Data Warehousing and Data Mining

Assignment 1.2

Name: Nabajyoti Borah
Id: cs16msiit035

ID3-A

ID3 algorithm, stands for Iterative Dichotomiser 3, is a classification algorithm that follows a greedy approach of building a decision tree by selecting a best attribute that yields maximum Information Gain (IG) or minimum Entropy (H).

What is Entropy and Information gain?

Entropy is a measure of the amount of uncertainty in the dataset.

Information Gain IG(A) tells us how much uncertainty in S was reduced after splitting set S on attribute A.

Steps in ID3 algorithm?

- Calculate entropy for dataset.
- For each attribute/feature
    - Calculate entropy for all its categorical values.
    - Calculate information gain for the feature.
- Find the feature with maximum information gain.
- Repeat it until we get the desired tree.

What are the characteristics of the ID3 algorithm?

- ID3 uses a greedy approach that's why it does not guarantee an optimal solution; it can get stuck in local optimums.
- ID3 can overfit to the training data (to avoid overfitting, smaller decision trees should be preferred over larger ones).
- This algorithm usually produces small trees, but it does not always produce the smallest possible tree.
- ID3 is harder to use on continuous data (if the values of any given attribute is continuous, then there are many more places to split the data on this attribute, and searching for the best value to split by can be time consuming).

CART

The CART algorithm is structured as a sequence of questions, the answers to which determine what the next question, if any, should be. The result of these questions is a tree like structure where the ends are terminal nodes at which point there are no more questions.

The main elements of CART (and any decision tree algorithm) are:

- Rules for splitting data at a node based on the value of one variable;
- Stopping rules for deciding when a branch is terminal and can be split no more; and
- Finally, a prediction for the target variable in each terminal node.

Features and advantages of CART:

- CART is nonparametric and therefore does not rely on data belonging to a particular type of distribution.
- CART is not significantly impacted by outliers in the input variables.
- You can relax stopping rules to "overgrow" decision trees and then prune back the tree to the optimal size. This approach minimizes the probability that important structure in the data set will be overlooked by stopping too soon.
- CART incorporates both testing with a test data set and cross-validation to assess the goodness of fit more accurately.
- CART can use the same variables more than once in different parts of the tree. This capability can uncover complex interdependencies between sets of variables.
- CART can be used in conjunction with other prediction methods to select the input set of variables.