PRNN 2023 - Assignment1

Prathosh A. P.

10 February 2023

1 Regression Analysis

- **p1:** In this problem, the task is to predict the current health (as given by the target variable) of an organism given the measurements from two biological sensors measuring their bio-markers (negative indicates that it is lesser than the average case). With this data, you are expected to try our linear regression models on the training data and report the following metrics on the test split: (a) Mean Squared Error, (b) Mean Absolute Error, (c) p-value out of significance test.
  DATA: p1train/test.csv
- **p2:** Here, you are expected to predict the lifespan of the above organism given the data from three sensors. In this case, the model is not linear. You are expected to try several (at least 3) non-linear regression models on the train split and report the following metrics on the test split (a) Mean Squared Error, (b) Mean Absolute Error, and (c) p-value out of significance test.
  DATA: p2train/test.csv

2 Multi-class classification

- **p3:** We have data from 10 sensors fitted in an industrial plant. There are five classes indicating which product is being produced. The task is to predict the product being produced by looking at the observation from these 10 sensors. Given this, you are expected to implement (a) Bayes' classifiers with 0-1 loss assuming Normal, exponential, and GMMs (with diagonal co-variances) as class-conditional densities. For GMMs, code up the EM algorithm, (b) Linear classifier using the one-vs-rest approach, and (c) Multi-class Logistic regressor with gradient descent. The metrics to be computed are - (a) Classification accuracy, (b) Confusion matrix, (c) Class-wise F1 score, (d) RoC curves for any pair of classes, and (e) likelihood curve for EM with different choices for the number of mixtures as hyper-parameters, (f) Emipiral risk on the train and test data while using logistic regressor.
  DATA: p3train/test.csv

- **p4**: In this problem, we consider an image dataset called Kannada-MNIST. This dataset contains images (60,000 images with 6000 per class) of digits from the south Indian language of Kannada. The task is to build a 10-class classifier for the digits. You are supposed to test the following classification schemes: (a) Naive Bayes' with Normal as Class conditional, (b) Logistic regressor with gradient descent, and (c) Multi-class Bayes' classifier with GMMs with diagonal co-variances for class conditionals. Report the following metrics on the test data: (a) Classification accuracy, (b) Confusion matrix, (c) Class-wise F1 score, and, (d) RoC curves for any pair of classes, (e) likelihood curve for EM with different choices for the number of mixtures as hyper-parameters, (f) Emipiral risk on the train and test data while using logistic regressor. In this problem, first split the data into train and test parts with the following ratios of 20:80, 30:70, 50:50, 70:30, and 90:10, and record your observations. Train the algorithms on the train part and evaluate over the test part.
  DATA:images.zip
- **p5:** In this part, the data from the previous problem is 'condensed' (using PCA) to 10 dimensions. Repeat the above experiment with all the models and metrics and record your observations.
  DATA:KannadaMNISTPCA.csv

General Instructions:

1. All the data files can be found here - data
2. For **p1:p3**, the last column in the csv file in the target variable, for **p4** it is the folder name and **p5**, it is the first column of the csv file.
3. You are supposed to submit a single Jupiter notebook with all the solutions made into separate blocks.
4. No ML library other than **numpy** and **matplotlib** should be used, failing which will attract zero marks.
5. A 4-6 page report has to be submitted that would list all the experiments, results, and your observations. It should be in double-column format in latex as specified here template. IISc has a subscription to overleaf and the report should be in the exact same format.
6. Use matplotlib for plotting the loss and RoC curves.
7. The final evaluation **does not** depend on the accuracy metrics but is based on the **quality of your experiments and observations thereof**.
8. 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.