**CS497 Data Mining Spring 2011**

**ASSIGNMENT#2**

***(45 points in total)***

**Due*: on Monday, March 7, at the beginning of class, hard copy submission***

**Problem 1** Given the following weather dataset:

**Outlook Temp Humidity Windy Play**

Overcast Hot Normal False Yes

Overcast Mild High True Yes

Sunny Mild Normal True Yes

Rainy Mild Normal False Yes

Sunny Cool Normal False Yes

Sunny Mild High False No

Overcast Cool Normal True Yes

Rainy Cool Normal True No

Rainy Cool Normal False Yes

Rainy Mild High False Yes

Overcast Hot High False Yes

Sunny Hot High True No

Sunny Hot High False No

Rainy Cool High True No

Sunny Hot Normal False Yes

1. **[10 points]** Use 1R method to construct classification rules for the predicted attribute **play. Show all your work!**
2. **[15 points]** Construct Naïve Bayes classifier. **( Compute all needed probabilities )**

Use this classifier to classify the new instances:

Overcast Cool Normal False ?

? Cool Normal False ?

Overcast ? ? False ?

**Show all your work!**

**Problem 2 [20 points] – Programming - Using WEKA.**

Go through all the steps bellow and answer all the questions.

1. Take the file [genes-leukemia.csv](http://www.kdnuggets.com/data_mining_course/data/genes-leukemia.csv) (genes-leukemia-description.txt - [description of the data](http://www.kdnuggets.com/data_mining_course/data/genes-leukemia-description.txt)) and convert it to Weka file genes-a.arff.   
   You can convert the file either using a text editor like emacs (brute force way) or find a Weka command that converts .csv file to .arff (a smart way).
2. Target field is CLASS. Use J48 on genes-leukemia with "Use training set" option.
3. Use genes-leukemia.arff to create two subsets:   
   genes-leukemia-train.arff, with the first 38 samples (s1 ... s38) of the data   
   genes-leukemia-test.arff, with the remaining 34 samples (s39 ... s72).
4. Train J48 on genes-leukemia-train.arff and specify "Use training set" as the test option.   
   What decision tree do you get? What is its accuracy?
5. Now specify genes-leukemia-test.arff as the test set.   
   What decision tree do you get and how does its accuracy compare to one in the previous question?
6. Now remove the field "Source" from the classifier (unclick checkmark next to Source, and click on Apply Filter in the top menu)   
   and repeat steps **4.** and **5**.   
   What do you observe? Does the accuracy on test set improve and if so, why do you think it does?
7. Which classifier gives the highest accuracy on the test set?

**Submit your Problem 1. and Problem 2 to me as a hard copy on Monday, March 7, at the beginning of class.**