Worksheet 2 – Information Theory and Decision Trees

Theory

Review the videos from week 2. Answer the following questions based on those lectures

1. Describe the main elements of a decision tree and its theory of operation
2. Explain the term “entropy” from information theory? How is it calculated?
3. What is meant by “information gain” and how is it used by an algorithm like ID3?
4. What is data partitioning in the context of decision tree construction?
5. Briefly describe some of the main advantages and disadvantages of decision trees as a machine learning approach
6. What is meant my “bootstrapping” and “subspace sampling”?
7. What is a model ensemble and what are its potential benefits?
8. Describe the operation of a random forest. What role does subsampling play?

Practice

Follow the tutorial videos from week 2 and carry out the following steps

1. Download the code archive and extract the file from the week 2 learning materials. Make sure that you can run the examples code as provided.
2. Use the code in **tree.py** as the starting point to build a full implementation of ID3 in in Python
3. Train and test your implementation against the heart diagnosis dataset. You should include all of the available What accuracy does your implementation achieve?
4. Run the **sklearn\_tree.py** implementation and compare with your result
5. An alternative measure to entropy for implementing decision trees is to use the **Gini Index**. The Gini index for a dataset **D** with respect to a target variable **t** over a feature value domain of **V** is defined as:

The Gini Index can be thought of as the probability that a feature would be misclassified and always has a value between 0 and 1. The advantage is that this is less intensive than calculating entropy because there is no logarithm computation required.

Modify your decision tree implementation to use the Gini Index instead of entropy. The calculation of information gain is similar. It is the Gini Index of the whole dataset minus the weighted Gini Index values for each partition.

* 1. Extend your decision tree implementation to implement **bootstrapping** and **subspace sampling**. Use these extensions to train and test a random forest for a set of different subspace sample sizes using the heart dataset. Compare your resulting model with the one in **sklearn\_forest.py**. How do they compare?