Week 2

**Theory：**

1. **Describe the main elements of a decision tree and its theory of operation**

Main elements include a root node, some interior nodes and leaf nodes.

root node: Contains the complete set of samples;

internal node: Corresponding feature attribute test.

leaf node : Represents the outcome of a decision.

During prediction, a certain value is used to make judgment at the inner node of the tree, and the branch node is determined according to the judgment result, until it reaches the leaf node, and the classiﬁcation result is obtained.

1. **Explain the term “entropy” from information theory? How is it calculated?**

In information theory, entropy is the average amount of information contained in each message received, also known as information entropy, source entropy, average self-information.

Entropy is actually the mathematical expectation of multiplying and summing up the number of bits and the probability of a random variable.

1. **What is meant by “information gain” and how is it used by an algorithm like ID3?**

Information gain is a concept in information theory. In order to explain its concept, the concept of information entropy is ﬁrstly deﬁned. Information entropy reﬂects the uniformity of the distribution of any kind of energy in space. Another term for information gain is mutual information.

The algorithm ID3 is based on information theory and takes information entropy and information gain as the measurement standard to realize the induction and classiﬁcation of data. First of all, the problem that ID3 algorithm needs to solve is how to choose features as the standard for partition data sets.

1. **What is data partitioning in the context of decision tree construction?**

According to information gain，partition the data set according to a given feature

1. **Briefly describe some of the main advantages and disadvantages of decision trees as a machine learning approach**

advantage:

1. Decision trees are easy to understand and interpret, can be analyzed visually, and easy to extract rules.

2. You can process both nominal and numeric data.

3. It is more suitable to deal with the samples with missing attributes.

4. Ability to handle unrelated features.

5. When testing the data set, it runs faster.

6. Capable of producing feasible and eﬀective results for large data sources in a relatively short me.

disadvantage:

7. Overfitting is easy to occur (random forest can greatly reduce overﬁtting. 8. It is easy to ignore the correlation of attributes in a dataset.

9. For the data with inconsistent number of diﬀerent kinds of samples, diﬀerent decision criteria will bring diﬀerent selection tendency of attributes when the attributes are divided in the decision tree.

10. When the information gain is calculated by ID3 algorithm, the result tends to be more numerical

1. **What is meant my “bootstrapping” and “subspace sampling”?**

In statistics, the bootstrap is any test or measure that relies on random sampling and substitution.

In machine learning the random subspace method, also called attribute bagging or feature bagging, is an ensemble learning method that attempts to reduce the correlation.

1. **What is a model ensemble and what are its potential benefits?**

A model ensemble is a model build from a set of cooperating models.

In model ensemble, each model is built separately and independently of the training data, usually involving a random sampling approach. The aggregated output of the model ensemble is the consideration of each of the ensemble’s member predictions and choosing the best result across these.

1. **Describe the operation of a random forest. What role does subsampling play?**

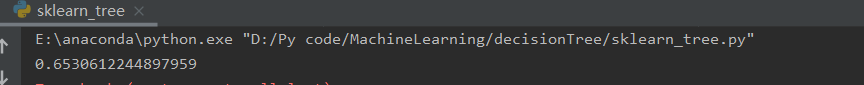
A random forest is an example of a model ensemble. The idea is to generate multiple random decision trees, ideally hundreds or even thousands depending on the dataset size, using random bootstrapping and random subsampling. Combining and comparing the outputs of aggregated bootstrapped models is called bagging.

**Practice：**

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**

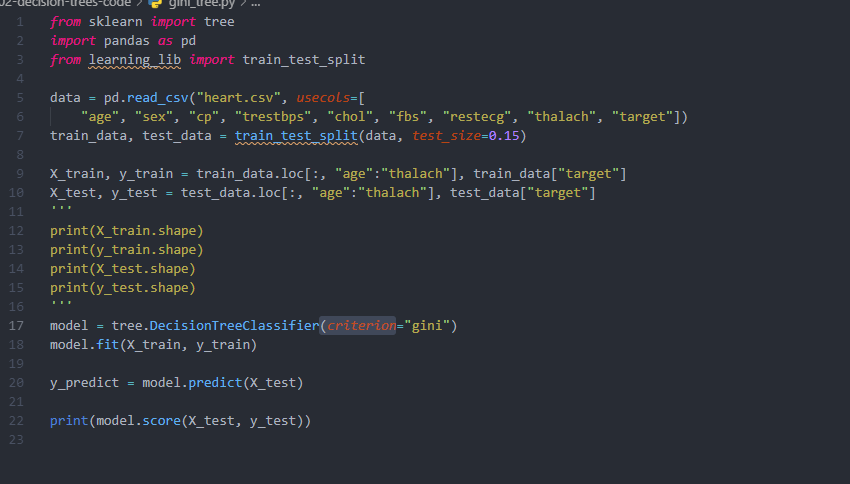
I try my best to learn something about the ID3, the file name is test.py. but there are also some problems I can’t handle.

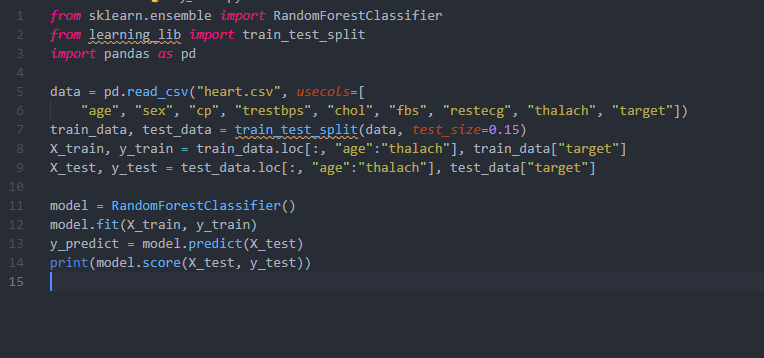
1. **Train and test your implementation against the heart diagnosis dataset. You should include all of the available What accuracy does your implementation achieve?**
2. **Run the sklearn\_tree.py implementation and compare with your result**



1. **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?**