SML Assignment 4
Report

Aditya Prasad
2022036

Q1)

Assumptions:

- Data quality: Assuming the MNIST dataset is clean and properly preprocessed.
- Feature representation: Assuming PCA is an appropriate method for dimensionality reduction.
- Algorithm assumptions: Assumption that boosting decision trees will lead to improved classification accuracy.

Approach:

- Data Loading and Preprocessing
  - Utilization of the MNIST dataset.
  - Selection of specific digit classes (0 and 1).
  - Stratified train-validation-test split for model evaluation.
  - Option to choose a random subset of training data.
- Dimensionality Reduction with PCA
  - Reduction of feature dimensions to improve computational efficiency.
  - PCA applied to training, validation, and test datasets.
- Model Training and Evaluation
  - Implementation of a Boosted Decision Tree algorithm.
  - Iterative addition of decision trees with adaptive boosting.
  - Training over multiple epochs.
  - Evaluation of model performance on validation and test sets.

Result:

```
Training: 100%|██████████| 50/50 [16:14<00:00, 19.49s/it]
Validation Accuracy: 98.18%
Test Accuracy: 98.47%
```

Q2)

Assumptions:

- Data integrity: Assuming the MNIST dataset contains reliable and properly preprocessed data.
- Algorithm suitability: Assuming that Gradient Boosted Decision Trees are appropriate for regression tasks on this dataset.

- Parameter tuning: Assuming the default and chosen parameters are suitable for the problem.
-

Approach:

- Decision Tree Regression
    - Implementation of a Decision Tree regressor to fit the training data.
    - Recursive partitioning of feature space based on best split points.
    - Calculation of residuals and updating them iteratively.
- Gradient Boosting
    - Utilization of Gradient Boosting for ensemble learning.
    - Addition of decision trees sequentially to minimize residuals.
    - Training over multiple epochs.
- Data Loading and Preprocessing
    - Loading the MNIST dataset.
    - Selection of specific digit classes (0 and 1).
    - Train-validation-test split for evaluation.
    - Reshaping of data for compatibility with the algorithm.
- Dimensionality Reduction with PCA
    - Dimensionality reduction using Principal Component Analysis (PCA).
    - Application of PCA to training, validation, and test datasets.

Result:

```
Validation MSE: 0.24411788737442988
Validation MSE: 0.23935634940175046
Validation MSE: 0.23468959225176034
Validation MSE: 0.23011572952391748
Validation MSE: 0.2256329123596728
Validation MSE: 0.22123932869535753
Validation MSE: 0.21693320252993933
Validation MSE: 0.21271279320735045
Validation MSE: 0.20857639471309727
Validation MSE: 0.2045223349848677
Validation MSE: 0.20054897523685794
Validation MSE: 0.19665470929754536
Validation MSE: 0.1928379629606398
Validation MSE: 0.18909719334895117
Validation MSE: 0.18543088829091658
Validation MSE: 0.1818375657095345
Validation MSE: 0.17831331849593948
Validation MSE: 0.17486132647788186
Validation MSE: 0.17147564733002646
Validation MSE: 0.16816514496057988
Validation MSE: 0.16491235870338983
Validation MSE: 0.16173197165936995
Validation MSE: 0.15860686486046882
Validation MSE: 0.15555148524871443
Validation MSE: 0.15254904894484314
Validation MSE: 0.14961376926987954
Validation MSE: 0.1467291909487842
Validation MSE: 0.14390929625302767
Validation MSE: 0.14113795214974972
Validation MSE: 0.1384289124619333
Validation MSE: 0.1357663601571756
```

```
Validation MSE: 0.004842296160238968
Validation MSE: 0.004828733405897978
Validation MSE: 0.0048137079231504905
Validation MSE: 0.004800674395304189
Validation MSE: 0.004786201326483359
Validation MSE: 0.004773676591713359
Validation MSE: 0.004759735248284443
Test MSE: 0.0031961953713614266
```