David Strickland

CS5402 Intro to Data Mining

Homework 2

Python file located at the following github repository: <https://github.com/ds5b4/CS5402-IntroToDataMining-Fall2018>

Task 1:

Classification is the process of learning to group of data into predefined classes. Clustering on the other hand seeks to group similar data points based on one or more of their features. Classification can be achieved through a decision tree based on what commonalities the features of the data of each class have. Clustering requires more advanced models and techniques like nearest-neighbor to group the data well.

Task 2:

Entropy in a data mining/statists sense is the measure of impurity or how well the classification/decision tree divides the data. For instance the entropy of a single class with ten C0 elements and ten C1 elements is 1 (the maximum value for a measure of entropy) while if there is a split which leads to a group with ten C0 elements in the first group and ten C1 elements in the second group then the entropy is now 0 meaning the groups are perfectly pure.

Task 3:

Gain is the difference between the impurity before a split and the impurity after the split. Gain ratio is . This is used to prevent splits which create a too many groups such that each group loses significance. Gini index is a measure of impurity, much like entropy, but with a maximum value of .5 rather than 1. This makes the gini index less sensitive than entropy but still useful.

Task 4:

ID3: information gain

C4.5: Gain Ratio

CART: Gini Index

Q1: Binary split on ‘Is Opponent in AP Top at Preseason’.

By Hand:

‘Is Opponent in AP Top at Preseason’ == Out

5 Win 0 Loss

‘Is Opponent in AP Top at Preseason’ == In

0 Win 1 Loss

Gini Index(Before) =

Gini Index(After, Out) = 1 – 1 = 0

Gini Index(After, In) = 1 – 1 = 0

ID3 = .27777 – (5/6 \* 0 + 1/6 \* 0) = .27777

C4.5 = = = = -1.419539

Python Generated:

Q2: