SML Assignment-3

Shobhit Raj (2022482)

1 Approach

The tasks include dimensionality reduction using Principal Component Analysis (PCA), decision tree learning with a maximum of 3 terminal nodes, and ensemble learning using bagging with decision trees.

1.1 Data Preparation and PCA Dimensionality Reduction

Samples belonging to classes 0, 1, and 2 are filtered out from the training and testing sets, and the dataset is preprocessed for further analysis. Principal Component Analysis (PCA) is applied to reduce the dimensionality of the training set to 10 principal components. The PCA process involves calculating the covariance matrix, eigenvalues, and eigenvectors, followed by selecting the top eigenvectors corresponding to the highest eigenvalues to form the PCA matrix.

1.2 Decision Tree Learning

Decision trees are learned with a constraint of 3 terminal nodes. The process involves finding the best splits across all dimensions to minimize the Gini index. For each split, the dataset is divided into two regions, and the process is repeated recursively until the entire space is partitioned into three regions.

1.3 Classification on Testing Dataset and Accuracy Evaluation

The trained decision tree model is used to predict the classes of samples in the testing dataset. The majority class in the region to which each test sample belongs is determined, and that class is assigned as the predicted class. Accuracy and class-wise accuracy are calculated for the testing dataset based on the predictions.

1.4 Ensemble Learning using Bagging

Bagging is employed to develop 5 different datasets from the original dataset by randomly sampling with replacement. Decision trees are learned for each dataset, and predictions are aggregated using majority voting among the trees. The predicted class for a test sample is determined by the majority vote from at least 3 trees. Total accuracy and class-wise accuracy are reported for the ensemble model.

2 Results

The overall accuracy achieved on the test set is calculated to be approximately 80.90%. Additionally, class-wise accuracy is reported, showing the performance of the classifier for each digit class. The accuracy for each class is as follows:

- Class 0 accuracy = 99.08163265306122%
- Class 1 accuracy = 90.30837004405286%
- Class 2 accuracy = 53.29457364341085%
- Overall accuracy = 80.90244677470606%

With Bagging :-

- Class 0 accuracy $\approx$ 99%
- Class 1 accuracy $\approx$ 90%
- Class 2 accuracy $\approx$ 54%
- Overall accuracy $\approx$ 80-81%

Hence, the bagging accuracy is close to the single decision tree's accuracy.