LI LIU

Assignment 6

MACS 30000

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

The submissions to the Netflix Prize open call contest were judged by the improvements in root mean squared error (RMSE), which is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

where y hat is the predicted rating and y is the actual rating in the test dataset (Bell et al., 2010).

The RMSE results would be compared with the baseline result by Cinematch, which is Netflix's algorithm for movies recommendation system. All submissions would be automaticaly judged by the system, but only the results with significant reduction of RMSE were likely to win the prize.

The nearest neighbors method was common at the beginning of the contest. This method predicts the individuals' ratings by reweighing their ratings for similar items (Bell et al., 2010).

Bell et al. (2010) describe that their submission was "a linear combination of 107 prediction sets, with weights determined by ridge regression". A model would improve the overall prediction of a blend when its correlation with other components is low (Bell et al., 2010).

Reference

Robert, B., Yehuda, K, & Chris, V. (2010) All Together Now: A Perspective on the Netflix Prize, *CHANCE, 23:1,* 24-29, DOI: 10.1080/09332480.2010.10739787

2. Collaborative problem solving: Project Euler

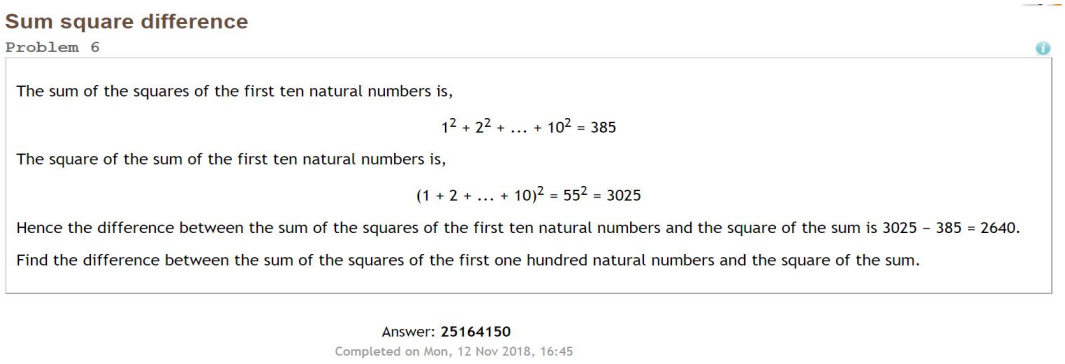
2.(a)

Username: liu431

Friend key: 1407332_NkO8kYQkJokCQm5LOKDr70049xnpnAxx

2. (b)

Problem 6:



Sum square difference

Problem 6

The sum of the squares of the first ten natural numbers is,

$$1^2 + 2^2 + \ldots + 10^2 = 385$$

The square of the sum of the first ten natural numbers is,

$$(1 + 2 + \ldots + 10)^2 = 55^2 = 3025$$

Hence the difference between the sum of the squares of the first ten natural numbers and the square of the sum is 3025 − 385 = 2640.

Find the difference between the sum of the squares of the first one hundred natural numbers and the square of the sum.

Answer: **25164150**

Completed on Mon, 12 Nov 2018, 16:45

Python Code:

```
def ssd(n):
    sum1,sum2=0,0
    for i in range(1,n+1):
        sum1+=i**2
        sum2+=i
    return(sum2)**2-sum1

print(ssd(100))
```

Answer: 25164150

2.c

Awards: Perfection (solve every problem); Gold Medal (the first to solve a problem); Master of Archives (solve every problem in the archives)

These awards motivate me to participate more in the platform and get intellectual rewards for solving hard problems. Having these rewards show credentials of my mathematics and coding skills.

3. Human computation projects on Amazon Mechanical Turk

3.a

Project: “Write the text shown in an image”

Description: The participants will be shown the image and asked to copy the “ID Number” and “Date of Birth” fields.

3.b

The reward is $0.01 for completing the task.

3.c

It is eligible for U.S. citizens or permanent residents who have master degrees.

3.d

Time allotted is 10 minutes. I can do at least 100 tasks in one hour. The implied hourly rate is $1.

3.e

The job expires on Nov 19th, 2018.

3.f

The project would cost at most $10,000 if 1 million people participate in the task.

4. Kaggle open calls

4.a

Profile page: https://www.kaggle.com/liu431

4.b

Website of the competition: https://www.kaggle.com/c/ga-customer-revenue-prediction

The title of the competition is “Google Analytics Customer Revenue Prediction”. The sponsor is RStudio, which is an open-source integrated development environment for R language. The company also partners with Google Cloud and Kaggle to host this competition.

The goal of this competition is to predict revenue per customer at the Google Merchandise Store. The submissions will be evaluated using root mean squared error. For winning submissions in the private leaderboard, the rewards are $12,000, $8,000 and $5,000 for the top three teams respectively. Moreover, they provide extra prizes for teams that use R as the modeling language. The total prizes amount is $45,000. There are two important honor code issues. One is participants shouldn’t submit from multiple accounts. The other is they are prohibited to share the code and data to other people or team. The participants need to register the competition and form the team before November 23nd, 2018. Then they should submit the final results in the following week. The test set to evaluate these submissions would be the actual transaction data in December 2018 and January 2019. Each team could submit at most 5 entries per day and 2 final entries before the deadline.

4.c

The Google Merchandise Store could use the machine learning models of the best submissions to predict future sales. This would help with the company’s marketing

strategy and supply chain operations. Google could also identify talented data scientists from the participants to join the company. RStudio could demonstrate its computing power and grow populaity in the data science community, as many are switching to Python nowadays.