Assignment 1

ITE 351: AI & Applications

Deadline: LMS for online submission and Hard copy in class

You are asked to write a small data analysis program for weekly lottery numbers (6 digits + 1 bonus digit; a single digit ranges from 1 to 45). The first place winner should get all the 6 digits correct, the bonus number is excluded. All the winning numbers are provided in 'lottery-2022.09.csv' file, containing around 1000 weekly rounds from 2002/12 till 2022/09. You will use this file to complete the following analysis tasks.

lottery.csv data format:

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

*Task-1: Write a statistical analysis script to display the most frequently appeared number to the least for the win numbers only. Use pandas (http://pandas.pydata.org/) for this task. Please print out your script for submission.

```
Example:        $> python ./your-program.py lottery-2022.09.csv
Sample output:  43 -> 180 times
                34 -> 178 times
                1 -> 175 times
                …
```

*Task-2: Create a modified lottery dataset by adding a new column (the 'win' column). This new column indicates '0' for lose and '1' for win. In order to do this, you need to create fake synthetic lose number (no win) to every round. These fake numbers should not be redundant throughout the entire dataset. You may use a random function to create, however, it does not guarantee the uniqueness of the fake numbers. The new dataset should be double in size. It look somewhat similar to:

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

945,2021.01.09,13,1765554491,9,10,15,30,33,37,26,1
945,2021.01.09,0,0,12,10,11,16,31,32,36,27,0
```

Please print the first and last 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-3: Repeat Task-1 with the fake numbers that you have created. Show the result.

*Task-4: K-mean clustering – Use any combination of features in your new dataset to group all weekly rounds. For example, you can use the values of 'first' and 'second' to create 2 clusters (K=2) 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 N-clusters. If you come up with your own and reasonable feature, that will be nice.

DO NOT use the example above:
- Do not use 'first' and 'second' as your feature.
- Do not use the average of all 6 digits as your feature.

Print out your K-mean analysis source code that creates a cluster figure (just one graph is fine). Within your code (comment line), please explain which features and how many clusters you use.

Overview - https://scikit-learn.org/stable/modules/clustering.html#k-means
Sample source code - http://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_iris.html

Python and scikit-learn are recommended. However, you can use any means possible (or comfortable for you. Do not use Neural Networks yet.) to complete the task. If you have a source code for doing this, please print them out for submission. R is also fine.

*Task-5: Write one paragraph explaining your tasks and any difficulties you had. (a couple of sentences should be enough.) Even if you can't do the whole assignment, submit as much as you can (with explanation why you can't do this).

SUBMISSION:

Put the outputs of Task-1 to 5 to a single PDF file: including source code (if any), steps, graphs, and your paragraph. Using Jupyter Notebook-IPython and converting it to PDF would be nicer as well.

DO NOT WORRY. We will discuss more in class.