PRNN - 24, Assignment 1

Prof Prathosh A P

February 19, 2024

General Instructions:

- For Each of the problems you can find the data in 'Data' folder shared with you, traverse to the relevant folder, and choose the data corresponding to your group as per the group ID in the Excel sheet.(Each data is different and if you choose the wrong data you will be penalized)
- You are supposed to submit a single Jupiter notebook with all the solutions made into separate blocks.
- No ML library other than **numpy and matplotlib** should be used, failing which will attract zero marks.
- The final evaluation does not depend on the accuracy metrics but is based on the quality of your experiments and observations thereof.
- We will run a plagiarism check on both your report and the codes. Any suspicion of copying would lead to a harsh penalty from negative marks in the assignment to a failing grade in the course, depending upon the severity. Therefore, kindly refrain from copying others' codes and/or reports.

1 Regression Task

- **Q1**: Multilinear Regression - You need to find the position of a particle in 3D space (y1, y2, and y3) given a 10-dimensional feature vector. The features are readings from 10 sensors in the experiment environment. It is experimentally seen that the position of the particle depends linearly on these readings. Can you figure out the relationship?
- **Q2**: Generalised Regression with polynomial kernel - Now, it turns out that the position only depends on the magnitude of the force along 2 basis vectors (features 1 and 2). That is, the recordings in Q1 are derived quantities from these two independent features and share a polynomial relationship. Hence, use a polynomial kernel to predict the position of the particle given the 2 features.
- **Q3** Generalised Regression with non-polynomial kernel - The probability of rain on a particular day depends non-linearly on 5 satellite readings. You need to figure out the generalised regression function that predicts the probability of rain given these features.

  Note: The bias/offset term is not provided in the training data for any of these regression questions. And data is not to scale.

Implementations:

1. For Q1 - Implement a linear regression with 3 targets.
2. For Q2 - Implement a generalised regression with a polynomial kernel with 3 targets.
3. For Q3 - Implement a generalised regression with a non-polynomial kernel with 1 target.
4. You need to generate a correlation plot (Figure 1) for each regression task, i.e., 3 independent plots for Q1 and Q2 and 1 plot for Q3. Metrics - Pearson Correlation, mean squared error, mean absolute error.

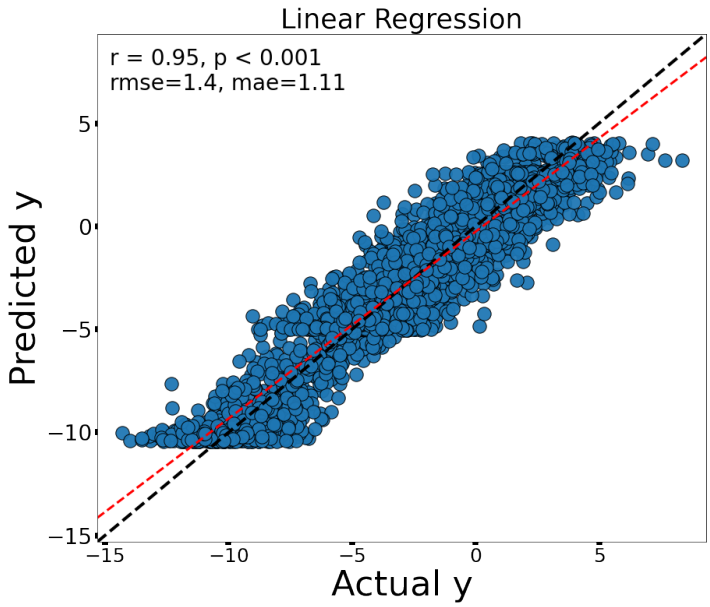

Figure 1: Shows performance of linear regression on arbitrary problem. The x-axis is the actual target, and the y-axis is the predicted target. The red line shows the line of best fit. The black line is the x=y line (prediction by oracle).

2 Classification Analysis

- **Q4** : Binary Classification : The data from 10 sensors planted at an industrial cite is tabulated. And depending on these sensor value we take a decision on the kind of product(one of 2 products). The task is to predict the product being produced by looking at the observation from these 10 sensors. Implement on all the relevant methods given below and for each of the method use all the relevant evaluation metrics. Data is present in /Data/binary-classification/. Last column is the label and remaining columns are the features.

- **Q5** : Multi class Classification problem (10 Classes): In the same industry from where we got the previous data has another big plant. There they have deployed 25 sensors and they produce 10 kinds of products. The task is to predict depending on the data from 25 sensors, predict which of these 10 products are. Implement on all the relevant methods given below and for each of the method use all the relevant evaluation metrics. Data is present in /Data/binary-classification/. Last column is the label and remaining columns are the features.

Implementations:

1. Bayes' classifiers with 0-1 loss assuming Normal, exponential, and GMMs (with diagonal covariances) as class-conditional densities. For GMMs, code up the EM algorithm,

2. Bayes' classifiers with non-parametric density estimators (parzen window) with 2 different kernels.

3. K-nearest neighbour classifiers with different K-values and 2 different distance metrics (Euclidean and Cosine-distances)

4. Linear classifier (O v R incase of multi class)

The metrics to be computed are:

1. Classification accuracy

2. Confusion matrix
3. Class-wise F1 score
4. RoC curves for any pair of classes
5. likelihood curve for EM with different choices for the number of mixtures as hyper-parameters,
6. Emipiral risk on the train and test data while using logistic regressor.