MACHINE LEARNING - 6363

PROJECT 1
LINEAR REGRESSION

Problem

- Classify the species of a flower based on its sepal length and width, and petal length and width.
- Three different species of flowers that need to be classified

Data

- A dataset named IRIS data is provided that contains attribute information about different flower species
- Link to the dataset: http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data
- Contains 150 samples:
  - Each data sample has 4 attributes:
    - Petal Length (cm)
    - Petal Width (cm)
    - Sepal Length (cm)
    - Sepal Width (cm)
  - Each sample is categorized to one of the following three species:
    - Iris-setosa
    - Iris-versicolor
    - Iris-virginica

Method

To approach the problem of classifying different species of flowers based on their leaves' lengths, a machine learning model called "Linear Regression" was used. The reason behind choosing linear regression is that the number of features in the dataset is small (4 features) and the features are correlated linearly. So, the dataset can be easily modeled with a linear function. To approximate this linear function, Linear Regression is used.

Linear Regression Model

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. For this task, I replaced the classes (species names) to continuous values **(-1, 0, 1)** and trained a linear regression model on the data.

While testing, I converted the predicted value back to the respective class based on **(-0.33, 0.33)** threshold. Anything left to this interval would be considered as “-1” class and anything right to the interval would be considered “1” class. Anything that falls in the interval would be considered “0” class. This way, three species of flowers can be classified using a regression model.

I used k-fold (**k=5**) cross validation to ensure that my model works correctly. For each fold, I randomly splitted 80% of the data samples into training data and the rest 20% into validation data.

Learning rate used for the model is: **0.001**
Number of iterations (epochs) the model was trained for: **1000**

Results

Based on a 5-fold cross-validation, the average classification **accuracy** obtained by my linear regression model is: **96.67%**

In the above result, each fold resulted in the following **accuracies**:
[1.0, 0.9667, 0.9667, 0.9333, 0.9667]

Mean Squared Error:

Average MSE = **0.0576 cm^2**

MSE for each fold was: **[0.041, 0.0674, 0.0539, 0.0651, 0.0606]**

NOTE: I have attached a pdf exported from jupyter notebook along with my code where I have printed my results when testing.