CAP 6610 – Machine Learning (Spring 2024)

Project 1

3/1/2024

Aaron Goldstein

Credit Card Application Acceptance Predictor Report

Brief Program Description

In short, this program trains various decision tree models using three importance methods and credit card application training data. It then uses these models to predict the acceptance or rejection of credit card applications on test datasets. The importance methods supported are Information Gain, Gain Ratio, and Gini Index.

The dataset used can be found at –

https://archive.ics.uci.edu/ml/datasets/Credit+Approval

Technical Details:

- Loads data two files in a data folder in the project directory
    - test.data
    - training.data
- Cleans and preprocesses the data.
    - Each categorical attribute is read in as such to the initial Pandas dataframe according to the dataset documentation.
    - All attributes are mapped to their type as categorical or continuous in a dictionary which is used in decision tree operations for determining the appropriate way to handle the attribute depending on the importance method.
    - For each attribute in both the test and training set, NaN values are replaced with the attribute median that was calculated from the training set.
        - This is done for both categorical and continuous data
- Splits the training data into ten sequential folds and uses the ten different smaller training sets to train ten different decision tree models for each importance method.
    - The best of these ten different decision three models is selected for each importance method using its calculated F1 score on its validation set. (There were thirty models trained in total, and three were kept, each trained with a different method)

- The best three models were then used to predict labels(acceptance or rejection of an application) on the test dataset and their corresponding f1 scores are calculated and printed to the console.
- All functions such as the recursive decision tree algorithm learning algorithm, the tree prediction and traversal algorithm, the importance method calculations (information gain, gain ratio, gini index), the f1 score calculations, etc are all calculated from scratch.
    - Note: Pandas is used extensively to store data in a way that can be effectively and efficiently manipulated.
- Two key random operations are used in the program.
    - During tree traversal (example prediction), when an example's categorical attribute value is not recognized when attempting to move to a subtree, the prediction algorithm attempts to use the modes for that attribute of the training data in the order they were calculated by pandas. If none of the modes can be used to traverse to a subtree, the subtree is randomly selected from the available ones.
    - In the decision tree learning algorithm, when attempting to get the plurality target either because examples was empty or the number of attributes left to split on is 0, if there are two plurality target values, the chosen label for that leaf node is randomly selected to avoid bias.
    - The Python random module is used as a tool to help make these operations possible. The seed was arbitrarily chosen to be 42 so results can easily be recreated.
- Results Interpretation: Based on the results of the image shown below we can interpret the following:

```
Decision Tree - Best Information Gain Method - ID3 Algorithm Result
F1 Score: 0.784
Decision Tree - Best Gain Ratio Method - C4.5 Algorithm Result
F1 Score: 0.8253968253968254
Decision Tree - Best Gini Index Method - Cart Algorithm Result
F1 Score: 0.8412698412698413
```

    - The Cart Algorithm, which is a decision tree strategy that utilizes the Gini Index importance method in determining the next attribute to split on has the highest F1 score, the harmonic mean of the model's precision and recall.
    - Precision simply means, out of the number of labels we predicted positive, what was the portion that was actually positive.
    - Recall on the other hand, means that out of the total number of positive labels, what was the portion that we were able to predict as positive.

- The F1 score is the harmonic mean of these two measures, which means it gives greater weight to low values, in other words, it will only be high if both precision and recall are high.
- Therefore, the higher F1 score of the Cart Algorithm says it is doing the best overall across these two metrics.
- The order of the algorithms in terms of result quality (best to worst) is as follows:
    - Cart Algorithm (Gini Index)
    - C4.5 (Gain Ratio)
    - Information Gain (ID3)
- In conclusion, for the given data, we should use the Cart algorithm for the best possible predictions in determining whether a credit card will be accepted based on given X features.