CS 6220 Data Mining — Assignment 9

Model Evaluation

For this assignment, you may use a dataset of your choice with binary classification. You can also manipulate your dataset forcing it to be binary. To do that you can select half of your classes and simply convert them to 0, and do the same for the other half, converting them to 1.

In this assignment, you will be using the Decision Tree classifier in two different settings, and compare the evaluation metrics for them.

Objectives:

Deliver a notebook containing a detailed evaluation report listing the metrics listed below.

1. The accuracy of your model on the test data
2. The precision and recall values
3. A classification report (scikit-learn has a function that can create this for you)
4. The confusion matrix for this experiment
5. An ROC curve
6. A Precision/Recall curve

Submission:

Through the assignment submission portal on Canvas, submit your ipynb with a pdf of your assignment solution; no need to zip the files. This is a norm for almost all assignments.

Grading Criteria:
Follow the instructions in the pdf, and complete each task. You will be graded on the application of the modules' topics, the completeness of your answers to the questions in the assignment notebook, and the clarity of your writing and code.

Assignment Description

1. Split the dataset into training set and test set (80, 20).
2. Using scikit-learn's DecisionTreeClassifier, train a supervised learning model that can be used to generate predictions for your data.
3. Similarly as in previous step, train another Decision Tree Classifier - but in this case set the maximum depth of the tree to 1 ($max\_depth = 1$). Use the same training and test set as you used for the Decision Tree in the previous step.
4. Report on the six evaluation metrics listed in objective for both the models, and compare their results.