Assignment 6

Fulin Guo

1. (a) The submissions would be judged based on the root-mean-squared error (Bell et al. 2010), which was to measure the differences between the predicted ratings and the real ratings. The criterion function was:

   RMSE=$\sqrt{\frac{\sum_{i=1}^{N}(y_i^p - y_i)^2}{N}}$

   where N was the number of test sets, y_i^p was the predicted value (rating) of the item *i* in the test set, and y_i was the corresponding real value.

   There were no cutoffs in this judging method. The smaller the RMSE, the better the model would be. Therefore, if the fit was very poor, the RMSE would be very large.

   (b) The most commonly method used at the beginning of the contest was "nearest neighbors" (Bell et al. 2010, p 25). That is, the predicted stars of a movie rated by a person was calculated by the weighted average of ratings of movies (by the same person) that were neighbors of the predicted movie (Bell et al. 2010). The formula was:

   $predicted\_s_{mi} = \frac{\sum_{o \in N(u,m)} w_{mo} s_{ou}}{\sum_{o \in N(u,m)} w_{mo}}$ (Bell et al. 2010)

   where $predicted\_s_{mi}$ is the predicted rating of the movie *m* by user i, which is calculated by the weighted average of s_{ou} (the ratings of movies that are in the neighbors of the movie *m*). Whether a movie belongs to the neighbors of movie *m* and how large the weight w_{mo} is depends on the similarity between the movie *m* and movie *o*, like whether they have similar actors, etc.

   (c) The characteristic that one model does not have a significant correlation with other models (those models are dissimilar and diverse) makes the ensemble methods improve the overall prediction (Bell et al. 2010).

2. (a) My username: fulinguo

   My friend key: 1410260_936CoWwAcDVL7d06voWGsTDpONoKGVTJ

   (b) I solved the Problem 30 "Digit fifth powers"

   (https://projecteuler.net/problem=30)

   The answer I got is 443839.

   I code in python to get the answer and here is the code:

```
ans=[]
for i in range(2,5*9**5):
        listi=list(str(i))
```

```
        sumi=0
        for j in listi:
                sumi+=int(j)**5
        if sumi==i:
                ans.append(i)
print(sum(ans))
```

(c) The three awards that I most aspire are “Fibonacci Fever”, “Gold Medal”, and “Master of Archive”. I like the award “Fibonacci Fever” because I like Fibonacci sequence very much and this is the prize that awards people who figured out the first twelve problems about Fibonacci numbers. “Gold Medal” awards the first people that solves a problem. It is a very fascinating prize since winning this award means having solved some problems that nobody has solved before. I am the kind of person that would like to solve difficult problems that defeated other people. “Master of Archives” awards to those who solved all problems in archives. I aspire to win this award because it is very amazing and challenging to solve all the problems in the archives. It would improve my mathematical and programming ability a lot after finishing all the problems. (https://projecteuler.net/awards)

3. (a) The HIT I selected was ScoutIt. In this project, participants are expected to identify the business of a receipt based on the image in that receipt. (https://worker.mturk.com/requesters/A3RRY7BIF8JDCS/projects?ref=w_pl_prvw)
   (b) The people who finished this project can get $0.03. That is, the people whose submitted work has been approved by the creator will get $0.03. If the submitted work is rejected, there would be no payment (https://www.mturk.com/help) (The creator also needs to pay fees to MTurk; the amount of MTurk fee is 20% of $0.03, https://www.mturk.com/pricing).
   (c) There are three qualifications: (1) “Location is US” (The Mailing Address the worker provided when creating the Worker Account should be within US) (2) “HIT approval rate (%) is greater than 97.” (More than 97% submissions of HITs were approved). (3) “Total approved HITs is not less than 1000” (The participant has submitted answers that have been approved to at least 1000 HITs)
   https://worker.mturk.com/qualifications/assigned
   (d) The allotted time for this task is 20 minutes. I think I could do ten items in an hour. This implies that the hourly rate is $0.3 per hour.
   (e) The job will expire on November 26, 2018.

(f) If one million people participate in this project, suppose every people finish one item, the creator needs to pay workers 30 thousand dollars. He/she also needs to pay the MTurk fee, which is equal to 20 percent of the amount of payments to workers (6 thousand dollars). Therefore, it would cost the HIT creator 36,000 dollars times the average number of items one people finishes. Therefore, the money would be the highest cost for the creator.

4. (a) I have registered a Kaggle account from the Kaggle home page. My display name is Fulin Guo and my user name is fulinguo.
(b) The competition I would like to describe is "Quora Insincere Questions Classification"(https://www.kaggle.com/c/quora-insincere-questions-classification). The Quora, Inc is the sponsor of the competition. Quora, Inc operates a website called Quora, where people can ask questions and answer others' questions (https://en.wikipedia.org/wiki/Quora). The competition expects participants to design methods that could be used to test whether a question posted by people in Quora is insincere. Specifically, in the competition, participants need to submit the predicted results of the authenticity of every test question (i.e., whether the question is insincere or not). Then, the kernels will evaluate the submission based on the F1 score between the predicted outcome and the real outcome (F1 score is a measure to test the accuracy of classification that takes both the precision and the recall of the test into consideration, https://en.wikipedia.org/wiki/F1_score). The team who ranks first in the competition will win 12,000 dollars, the second will win 8,000 dollars and the team who ranks third will win 5,000 dollars. Regarding the honor code, one people can only register for one Kaggle account and submit outcomes from one account. Private sharing source, including codes and data, cannot be shared outside the team. Participants can only use the competition data for non-commercial purpose and cannot use external data source to complete the competition. The competition started on November 6, 2018 11:59 PM UTC. The deadline of team mergers and entering the competition is January 29, 2019 11:59 PM UTC. This means participants must merger team and accept the competition rule before this date. Participants must submit their solutions before February 5, 2019 11:59 PM UTC. To submit the solution, teams are supposed to store their predictions in a csv file, named submission.csv, then commit the kernel and finally click "Submit to Competition". Teams can submit their

solutions from a script or from a notebook. One important thing is that the csv file they submit should have the correct format and correct number of rows.
(https://www.kaggle.com/c/quora-insincere-questions-classification/rules)
(c) The sponsor will use the best solution to predict whether the questions (or answers, or other comments) that someone posts in Quora is insincere. For instance, the question in the Quora might be the input of the best model, and the model will predict whether this question is insincere or not. If the winning model predicts that it is insincere, the Quora may remove this question from the Quora.

References:

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

https://projecteuler.net/problem=30
https://projecteuler.net/awards
https://worker.mturk.com/requesters/A3RRY7BIF8JDCS/projects?ref=w_pl_prvw
https://www.mturk.com/help
https://www.mturk.com/pricing
https://worker.mturk.com/qualifications/assigned
https://www.kaggle.com/c/quora-insincere-questions-classification
https://en.wikipedia.org/wiki/Quora
https://en.wikipedia.org/wiki/F1_score
https://www.kaggle.com/c/quora-insincere-questions-classification/rules