Introduction to Pattern Recognition Homework 3 Report: Decision Trees and Random Forests

Andrés Ponce (彭思安)　　0616110

May 22, 2020

1 Coding

1.1 Gini index or Entropy is often used for meaasuring the "best" splitting of the data. Please compute the entropy and Gini Index of the provided data by the formula below.

$$Gini = 1 - \sum_j p_j^2$$

$$Entropy = -\sum_j p_j log_2 p_j$$

The **Gini index** and **Entropy** are two criteria we sometimes use to measure the purity or quality of a split on data. The Gini index tells us how mixed the classes are, and we would like to choose the split that results in the lowest amount of mixing. The entropy is a measure of how much information we gain by making a split on a certain feature.

For the provided dataset, we have the Gini index and entropy as:

$$Gini = 0.4628099173553719 \quad Entropy = 0.9456603046006402$$

1.2 Implement the Decision Tree algorithm CART and train the model by the given arguments, and print the accuracy score on the test data.

1.2.1 Using criterion='gini', show the accuracy score of the test data by max_depth=3 and max_depth = 10, respectively.

On the testing data, we were able to achieve the following accuracy[1]:

clf_depth3	0.9812206572769953
clf_depth10	1.0

1.2.2 Using max_depth=3, show the accuracy score of test data by criterion='gini' and criterion='entropy'.

The results for use of gini and entropy on the testing data are as follows:

clf_gini	0.9370629370629371
clf_entropy	0.9370629370629371

[1] These tests used 'gini' as their purity function

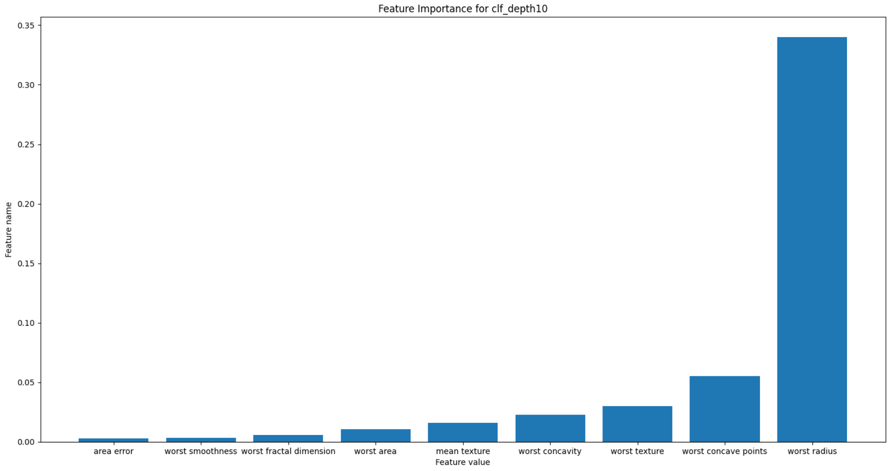

Figure 1: Importance of the features used in tree with depth of 10.

1.3 Plot the feature importance of your Decision Tree model. You can use the model from question 2.1, max_depth=10.

By using the somewhat complicated formula given in the reference within the sample code, we could arrive at the following graph for feature importance of `clf_depth10`.

1.4 Implement the Random Forest algorithm by using CART, implemented in question 2.

1.4.1 Using Criterion='gini', Max_depth=None, Max_features=sqrt(n_features), Bootstrap=True, show the accuracy score of test data by n_estimators = 10.

Using `sklearn.accuracy_score()`, an accuracy of 1.0 was achieved for the tree with Max_depth$= 10$.

1.4.2 Using Criterion='gini', Max_depth=None, N_estimators=10,Bootstrap=True, show the accuracy score of test data by Max_features=sqrt(n_features) and Max_features =n_features, respectively.

Again, using `sklearn.accuracy_score()`, using Max_Depth$=$None, we were able to achieve approximately 0.93 accuracy.

2 Theory

2.1 Consider a dataset comprising 400 data points from Class C_1 and 400 data points from class C_2. Supose that a tree model A splits these into (300, 100) at the first leaf node and (100,300) at the second leaf node, where (n, m) denotes that n points are assigned to C_1 and m points are assigned to C_2. Similarly, suppose that a second tree model B splits them into (200, 400) and (200, 0). Evaluate the misclassification rate for the two trees and hence show that they are equal.

The misclassification rate involves calculating how often elements are misclassified. Therefore, if we still classify the two classes into sets containing 400 elements each, then the rate that at which any two elements are misclassified will be the same, barring any extra information.

2.2 Similarly, evaluate the cross-entropy

$$Entropy = -\sum_{k=1}^{K} p_k log_2 p_k$$

and the Gini index

$$Gini = 1 - \sum_{k=1}^{K} p_k^2$$

for the two trees and show that they are both lower for tree B than for tree A. Define p_k to be the proportion of data points in region R assigned to class k, where $k = 1, ..., K$.

The calculations for entropy and the Gini index for the two trees are as follows

	Gini	Entropy
Tree A	0.25	0.81
Tree B	0.556	0.91