Assignment #6
MACS 30000, Dr. Evans
Due Monday, Nov. 19 at 11:30am
Haowen Shang

1. Netflix Prize and Bell, Koren, and Volinsky (2010)

The Netflix Prize is an open call contest on quality of predictions. Submissions to this contest would be judged by how much improvement in root mean squared error (RMSE) a team achieves over Netflix's internal algorithm, Cinematch (Bell et al., 2010, p.24). The RMSE is the root mean squared error of the movie rating prediction model. The hold-out dataset contains around 3 million ratings on 18,000 movies (Bell et al., 2010, p.24). The team which achieves the greatest improvement in RMSE would be the winner. The formula to calculate RMSE is as follows:

$$RMSE = \sqrt{E((predicted\ rating - actual\ rating)^2)}$$

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(predicted\ rating_i - actual\ rating_i)^2}$$

Only the team who improved Netflix website's movie recommendation system by 10% or more would be considered a candidate to win the prize (Steve Lohr, 2010, p.25).

(b)

At the beginning of the Netflix Prize contest, the most commonly used method for predicting ratings (stars) on movies was "nearest neighbors" (Bell et al., 2010, p.25). With this method, the predicting ratings on a movie by a person will be "a weighted average predicting rating" of similar movies by this person (Bell et al., 2010, p.25).

(c)

The author used the "ensembles of heterogeneous methods" to blend multiple models into prediction sets (Bell et al., 2010, p.29). If a model is not highly correlated with other models, blending this model with the other models would improve the overall prediction (Bell et al., 2010, p.28).

Reference:

Robert M. Bell, Yehuda Koren and Chris Volinsky, "All Together Now: A Perspective on the Netflix Prize", Chance, 2010, 23 (1), 24–29.

Steve Lohr, "The Contest That Shaped Careers and Inspired Research Papers", Chance, 2010, 23 (1), 25-26,

2. Collaborative problem solving: Project Euler

(a) My Project Euler user name is 'haowen'.

My friend key is '1408032_5lIJ7aUXGNSKLf9KG8LfX4yrWYUlrbx9'.

(b) The problem I choose to answer is Problem 1:

"If we list all the natural numbers below 10 that are multiples of 3 or 5, we get 3, 5, 6 and 9. The sum of these multiples is 23.

Find the sum of all the multiples of 3 or 5 below 1000."

The code is:

```
In [1]: lst = []
        for i in range(1000):
            if i % 3 == 0 or i % 5 == 0:
                lst.append(i)
        sum(lst)

Out[1]: 233168
```

The answer is: 233168 .

(c) I would most aspire to achieving "High Flyer", "One Percenter", and "Perfection" awards.

When I process to the maximum level, I'll achieve "High Flyer" award. I like this award because it encourages me to solve more complicated and challenging problems.

When I solve 111 problems, I'll achieve "One Percenter" award, which means I am better than 99% problem solver. I like this award because it represents that I achieve a higher-level in coding and make much more efforts than most people.

When I solve 643 problems, I'll achieve "Perfection" award, which means I have solved every problem. I want to achieve this award because I think this is the most challenging award. If I can achieve it, I'll improve my coding, modeling, and problem solving ability a lot.

3. Human computation projects on Amazon Mechanical Turk

I select the task named "Find phone numbers for property owners", which is created by Chad Collishaw. This task asks participants to find the phone numbers for the property owners when giving the property address, owners' names, and owners' addresses.

(a) The full payment structure is $1.00 if participants find the phone numbers for the property owners.

(b) There are two qualifications. The first one is participant's HIT approval rate (%) should be greater than 75. The second one is the participant should be granted the "Master" status. Participants can request qualification when they are not granted the "Master" status.

(c) The allotted time for this task is 20 minutes. There are at most three phone numbers of the property owners I need to find. I think I can do it all in one hour. Indeed, I need to do it all in 20 minutes. Therefore, the implied hourly rate is $3.00 per hour.

(d) This job expires on 11/22/2018.

(e) This project would cost the HIT creator $1,000,000 at most if 1 million people participated in the task.

Reference:

Amazon Mechanical Turk website:

https://worker.mturk.com/?filters%5Bsearch_term%5D=Find+phone+numbers+for+property+owners

4. Kaggle open calls

(a) I registered for the Kaggle account and my username is 'haowens'.

My Kaggle website is: https://www.kaggle.com/haowens

(b) I'm interested in the competition titled "Two Sigma: Using News to Predict Stock Movements". In this competition, participants use the data from news and create models to predict stock price.

A technology company named "Two Sigma" sponsored this competition. This company aimed at using scientific method in investment management. They use machine learning, distributed computing and other technologies to combine data with financial expertise and create sophisticated trading models. By applying technology and data science, they produced breakthroughs in financial forecasts, risk management and related fields.

The submissions of this competition will be evaluated by a score based on the quality of prediction on the 10-day return for a stock.

$$score = \frac{\bar{x_t}}{\sigma(x_t)}$$

$$x_t = \sum_i \hat{y}_{ti} r_{ti} u_{ti}$$

$\hat{y}_{ti}$ is a confidence value between -1 to 1. Based on participant' expectation of a stock return, he would assign the value of $\hat{y}_{ti}$. A large and positive value which is near 1 will be assigned to $\hat{y}_{ti}$ when he expects a stock to have a large positive return over the next ten days. And a large negative value will be assigned to $\hat{y}_{ti}$ when he expects the stock to have a negative return, which is near -1. r_{ti} is the 10-day market-adjusted leading return for the stock the participant predicted and u_{ti} is a 0/1 universe variable.

The final score represents the precision of participant's prediction, which is calculated by the mean divided by the standard deviation of participant's daily x_t value.

Participant who gets the score in the first place will earn a prize of $25,000. Participant who gets the score in the second place will earn a prize of $20,000. Participant who gets the score in the third place will earn a prize of $15,000. And participant who gets the score in the 4th through 7th place will earn a prize of $10,000.

Honor code issues are of importance in this competition. Using information outside the provided dataset and references and abusing the competition infrastructure to gain an edge in accuracy are not allowed. The sponsor reserves the right to determine and disqualify any contributions that he believes demonstrating cheating.

There are two stages for the competition. The first stage is submission period from 9/25/2018 to 1/8/2019. Participant should accept rules and merge team no later than 1/2/2019. The second stage is scoring period ending at 7/15/2019.

At the end of stage one, participants select two best submissions to be rescored. Submission files will be evaluated by the entire time span of stage one and stage two. Specifically, participants use the Kernels environment to analyze market and news data and build models for the historical, stage 1 time period and the future, stage 2 time period at the same time. In stage one, the leaderboard will show scores on historical data while at the end of stage one, participants are not allowed to change their code and then leaderboard will transit to show scores on future data. The Kernels limits of constraints for this competition are 16 GB Memory, 6 Hours total runtime and 4 CPU cores.

(c)

The sponsoring entity, "Two Sigma", is a technology company aimed at using data and model to improve investment management, insurance and related fields, so I think they may use the results of this competition to understand the predictive power of the news. They may use the winning submission answer as references to find how to use news data to predict financial outcomes and improve investment strategies.

Reference:

https://www.kaggle.com/c/two-sigma-financial-news