Midterm Take Home Exam

AI+X: Deep Learning

제출 마감일: Nov. 22 *Submit LMS online*

You are asked to write a small data analysis program for weekly lottery numbers (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 around 900+ weekly rounds from 2002/12/07 till 2021/11/13. You will use this file (lottery.csv) to complete the following analysis tasks.

lottery.csv data format:

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

**Task-1: 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. An alternative way to do this is using Excel, but I strongly suggest that you try python and pandas for fun. If you are using Excel, show the step by step making your file.

Example: $> python ./program.py lottery.csv
Sample output: 1 -> 100 times
2 -> 99 times
3 -> 98 times …

You may find that something like this:
https://www.dhlottery.co.kr/gameResult.do?method=statByNumber

**Task-2: Create a modified lottery data format by adding a new column. Let's make a simple assumption that winning here means get all six numbers correctly (excluding the bonus number). For example, you can add "win" column indicating '0'-lose and '1'-win. Please add not-winning (fake) numbers to each round to your modified lottery dataset. You will be adding '0' for lose for every round: '1' for each round. So your new dataset size should be double. Your new lottery data set would look like the following:

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

663,2015.08.15,3,5,8,19,38,42,20,1
663,2015.08.15,1,2,3,4,5,6,7,0

Please print the first 10 and last 10 lines of your modified data set for submission. If you have a source code for doing this, please include them for submission.

**Task-3: Feature engineering: Create a new feature and add it to the column list (to the dataset from Task-2). For example, you can compute the average value of all 7 numbers (including the bonus) and maybe use it as an extra feature to consider. Do not use this average example as your feature. Come up with your own and explain your feature. If you have a source code to compute your feature, please include them for submission.

**Task-4: Explain your plan how you use the data file from Task 2 or 3 to create the smart lottery prediction agent. You can explain how you create training and test dataset. And show how to use ML (using random-forest exercise from our class or like PCA, or K-mean clustering, or anything that you like) to create the learning agent. If you can provide a code and running example (by attaching the screenshots), that will be the best. If you think that other decision making (statistic-based) algorithms are suitable, then do so and explain how it worked. Be clear and show code if

you can. Accuracy does not matter. Take a good look at the samples that we have seen in the class, perhaps use them wisely.

Python and scikit-learn are recommended. However, you can use any means possible (or comfortable for you, like Excel or R or google spreadsheet) to complete the task. If you have a source code for doing this, please print them out for submission.

ML regression in python - https://plotly.com/python/ml-regression
ML regression in python - https://scikit-learn.org/stable/
R - https://cran.r-project.org/web/views/MachineLearning.html
Weka - https://www.cs.waikato.ac.nz/ml/weka/

**Task-5: Write one paragraph explaining your tasks and any difficulties you had. (several 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:

Do the above tasks. Put the outputs of all tasks to a single PDF file: including source code (if any), steps, graphs, and your paragraph. Perhaps, using Jupyter Notebook may be helpful to export to a single PDF.

DO NOT WORRY. We will discuss more in class.

Suggestions:

Take a very similar step like the examples we saw in class. You will show me how to investigate the dataset. You can choose one of the methods from the AI/Machine learning/DL above. Pandas should be really helpful as well. You are showing a step-by-step procedure on how to do learning. Accuracy of your method DOES NOT matter in this exam.

It's completely up to you for your choice of tools.

Using Excel to complete the exam is also an option.