PA-2 REPORT-DT

Name: Abhay Singh

ID: 1001669333

* describe the Decision Tree method.

A decision tree is a popular method of creating and visualizing predictive models and algorithms. Decision trees tend to be the method of choice for predictive modeling because they are relatively easy to understand and are also very effective. The basic goal of a decision tree is to split a population of data into smaller segments. There are two stages to prediction. The first stage is training the model—this is where the tree is built, tested, and optimized by using an existing collection of data. In the second stage, you actually use the model to predict an unknown outcome. We’ll explain this more in-depth later in this post.



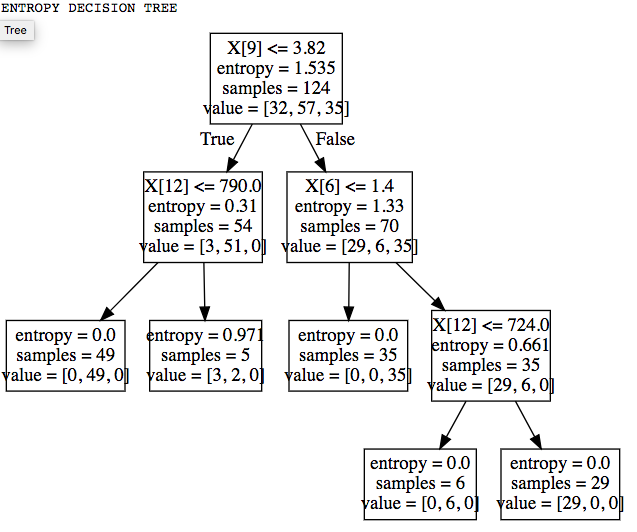
In the table above, Column Z is the target indicator; the piece of information that is being predicted by the model. Alternate terms: class, predicted variable, target variable

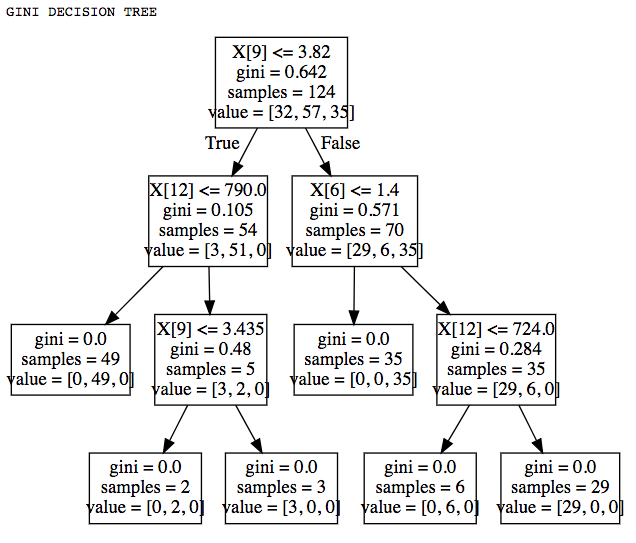
The data in columns A, B, C, and so on are called indicators. Alternate terms: feature, dimension, variable

Rows 1, 2, and 3 represent what we refer to as atoms. Alternate terms: instances, examples, data points

A mouthful, to be sure, but the simple calculations of entropy followed by the required number of information gain comparisons is enough to tell us how much is actually gained if we were to branch on a given attribute,

* the dataset and if you have done any pre-process. your code.
  + The instaltion of graphviz and pydotplus is required.
    - pip -install pydotplus
    - pip -install graphviz
  + The data in Wine dataset is Bunch. Preprocessing-> Convert to Series and Dataframe od Target and Data of data.
    - da1=pd.DataFrame(data.data, columns=data.feature\_names)
    - da1['target'] = pd.Series(data.target)
* Visualization of the decision tree for gini and entropy.[8 points]





* Interpret your results, and do not forget to compare gini and entropy [10 points]
  + From the decision tree above, we can say that both are good for presentation. We can observe a differences with the GINI in second iteration.
  + Gini is intended for continuous attributes and Entropy is for attributes that occur in classes
  + Gini is to minimize misclassification
  + Entropy is for exploratory analysis
  + Entropy is a little slower to compute for big dataset.

Reference:

<https://www.aunalytics.com/2015/01/30/decision-trees-an-overview/>

https://datascience.stackexchange.com/questions/10228/gini-impurity-vs-entropy