CS559 Assignment 2

Due 3/11/2022 Friday 11:59 PM

Problem 1 and 2 are fundamental questions. Please do on paper and submit in pdf format.

Problem 3, 4, and 5 are implementing questions. Please work on notebook and submit in notebook file. **Do not use any packages/tools except numpy.**

**Problem 1 – Linear Regression** (15 pts)

Consider a linear model of the form

together with a sum of squares error function of the form

Suppose that Gaussian noise is added independently to each of the input variables . By making use of the expected value of noise, , and , where , show that minimizing averaged over the noise distribution is equivalent to minimizing the sum-of-square for noise-free input variables with the addition of a regularized term .

**Problem 2 – Linear Discriminative Analysis** (15 pts)

Using the definitions of the between-class and within-class covariance matrices

and

respectively, together with

where is the total data set mean, show that the expression

that minimizes the sum-of-squares error function can be written in the form of

**Problem 3. KNN** (20 pts)

Download the yeast dataset from the link: <https://archive.ics.uci.edu/ml/datasets/Yeast>.

1. Implement a basic KNN model on the yeast dataset. The task is to predict the compartment in a cell that a yeast protein will localize to base on properties of its sequence.
2. To optimize the results, test with Manhattan and Euclidean distance metrics.
3. Report the accuracy of the model for both distance metrics with k values from 5 to 20.

**Problem 4. Perceptron** (20 pts)

Using the data table below

Table 1

|  |  |  |  |
| --- | --- | --- | --- |
| Data |  |  |  |
|  | 4  4  1  2.5  4.9  1.9  3.5  0.5  2  4.5 | 2.9  4  2.5  1  4.5  1.9  4  1.5  2.1  2.5 | 1  1  -1  -1  1  -1  1  -1  -1  1 |

1. Write a pseudocode of the Perceptron algorithm
2. Implement the Perceptron algorithm to find the hyperplane equation, , and report the hyperplane.
3. Then, classify the point using from above.

**Problem 5. Logistic Regression** (20 pts)

Download the breast cancer data set from UCI Machin Learning repository.

1. Implement a Logistic Regression classifier with ML estimator using gradient descent algorithm.
2. Implement a 5-fold cross validation algorithm to report the model and accuracy of each fold.
3. Report the result.