Midterm Exam

CSE 251A: ML – Learning algorithms

Start Time: May 11th, 2023, 9:30am (Pacific Time)
End Time: May 12th, 2023, 9:30am (Pacific Time)

Instructions: Please answer the questions below and create **a single PDF file**.

Grade: ____ out of 100 points

1 (10 points) Comparison

Please answer the following questions.

1. What's the relationship between Nearest Neighbor Classifier and K-Means Clustering? Can K-Means clustering inference (i.e., finding a cluster from the existing ones for a new data point in the test set) be viewed as Nearest Neighbor Classifier?

2. Suppose we have N data points $\{(x_i, y_i)\}_{i=1}^{N}$ and we want to build a regression model to predict Y using X. After visualizing these data point, we observed an exponential relationship between X and Y. In this case, can we leverage the LinearRegression model from sklearn to quantify the exponential relationship? (1) Yes or No? (2) If Yes, how? If No, why?

2 (15 points) Linear Regression with "Exponential Regularization"

Let's consider the house rent prediction problem — we are supposed to predict the price of a house based on just its area. Suppose we have n samples with their respective areas, $x^{(1)}, x^{(2)}, ..., x^{(n)}$, and their true house rents, $y^{(1)}, y^{(2)}, ..., y^{(n)}$. Let's say, we train a linear regression model that predicts $\hat{y}^{(i)} = \theta_0 + \theta_1 x^{(i)}$. The parameters θ_0 and θ_1 are scalars to be learned through minimizing the loss function L using gradient descent with a **learning rate** α.

The **loss function** L here is the **mean-squared error** plus an **exponential regularization** — a regularization TA invented for this question. Mathematically, if β is the scalar parameter to control the regularization strength, and w_1, w_2, ..., w_d (d is the number of dimensions) are the coefficients, $\beta \sum_{j=1}^{d} e^{w_j}$ is added as a penalty. In the context of this question, the added penalty shall look like $\beta(e^{\theta_0} + e^{\theta_1})$.

Given all the information above, please answer the following questions.

1. Express the loss function(L) in terms of $x^{(i)}$, $y^{(i)}$, n, θ_0, θ_1, β, α.

2. Compute $\frac{\partial L}{\partial \theta_0}$.

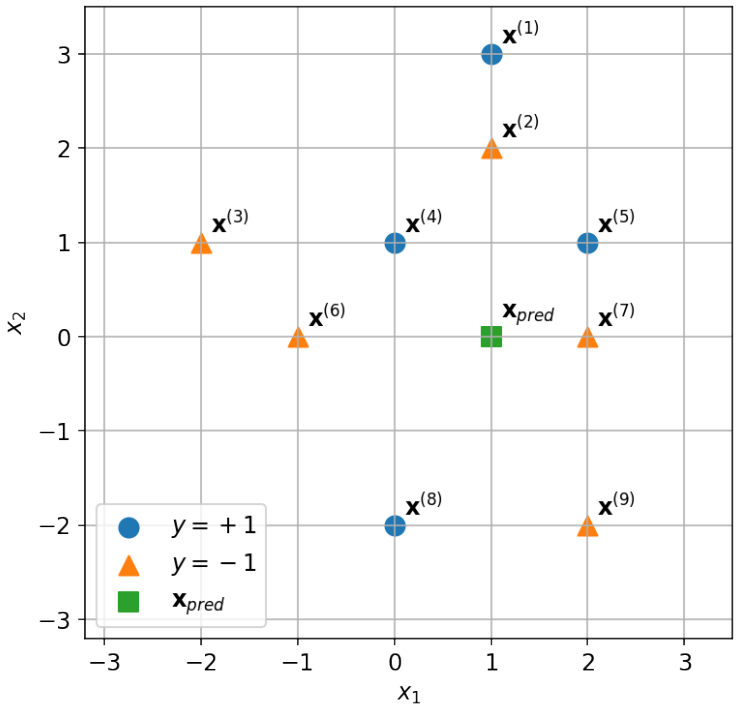
3. Compute $\frac{\partial L}{\partial \theta_1}$.

4. Write update rules for θ_0 and θ_1

5. How does the above regularization perform in experiments? Does this work or fail? Write your thoughts and reasoning. (This is an open-ended question. You do not need to do any experiments for this)

3 (15 points) k Nearest Neighbour Classification

Assume there is a training dataset $S = \{(\mathbf{x}^{(i)}, y^{(i)}), i = 1, ..., 9\}$ where each data point $(\mathbf{x}, y)$ contains a feature vector $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$ and a ground-truth label $y \in \{+1, -1\}$. The dataset is visualized in the following figure, where data points with label $+1$ are marked as circle (●) and data points with label -1 are marked as triangle (▲):



In this problem, we would like to find k nearest neighbors using different distance metrics to predict the label of a data point with feature vector $\mathbf{x}_{\text{pred}} = (1, 0)$, which is marked as square (■). Please answer the questions below:

1. What is the predicted label of $\mathbf{x}_{\text{pred}}$ if we use 3 nearest neighbor with Euclidean distance as the distance metric?

2. What is the predicted label of $\mathbf{x}_{\text{pred}}$ if we use 5 nearest neighbors with Euclidean distance as the distance metric?

3. What is the predicted label of $\mathbf{x}_{\text{pred}}$ if we use 5 nearest neighbors with Manhattan distance as the distance metric?

4 (10 points) Matrix Computation

1. If A is an arbitrary square matrix, determine whether $(A - A^T)$ is a symmetric matrix.

2. Given $\mathbf{x}$ an N-dim vector and $||\mathbf{x}||_2 = \sqrt{2}$, what is the result of $\mathbf{x}^T\mathbf{x}\mathbf{x}^T\mathbf{x}\mathbf{x}^T\mathbf{x}\mathbf{x}^T\mathbf{x}$

5 (10 points) Clustering

Assume there is a dataset $S = \{x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}, x^{(5)}\}$ where:

$$x^{(1)} = -8, \qquad x^{(2)} = -10, \qquad x^{(3)} = 0,$$
$$x^{(4)} = 11, \quad x^{(5)} = 12.$$

1. If we have initial centroids $c_1 = 0$, $c_2 = 11$, $c_3 = 12$ and perform ***K*-means** clustering algorithm on dataset S using Euclidean distance. Fill in the table until the algorithm converges. To show the algorithm has converged, it is necessary to fill in one more iteration that duplicates the previous one. You may not use all the rows. Some example entries are filled. For each of the iteration, find the positions of the centroids, assign the data points to the centroids, and compute the loss.

	centroid positions			data points assignment					
iter	c_1	c_2	c_3	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	SSE
1	0	11	12	c_1					
2									
3									
4									
5									

2. What is the lowest SSE of the most optimal assignment of these data points?

3. Where are the optimal centroids?

6 (15 points) SVM and Kernel

1. (Solving a SVM) In class, we have learned the formulation of a soft margin SVM:

$$\min \lambda|w|_2^2 + \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i(w^T x_i - b))$$

SVM are solvable via gradient descent. Write down the gradient for the parameter w. If the gradient at a certain location is not defined, point it out, and ignore it in later computations.

2. (Support vectors) Construct a sufficiently large, linearly separable binary-label dataset, such that when we train a hard SVM model based on this dataset, the number of support vectors is minimized. You need to describe breifly how the dataset looks like and answer the minimum number of support vectors.

3. (Understanding kernels) Kernels is a core concept in Dual SVM. Suppose k_1 and k_2 are two **kernels** with **feature map** ϕ_1 and ϕ_2, that is, for feature vectors $\mathbf{x}, \mathbf{z}$,

$$k_1(\mathbf{x}, \mathbf{z}) = \phi_1(\mathbf{x})^T \phi_1(\mathbf{z})$$
$$k_2(\mathbf{x}, \mathbf{z}) = \phi_2(\mathbf{x})^T \phi_2(\mathbf{z})$$

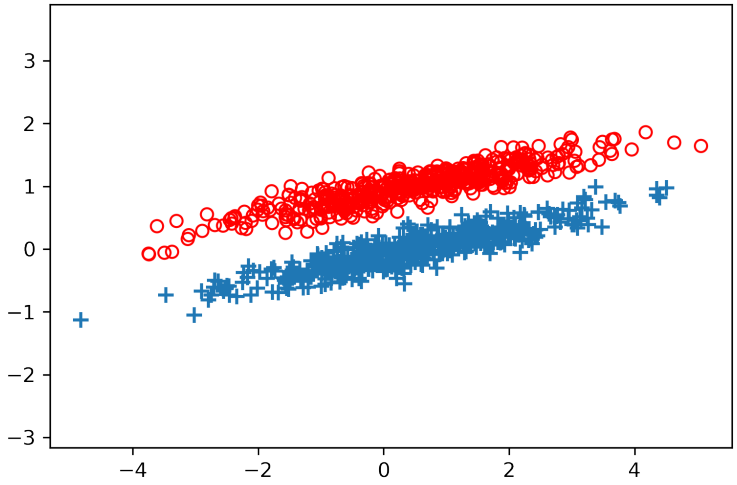
Now we construct a new **kernel**,

$$k'(\mathbf{x}, \mathbf{z}) = k_1(\mathbf{x}, \mathbf{z}) k_2(\mathbf{x}, \mathbf{z})$$

Please write a corresponding **feature map** ϕ' for **kernel** k', in terms of ϕ_1 and ϕ_2. You need to elaborate your answer in details. There could be multiple correct solutions and you will only need one feasible ϕ' here.

7 (10 points) 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?

8 (5 points) Loss Function

Can Maximum Absolute Error (MAE), Rooted Mean Square Error (RMSE) and Max Error (ME) between two vectors $y \in \mathbb{R}^n$ and $\hat{y} \in \mathbb{R}^n$ ($n > 1$) have the same **non-zero** value? If yes, give an example. If no, explain why. Recall the definitions of these loss functions as follows.

- MAE $= \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i|$
- RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}$
- ME $= \max_{i=1}^{n} |\hat{y}_i - y_i|$

9 (10 points) Kernel or Not

Prove the following functions are valid kernels, or disprove it, $x, y \in \mathbb{R}^d$.

1. $k(x, y) = (x \cdot y + 2)^2$
2. $k(x, y) = (x \cdot y - 1)^3$

Hint: A kernel is considered valid when it satisfies Mercer's theorem, i.e., $\forall x_1, x_2, \ldots, x_n \in \mathbb{R}^d$, $K \in \mathbb{R}^{n \times n}$ *is a positive semi-definite matrix, where its element at the* i*-th row and* j*-th column is equal to* $k(x_i, x_j)$*, namely* $K_{i,j} = k(x_i, x_j),\ i, j \in \{1, \ldots, n\}$*. This property can also be shown by finding a function* $\phi(\cdot)$*, such that* $k(x, y) = \phi(x)\phi(y)$*.*