Rhodes College

**Topics in Computer Science: Data Mining** **Assignment 4**

COMP 345 – CRN 19746/19839 Fall 2018

---

**Classification** **Due Mon. Oct. 1st/Tues. Oct. 2nd**
**(at the beginning of class)**

1. Read Chapter 8 in Han Book

2. We want to predict the outcome of the next tennis match between the two top-ranked female tennis players: Serena Williams and Dinara Safina. The outcomes for each match are labeled as S for Serena or D for Dinara. Other information we know about each match is the time of day it was played, Match Type, and Court Surface.

   Using the following training set, determine which attribute would be selected for the first split using the ID3 decision tree induction algorithm. Show your work, and draw the resulting tree after the first split.

| Time | Match Type | Court Surface | Outcome |
|---|---|---|---|
| Morning | Master | Grass | S |
| Afternoon | Grand Slam | Clay | S |
| Night | Friendly | Hard | S |
| Afternoon | Friendly | Mixed | D |
| Afternoon | Master | Clay | D |
| Afternoon | Grand Slam | Grass | S |
| Afternoon | Grand Slam | Hard | S |
| Afternoon | Grand Slam | Hard | S |
| Morning | Master | Grass | S |
| Afternoon | Grand Slam | Clay | D |
| Night | Friendly | Hard | S |
| Night | Master | Mixed | D |
| Afternoon | Master | Clay | D |
| Afternoon | Master | Grass | S |
| Afternoon | Grand Slam | Hard | S |
| Afternoon | Grand Slam | Clay | S |

3. Imagine that you are given the following set of training instances. Each feature can take on one of three nominal values: a, b, or c.

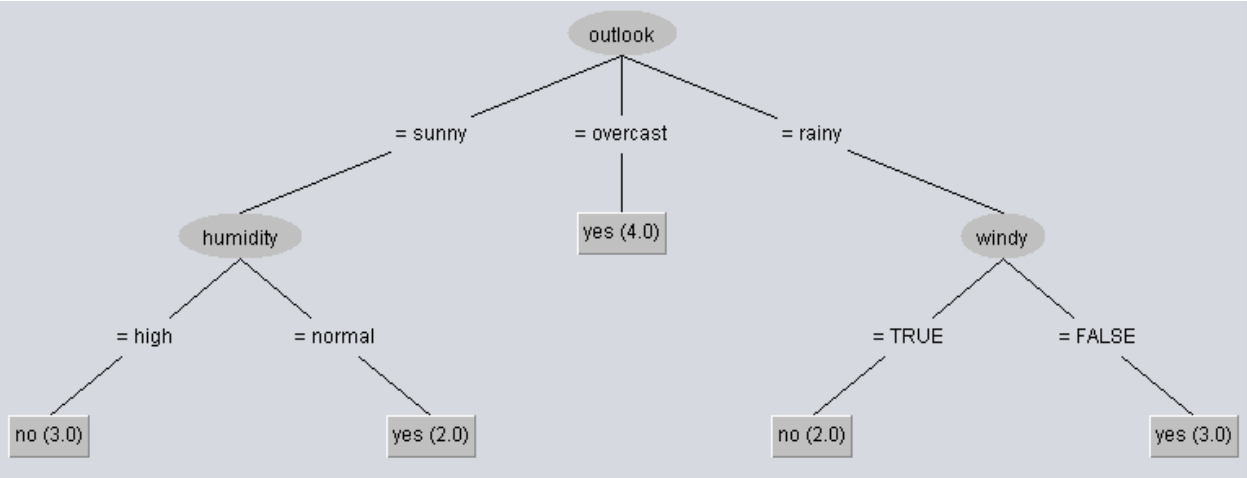
| F1 | F2 | F3 | Category (Class) |
|---|---|---|---|
| a | c | a | P |
| c | a | c | P |
| b | c | a | N |
| c | c | b | N |
| a | a | c | N |

How would a Naïve Bayes system classify the following test example? Be sure to show your work.

F1 = a, F2 = c, F3 = b

4. Suppose that there are a total of 50 data mining related documents in a library of 200 documents. Suppose that a search engine retrieves 10 documents after a user enters 'data mining' as a query, of which 5 are data mining related documents. What are the precision and recall?

5. The following decision tree is the tree produced by running the J48 algorithm in Weka on the weather.nominal.arff file.




Use the decision tree to predict the class labels for the following test set of data and fill the correct value in for the Predicted Class value for each instance.

| Outlook | Temperature | Humidity | Windy | Play (Actual) | Predicted Class |
|---|---|---|---|---|---|
| sunny | hot | high | FALSE | no | |
| overcast | hot | high | TRUE | yes | |
| rainy | hot | normal | FALSE | no | |
| overcast | mild | high | FALSE | yes | |
| rainy | cool | normal | TRUE | yes | |
| rainy | cool | high | TRUE | no | |
| overcast | mild | normal | FALSE | yes | |
| sunny | hot | high | FALSE | no | |
| sunny | cool | normal | FALSE | yes | |
| rainy | mild | normal | FALSE | no | |
| sunny | mild | normal | TRUE | yes | |
| overcast | mild | normal | FALSE | no | |
| overcast | hot | high | FALSE | yes | |
| rainy | mild | high | TRUE | no | |

Using the predicted and actual class values above, fill in the correct values for the following confusion matrix.

| Actual Class \ Predicted Class | Play = yes | Play = no | Total |
|---|---|---|---|
| **Play = Yes** | | | |
| **Play = No** | | | |
| **Total** | | | |

Once you've filled in the values for the confusion matrix above, use those values to evaluate the performance of the decision tree model.

a. Accuracy

b. Specificity

c. Sensitivity

d. Precision

e. F-measure