Assignment 1

Introduction to Artificial Intelligence

DUE: Oct. 25. *Submit in Class*

(Data Analysis & Machine Learning)

First, you will try a titanic survival prediction script. The script is provided. Refer to Task-1 below.

Next, you are asked to write a small data analysis program for predicting next lottery number (6 digits including the bonus number; a single digit ranges from 1 to 45). The previous winning numbers are provided in 'lottery.csv' file, containing all 761 weekly rounds from 2002/12/07 till 2016/08/20. You will use this file to complete the following analysis tasks.

lottery.csv data format:

round, date, first, second, third, fourth, fifth, sixth, bonus

**Task-1: Download 'titanic-assign1.R', 'train.csv', and 'test.csv'. Install R (or R studio) and the required packages to run the given script (study it!). Please create the graphs in the script and print them out (black and white printing is just fine) for submission.

For example, I use homebrew to install R or you can download R.

$> brew install r
$> sudo r
$> install.packages("ggplot2")
$> ….

**Task-2: Write a statistical analysis script to display the most frequently appeared number to the least. Use pandas (http://pandas.pydata.org/) for this task. Please print out your script for submission.

Example: $> ./your_script.py lottery.csv
Sample output: 1 -> 134 times
2 -> 117 times
3 -> 112 times …

**Task 3: Create a modified lottery data format by adding at new columns. For example, you can add "win" column indicating '0'-lose and '1'-win. Then your lottery data set would look like the following:

round, date, first, second, third, fourth, fifth, sixth, bonus, win

Likewise, please add at least three columns (including 'win') that you like, such as 'weather', 'daysofweek', and whatever. What you want to add and how many choices for a single column are completely up to you. Please print the first 20 lines of your modified data set (including csv header) for submission. If you have a source code for doing this, please print them out for submission.

**Task 4: K-mean clustering – Use any combination of features in your lottery.csv to group all weekly rounds. For example, you can use the vectors of 'first' and 'second' to create 3 clusters (K=3) and provide a clustering graph like the below scikit-learn sample source code. Another example is to use the average of all 6 digits to create 4 clusters (K=4). Write a K-mean analysis

script that creates a cluster figure (just one graph). Within your source code (comment line), please explain which features and how many clusters you use.

Overview - http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

Sample source code - http://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_iris.html

SUBMISSION:

Print out source code and figures from your Task-1, 2, 3, and 4.

We will discuss more in class.