HW2 individual assignment

Q) What is causing most of the errors in the Decision Stump?

A) Since decision stump is a single level, we are not asking many questions to classify the data. Thus errors occur in decision stump.

Q) What is causing most of the errors in the Unpruned Decision Tree?

A) In the unpruned decision trees, too much questions are being asked, thus the errors.

Q) What happened when you allowed the tree to get pruned?

A) After pruning, the accuracy increased. The depth, max-features and min-leaves were assigned.

Churning data

Q) Using a decision tree implementation and cross-validation produce a classification model for this data. What is noticeable in the data? Is this model good?

A) The data is imbalanced. The model is bad on the imbalanced data.

Q) Using undersampling, oversampling, and Smote (separately) balance the data and reapply the decision tree software (pruned) to the data. What do you notice? Report your findings

A) In this particular data oversampling got bad results, undersampling was better than oversampling. But SMOTE was way better than both samplings. It balanced the data better.

Pixel data

Q) Assess the “usefulness” of the features using the attribute evaluator Information Gain. How does this “attribute evaluator” work? Remove all the attributes that have 0 information gain. Is it safe to perform this operation? Save the reduced dataset.

A) “attribute evaluator” here reduces the impurity in the data. It is safe to perform this operation as it removes data that gives 0 information gain.

Q) For each of the datasets (the one we provide, the ones you obtained in Steps a. and b.) use SVM to build a classifier. Retain 20% of the data for test purposes (not to be used in the building of the classifier). Use cross-validation on the rest of the set to find the best parameters for SVM (grid search). Report your results.

A) Using grid search with the kernel =”linear” on

1) the given data I found the f1-score to be high

2) removing 0 information gain, the score was the same, since the 0 information gain was of no use.

3)After applying PCA and retaining 90% variance, the f1-score reduced as we had less data.