BASIC ASSIGNMENT 2
HUDK4050: CORE METHODS IN EDUCATIONAL DATA MINING

PROFESSOR RYAN BAKER

DIAGNOSTIC METRICS
DUE NOON, WEDNESDAY SEPTEMBER 24

In this assignment, you need to apply a set of metrics to two data sets, classifier-data-asgn2.csv and regressor-data-asgn2.csv

These are not real data sets, but are simulated data, generated for the purpose of this assignment.

You can use any tool you want to complete this assignment. This includes Microsoft Excel, and statistical or data mining package you choose, and other tools available on the internet.

Question 1:

Using regressor-data-asgn2.csv, what is the Pearson correlation between data and predicted (model)? (Hint: this is easy to compute in Excel)

Question 2:

Using regressor-data-asgn2.csv, what is the RMSE between data and predicted (model)? (Hint: this is easy to compute in Excel)

Question 3: Using regressor-data-asgn2.csv, what is the MAD between data and predicted (model)? (Hint: this is easy to compute in Excel)

Question 4: Using classifier-data-asgn2.csv, what is the accuracy of the predicted (model)? Assume a threshold of 0.5. (Hint: this is easy to compute in Excel)

Question 5: Using classifier-data-asgn2.csv, how well would a detector perform, if it always picked the majority (most common) class? (Hint: this is easy to compute in Excel)

Question 6: Is this detector's performance better than chance, according to the accuracy and the frequency of the most common class?

Question 7: What is this detector’s value for Cohen’s Kappa? Assume a threshold of 0.5.

Question 8: What is this detector’s precision, assuming we are trying to predict “Y” and assuming a threshold of 0.5?

Question 9: What is this detector’s recall, assuming we are trying to predict “Y” and assuming a threshold of 0.5?

Question 10: Based on the precision and recall, should this detector be used for strong interventions that have a high cost if mis-applied, or fail-soft interventions with low benefit and a low cost if mis-applied?

A) STRONG

B) FAIL-SOFT

C) EITHER

D) NEITHER

Bonus Question 11: What is this detector’s value for A’? (Hint: There are some data points with the exact same detector confidence, so it is probably preferable to use a tool that computes A’, such as http://www.columbia.edu/~rsb2162/computeAPrime.zip -- rather than a tool that computes the area under the ROC curve).