**Machine Learning Homework 1 Due 21 Feb Midnight (PST)**

1. We have 2010 World Cup data including last name, team, position, minutes played, and game statistics for each player (Players.csv) as well as world ranking, games played in tournaments, and game statistics for each team (Teams.csv)

Note: Statistics, including yellowCards and RedCards, are for the entire tournament (excluding final game). Team ranking is the world ranking going into the tournament so may not be 1-32 even though there are only 32 teams. For the joined dataset (PlayersExt.csv), keep in mind that since the tables are joined, country data will show up for each player.

We need to set up machine learning model to predict whether team ends in group or knockout stage based on number of yellow cards per game and number of red cards per game. Try to use at least two models you learned so far and give predictions and give evaluations and explain why use these two models for prediction.

(data files found in classroom) (50’)

2. True of False (5’\*4)

a) SVM is best suitable for small set of data and can achieve better performance with a small amount of training data.

b) Logistic Regression is normally used to solve problems of regressions which is not linear.

c) Meanshift needs to specify how much clusters before training

d) Naïve Beyes is based on very unrealistic assumptions so we really do not use it a lot in real world.

3. Questions (10’\*3)

List all the models we learned so far which

a) has demand in using huge computer memory when data sets grows in huge amount.

b) Need heuristic algorithm to solve the parameters of the model during training

c) Only limits to the convex optimal problem ( means data set should be convex when classified)