Machine Learning Final Project (Due May 12 11:59PM - Hard deadline)

Total Points on this project is 20. You are allowed to work in teams. Maximum team size is 4. However, the grading is relative. The best project (if for instance, executed by a single person) gets 20, your team size of 4 is expected to do 4 times the work to get a grade of 20. I would expect most project teams to be either 2 or 3. You do not get special consideration for going alone on a project.

The aim of the project is to demonstrate your understanding of machine learning. There are two different types of projects you could submit.

1. In the first type, you will identify a prediction problem of interest. For example, a problem could be to predict if the young generation would buy a new product. In order to predict this, I could imagine collecting data from 50 students who are either at junior year or below. I can collect information such as major, the gender, their interest in computing and the most important feature that they like in the new product. Identify several features (attributes) that are most useful for this prediction task. Then collect the data. Note that to collect data, you could use tools such as google forms or simply collect the data from web (previously played games, previous elections, census information that is open access etc).

2. In the second type, you could use any of the UCI or Kaggle data sets. If you are preferring to use the UCI data sets, I require at least 2 different data sets, if it is Kaggle, then one should suffice.

As with the data set, there are two different approaches you can take with the algorithms.

1. You can use your own implementation of the algorithms. In this case, you are required to use two algorithms. The analysis can be done reasonably well - understanding when the algorithms work, when it fails, what issues are present with your implementation and what you could do potentially better are sufficient. You should submit your code in this case.

2. You can use Weka/Scikit or any other implementation. In this case, you need at least 5 algorithms. The analysis should be deeper. In addition to above questions, I want you to present the confusion matrices, the potential pitfalls, different parameter settings etc. In summary, this must be a comprehensive analysis. You should talk to me about your potential submission before you submit the final project.

You will submit the following as a report in class.

1. The data set description.

2. Results of running the experiments that contain the average of the confusion matrix entries, accuracy, true positive rate, false positive rate and the area under the ROC curve.

3. Brief discussion of the performance of the algorithms. This is very important and contributes to majority of the points in the project. You are allowed to speculate on why a particular algorithm is better than the other or why they exhibit similar results.

4. Screen shots from Weka/Scikit etc.(the last part of the output with the results).

Remember, we require hard copies of the project report in class.
You are required to submit the code online in elearning.