SML assignment-3 Report

Ritvik Shekhar-Roll no. 2023440

March 31, 2025

1 Introduction

This report describes the implementation of a decision tree, Bagging, and random forest algorithm, that is, Q3 and Q4. It also discusses Q5.

2 Q4 and Q5 - Data Description

The training data consist of 8 samples with 5 features:

- 1 index-Age(Numerical value)
- 2 index-Income(low-0 ,medium-1 ,high-2)
- 3 index-Student (no-0,yes-1)
- 4 index-Credit Rating (fair-0, excellent-1)
- 5 index-Label (Target variable)

3 Decision Tree Implementation

Decision tree is constructed based on the following steps:

1. Compute potential splits using the features.
2. Calculate the Gini index for possible splits.
3. Find the best split that minimizes the Gini index.
4. Recursively build the tree until reaching a depth = 2 / node becomes pure / node has 2 or fewer number of samples.

3.1 Splits

Best split is selected on the basis if Gini index For splits, my approach is as follows:

1. For numerical attributes (e.g., Age): I calculate all possible split points and determine the best one. Values less than or equal to the chosen split go to the left, while greater values go to the right.

2. For categorical attributes (e.g., Income, Student, and Credit Rating): I group instances with the same category on the left and those with different categories on the right.

3.2 Gini Index Calculation

The Gini index is calculated as:

$$Gini = \sum P_{mk}(1 - P_{mk}) \tag{1}$$

where P_{mk} is the probability of class m in node k.

4 Bagging

To reduce the variance, bagging (Bootstrap Aggregating) is employed. In the code I have trained 10 models on 10 bootstrap samples of the dataset and averaging their predictions. The key steps include:

- Generate multiple bootstrap samples from the original dataset.
- Train a separate decision tree on each sample.
- Aggregate the predictions using majority voting for classification tasks.

4.1 Out-of-Bag (OOB) Error

The OOB error is calculated as: we have 8 sample so it can happen in random sampling with replacement that some samples are left so they are used for test purpose and the tree is trained on the sampled dataset

$$OOBError(for - binary - classificaton) = \frac{\sum(\hat{f}_i - f_i)^2}{N} \tag{2}$$

$$if - missclasified = (\hat{f}_i - f_i) = 1/-1 \tag{3}$$

$$correctly - classified = (\hat{f}_i - f_i) = 0 \tag{4}$$

where $\hat{f}_i$ is the predicted label from tree trained on sampled dataset and f_i is the actual label.(which will be 0/1 in this case)

5 Random Forest Implementation

A random forest is built using bagging (bootstrap aggregating):

- Generated 10 bagged datasets using bootstrap sampling.
- Constructed 10 decision trees with a random subset of features.
- Computed Out-of-Bag (OOB) error.
- Performed majority voting to get final predictions.

5.1 Out-of-Bag (OOB) Error

Same as done in 4.1

6 Results

6.1 Decision Tree Results

A decision tree is trained on the dataset, and predictions are made for a sample with the following features:

- Age: 42
- Income: 0
- Student: 0
- Credit Rating: 1
- Depth 0: Best Split → Feature=2, Threshold=2
- Depth 1: Best Split → Feature=1, Threshold=47.5
- **By tree predicted sample would belong to [2, 3, 4] node which gives YES as answer i.e. the person will buy computer**

6.2 Random Forest and Bagging Result

After training and testing the random forest and bagged tree:

- **OOB of random forest is less than the OOB of bagged tree**

7 K fold Cross Validation

The evaluation of a regression model using 5-fold cross-validation on synthetic data generated from a known function. The goal is to determine the optimal polynomial degree that fits the model.

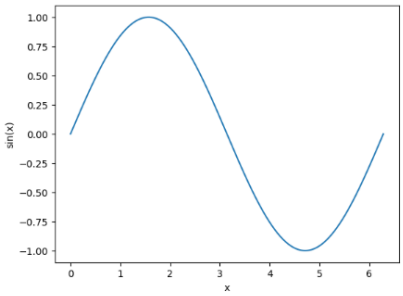

Figure 1: Sine Function

8 Polynomial Regression Models

Polynomial regression models of degrees 1 to 4 are considered. Each model is trained and evaluated using 5 fold cross-validation. **The best model is selected according to Mean squared error.**

9 5-Fold Cross-Validation

The dataset is split into 5 folds i.e each fold contains 20 dataset, and the following steps are performed:

1. Divide the dataset into 5 folds.
2. Train the model on 4 folds and test on the remaining fold.
3. Repeat the process for all 5 fold combinations.
4. Compute the average error to assess model performance.
5. Identify the polynomial degree that yields the lowest cross-validation error.

10 Results and Visualization

The best model is 3deg polynomial . For 3 degree, the following plots are generated:

- The true function $y = \sin(x)$.
- The noisy training points.
- The regression model's predictions.

These visualizations help in understanding how well the model approximates the underlying function.

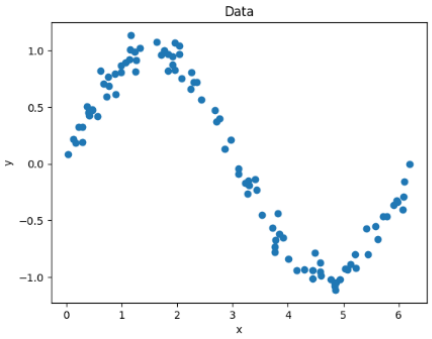

Figure 2: Noisy training points

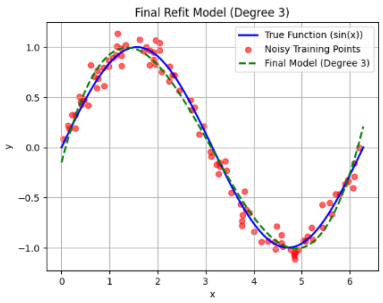

Figure 3: I refitted the 3 deg polynomial model on the entire dataset given for training and plotted for new data points. The Green curve represents the 3 deg model that predicts for new data points. Blue line is true function and orange dot is noisy data points.

11 Conclusion

- Using 5-fold cross-validation, the **optimal polynomial degree is 3** based on minimizing squared error. The results demonstrate the balance between underfitting and overfitting in polynomial regression.
- **Deg 1,2 will under-fit the sin function and the deg 4 will overfit the model.**