Data Mining Project 2

Alay Parikh (1001774636)

Vianny D'souza (1001770237)

Report For DT and NB

Q-1 Describe the Decision Tree methods, and Naive Bayes classifier. Dont copy paste it from the internet. Write it on your own.

Decision tree methods:

In the decision tree we calculate the information gain to split a node. There are 2 methods to measure impurity : gini and entropy. Entropy is measurement of impurity in data. This is to identify which feature provides max information about classification. Whereas the gini index is the amount of probability of a feature that will be incorrectly classified when selected randomly.

Naive Bayes classifier:

- This means to classify the data based on certain features.
- It uses bayes algorithm :
  Probability of A happening when B has occurred. B is evidence and A is hypothesis.

  $P(A/B)= P(B/A)P(A)/P(B)$

- The features here are independent, that means presence of one feature does not affect the other so it is called naive.

Q-2 Describe the datasets like what do you understand from the dataset?

It uses various attributes about different people such as active life style, Smoking, Alcohol, height, weight and things like that. Which is used to predict weather the person is a cardiac patient or not. The other data set is Titanic Dataset which uses various attributes like Name, Ticket, Embarked, Etc. to show weather person has survived or not.

If you have done any pre-processing , and your code, please write down your observation.

For Titanic Dataset, We had to perform various pre-processing operations to eliminate redundant data. One of those operation was to find out weather there are many columns with null attributes. Such columns are delated from the dataset

Q-4. Visualization of the decision tree for gini and entropy.

1. Ap_hi attribute has the highest entropy of 1.0 Hence, it is taken as a root node.
2. The left child is taken as Age, as it is the second highest (entropy of 0.901)
3. The right child of Ap_hi is the third highest. As the entropy of it is 0.791.

Q-5. Interpret your results, compare gini and entropy.

We have compared both Gini and Entropy. And accuracy for Gini has come upto 0.5614 and for Entropy has come upto 0.7322.

Gini is the probability of a random sample being classified incorrectly. It is calculated by subtracting the sum of each class from 1.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Entropy is measurement of impurity in information. Here we minus the entropy of parent node minus the entropy of the child node.

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

Q-6. Visualize the dataset, for the target variable.

We have even generated classification reports for both Gini and Entropy. F-1 Scores for cardio = No in Gini is about 0.68 where as in Entropy it is about 0.75. Also, F-1 score for cardio = Yes in Gini is around 0.29 where as in Entropy it is about 0.72.

Citations

We referred to https://scikit-learn.org/ website to understand and check various libraries and how they are implemented.