Introduction to Artificial Intellligence Homework 4: Random Forests

Andrés Ponce (彭思安) 0616110

June 28, 2020

1 Introduction

For the fourth and last homework assignment, the task involved making a decision tree that could classify data. This approach works by having a tree, in this case a binary tree, in which internal node splits the dataset. Once we reach a leaf node, we have a class label which we assign to the data point.

The process for choosing the best attribute of the dataset to split on invovlves trying every possible attribute in the dataset and every possible threshold value. For a given threshold value, we have to compare the purity of the resulting split dataset. There are several methods to check the purity of a dataset, however for this assignment the Gini impurity is utilized. This metric relies on the percentage of the data points that belong to a certain class, and is given by the formula

$$G(x) = 1 - \sum_{n=1}^{N} p(C_j) \tag{1}$$

where $p(C_j)$ refers to the probability of a data point belonging to class C_j. This is just a simple division of the amount of values with label j and all the amounts in the given dataset. Since we split the dataset into two, those with a threshold value greater and lower than the threshold currently being tested. Then the sum of the gini of the left dataset and the right dataset have to be taken into account when deciding if one threshold leads to a purer split.

As for the dataset in use, the famous iris dataset is used, which provides information concerning sepals and pedals of 150 iris flowers. There are three different class labels present. One of them is linearly separable from the other ones. The other two have to be separated with a guess from the tree.

2 Procedure

3 Conclusion