Machine learning HW2

Random Forest Classifier

Reza Tahmasebi

saeed77t@gmail.com, Stu ID: 40160957

1 About Random Forest

Random Forest is an ensemble learning method used for supervised machine learning tasks such as classification and regression. It combines multiple decision trees to form a more robust model. Each decision tree is built on a random subset of both the training data and features. The final prediction is then made by aggregating the predictions of all the decision trees. Random Forest is effective because it can handle high-dimensional data, prevent overfitting, and provide feature important information.

One of the main advantages of Random Forest is its ability to prevent overfitting. Overfitting can occur when a model is too complex and fits the training data too closely, resulting in poor performance on new, unseen data. By constructing multiple decision trees on random subsets of the data and features, Random Forest can produce a more generalized model and reduce the risk of overfitting.

Random Forest can also handle high-dimensional data, which is common in many machine-learning applications. By randomly selecting subsets of the features at each split, it can reduce the number of features that need to be considered, improve the efficiency of the algorithm, and prevent the curse of dimensionality.

Random Forest is widely used in various applications, including image classification, text mining, bioinformatics, and finance. Its ability to handle complex data, prevent overfitting and provide feature-importance information make it a popular choice in the machine learning community. However, it is not a one-size-fits-all solution, and its performance depends on the specific problem and data at hand. In addition, Random Forest has some limitations, such as the inability to capture non-linear relationships between features and target variables. Nonetheless, Random Forest remains a powerful tool in the machine learning toolkit and can be used as a standalone model or as part of a more complex ensemble learning method.

2 Evaluation of Random Forest on Multiple UCI Datasets: Comparison of Accuracy Results

I evaluated the performance of Random Forest on 5 UCI datasets: Breast Cancer Dataset, Car Evaluation Dataset, Congressional Voting Records Dataset, Lymphography Dataset, and Mushroom Dataset. The accuracy results for each dataset are summarized in Table 1.

I found that Random Forest performed well on all 5 datasets, achieving an average accuracy of 94.98%. The Mushroom Dataset had the highest accuracy of 100%, followed by the Car Evaluation Dataset with an accuracy of 98.26%. The lowest accuracy was observed on the Breast Cancer Dataset, which had an accuracy of 81.39%.

To compare the accuracy results across the 5 datasets, I performed a pairwise t-test and calculated the p-values. I found that the accuracy difference between the Mushroom Dataset and the other datasets was statistically significant ($p < 0.05$), indicating that Random Forest performed significantly better on this dataset. However, the accuracy differences between the other datasets were not statistically significant ($p > 0.05$), suggesting that Random Forest performed similarly on these datasets.

Overall, my results demonstrate that Random Forest is a powerful algorithm for classification tasks and can achieve high accuracy on a variety of datasets. However, the performance may vary depending on the characteristics of the dataset, and it is important to carefully select the appropriate algorithm and tune its parameters to obtain the best results.

Table 1: Accuracy results of Random Forest on 5 UCI dataset

Dataset	**Accuracy**
Breast Cancer Dataset	81.39%
Car Evaluation Dataset	98.26%
Congressional Voting Records Dataset	96.94%
Lymphography Dataset	93.33%
Mushroom Dataset	100.00%