**CS497 Data Mining Spring 2011**

**ASSIGNMENT#1**

***(30 points in total)***

**Due*: on Wednesday, February 24, at the beginning of class, hard copy submission***

**Problem 1 [10 points] – Reading/Writing:**

Use two different search engines and find two different recent stories involving data mining. Describe the role of "data mining" in the story using your own words.

***Example:*** Searching [Google News](http://news.google.com) for [data mining](http://news.google.com/news?hl=en&ned=us&q=data+mining&btnG=Search+News) on August 9, 2004 (*tell what engine was used, when*) gives us [***IBM, Mayo Clinic Take Next Data Mining Step DB2-based system contains records on 4.4M patients***](http://www.computerworld.com/industrytopics/healthcare/story/0,10801,95115,00.html)(*tell what was found*).

This story describes how Mayo clinic uses IBM tools to analyze the data on its patients in order to develop indivudualized care. ... a doctor treating a patient for cancer could use data mining to find the results treatments given to the similar patients and use this information to find the best treatment course for new patient (*tell what the story was about. Use your own words. One-two paragraphs are enough.*)

**Problem 2 [20 points] – Programming - Using the WEKA Workbench - Become familiar with the use of the WEKA workbench to invoke several different machine learning schemes.**   
Use latest stable version of WEKA. Use both the graphical interface (Explorer) and command line interface (CLI). (See [Weka home page](http://www.cs.waikato.ac.nz/~ml/weka/) for Weka documentation.)

1. **Use the following learning schemes**, with the default settings to analyze the weather data (in weather.arff). For test options, first choose "Use training set", then choose "Percentage Split" using default 66% percentage split. Report model percent error rate for each scheme and each size of training set used.
2. **The following learning schemes must be used:**

ZeroR (majority class)

OneR

Naive Bayes Simple

J4.8

1. **Based on the results (error rates) answer the following questions:**
2. Which of these classifiers are you more likely to trust when determining whether to play? Why? Explain your answer!
3. What can you say about accuracy when using training set data and when using a separate percentage to train? Try to explain your answer!

**Submit your Problem 1. and Problem 2 to me as a hard copy on Wednesday, February 24, at the beginning of class.**

**The dataset for Problem 2 may be found at:** [www.kdnuggets.com/data\_mining\_course/data/weather.arff](http://www.kdnuggets.com/data_mining_course/data/weather.arff)