# CSCI 6333 Data Mining & Warehousing

**Module 2: Classification**

**Homework Assignment One**

# Problem: This assignment has two parts:

* Part 1: I ask you to write a program to build a decision tree using the entropy impurity measurement to guide tree generation. The data set is the poker hand data set archived at UCI Machine Learning Repository:

|  |  |
| --- | --- |
| **Poker Hand Data Set**  *Download*: [Data Folder](https://archive.ics.uci.edu/ml/machine-learning-databases/poker/), [Data Set Description](https://archive.ics.uci.edu/ml/machine-learning-databases/poker/poker-hand.names)  **Abstract**: Purpose is to predict poker hands | https://archive.ics.uci.edu/ml/assets/MLimages/Large158.jpg |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 1025010 | **Area:** | Game |
| **Attribute Characteristics:** | Categorical, Integer | **Number of Attributes:** | 11 | **Date Donated** | 2007-01-01 |
| **Associated Tasks:** | Classification | **Missing Values?** | No | **Number of Web Hits:** | 212827 |

You shall use the training data set to build your decision tree and then use the testing data set to evaluate your decision tree. You need to report classification accuracy using a bar chart and compare it with AdaBoost classification which is given in Part II.

**Remark:** For the Porker Hand Data Set, the 11th column is the “Poker Hand” Classification attribute with 0 to 9 values. This classification is not binary. For the sake of our homework assignment, we consider two classes: 0, which means “nothing in hand”; and nonzero (1 to 9), which means “something in hand”. By grouping 1 to 9 into one class, we will work on a binary classification problem, which is easier to work with. When you test your classifier, you shall consider grouping the original classes 1 to 9 into one class “something in hand.” You shall also consider the binary classification for Part II as well.

* Part II: For this part, I ask you to use the same training data set in Part I (ref. to the Remark above) to build a classifier using AdaBoost. Here, I ask you to use a perceptron to build a weak classifier at each iteration of AdaBoost. Run 15 iterations. Then, you need to apply the final classifier, which is a boosted combination of these 15 perceptron classifiers to the testing data set to evaluate your final classifier. You shall record the classification accuracy and compare it in a bar chart with that of the decision tree model built in Part I.

**Programming Language:**

C++, or Java, or Python

# Alternative Approaches to Homework Assignment One

The background of students in this class is diverse. If your programming skill is not good enough at this point, I suggest you choose one of the following two alternative approaches:

* You can team with one student who is good at programming to complete this assignment.
* You can research for existing data mining tools/software packages and use these to solve Part I and Part II. In this case, you shall cite clearly what tools/software packages are used in your solutions.