SML

Assignment 3

Mar 2024

1 Instructions

- You can use inbuilt libraries for Math, plotting, and handling the data (eg. NumPy, Pandas, Matplotlib).
- Usage instructions for other libraries can be found in the question.
- Only (*.py) files should be submitted for code.
- Create a (*.pdf) report explaining your assumptions, approach, results, and any further detail asked in the question.
- You should be able to replicate your results during demo.

---

2 Question-1

Use https://storage.googleapis.com/tensorflow/tf-keras-datasets/mnist.npz MNIST dataset for this question and select class 0, 1 and 2. Note you are not allowed to use libraries which can take data, fit the model, predict the classes and give accuracy. Perform following tasks.

- Apply PCA and reduce the dimension to $p = 10$. You can use the entire train set of these 3 classes to obtain PCA matrix. For the remaining parts, use the reduced dimension dataset.
- Now learn a decision tree using the train set. You need to grow a decision tree with 3 terminal nodes. This is similar to what we did in the baseball salary example. For the first split, consider all p dimensions. For each dimension, consider one split which will divide the space into two regions. Find the total Gini index. Similarly find the total Gini index for all 50 dimensions. Find the best split by searching for minimum Gini index. Suppose, you split across 10th dimension. Choose one of the splits, and repeat the steps to find best split. Once you find it, the entire p dimensional space is divided into three regions. [2]

- Find the class of all samples in test set of these 3 classes. For a particular test sample, check where the samples lies in the segmented space. The class for a particular sample is the class of sample which is in majority in the region to which the test sample belongs. Report accuracy and class-wise accuracy for testing dataset. [1]

- Now use bagging, develop 5 different datasets from the original dataset. Learn trees for all these datasets. For test samples, use majority voting (atleast 3 trees should predict the same class) to find the class of a given sample. In case there is a tie, that is two trees predict one class and other two trees predict another class, then you can choose either of the classes. Report the total accuracy and class-wise accuracy. [1]