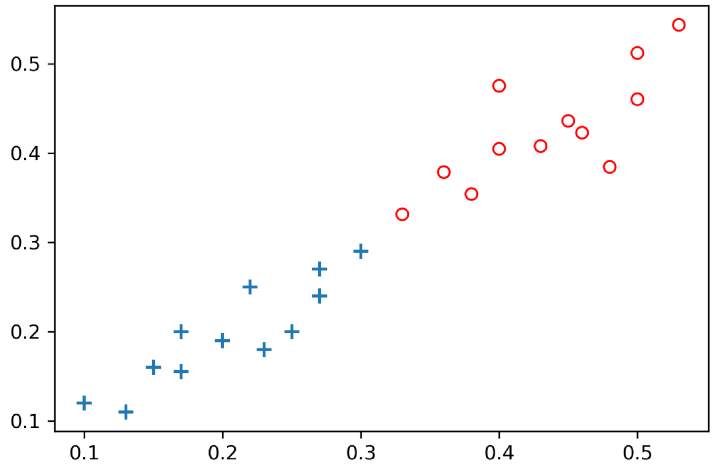
3 (10 points) Gaussian Mixture Models

Suppose we have a set of probabilistic clusters, $\mathbf{C} = (C_1, C_2)$ .C_1 has probability density function $f_1 = \mathcal{N}(0, 1)$, that is, a Gaussian distribution with mean 0 and variance 1 , and C_2 has probability density function $f_2 = \mathcal{N}(1, 1)$. The weights of these clusters are given as: $\mathbb{P}(C_1) = \mathbb{P}(C_2) = 0.5$. Given the following data points in 1D, calculate the probability $\mathbb{P}(x \mid \mathbf{C})$ that a data point x is generated by this set of clusters $\mathbf{C}$, for the following 2 data points: $(1)x = 0.7, (2)x = 1.5$.

Hint: The probability density function of $\mathcal{N}(\mu, \sigma)$ is: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x-\mu)^2}{2\sigma^2}$

4 (10 points) Principle Component Analysis (PCA)

You are now given some data points in 2D, with 2 kinds of labels ("+" and "o").



(1) (5 points) On the scatter plot, roughly draw the direction for the 1^{st} and 2^{nd} principle axes for this dataset.

(2) (5 points) Which principle axis would you choose to project onto, if we want to project the 2D features onto 1D, and still want good separation between the 2 labels?

5 (50 points) Implementing the K-means algorithm

Now, you will implement a K-Means model from scratch. We have provided a skeleton code file (i.e. KMeans.py) for you to implement the algorithm as well as a notebook file (i.e. KMeans.ipynb) for you to conduct experiments and answer relevant questions. Libraries such as numpy and pandas may be used for auxiliary tasks (such as matrix multiplication, matrix inversion, and so on), but not for the algorithms. That is, you can use numpy to implement your model, but cannot directly call libraries such as scikit-learn to get a K-Means model for your skeleton code. We will grade this question based on the three following criteria:

1. Your implementation in code. Please do not change the structure of our skeleton code.
2. Your model's performance (we check if your model behaves correctly based on the results from multiple experiments in the notebook file).
3. Your written answers for questions in the notebook file.

6 (20 points) Exponential Mixture Model

Suppose there exists a distribution called **Exponential Distribution** whose probability density function is described as:

$$\text{Exp}(\beta)\text{: } f(x;\beta) = \begin{cases} \beta e^{-\beta x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

where β is the parameter.

Suppose $X \in \mathbb{R}^+$ is sampled from a mixture of K clusters and each cluster Z_i is characterized by an Exponential Distribution with a parameter β_{Z_i}. The corresponding cluster priors are given as

$$P(Z_i) = \Pi_i \text{ for } i = 1, 2, \ldots, K$$

For example, given $K = 2$, $P(Z_1) = \Pi_1, P(Z_2) = \Pi_2$, X is sampled from a mixture of $\text{Exp}(\beta_{Z_1})$ and $\text{Exp}(\beta_{Z_2})$.

1. Given $K = 3$ and $\beta_{Z_1} = 1$, $\beta_{Z_2} = 2$, $\beta_{Z_3} = 4$, what is $P(Z_1|X = 1)$? **(5 points)**
2. Describe the E-step. Write an equation for each value that is computed. **(15 points)**