

Assignment 6

Jiaxu Han

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

Netflix announced a contest in October 2006 and the winner of this contest would receive 1-million dollars. Netflix has its own algorithm - Cinematch - to predict and recommend movies to users at the time. However, “the progress on Cinematch had plateaued” after many years of improvement (Salganik, 2018, Ch 5.3.1), which inspired the company to initiate an open call contest to find better solutions. Netflix asked very structured questions and gave a simple evaluation procedure: they “released a training set consisting of...more than 100 million ratings”, and “the task was to use these data to build a model to predict ratings for a hold-out set of 3 million ratings” (Bell, Koren, & Volinsky, 2010, p24); and the winner would be the first team to submit an algorithm that has 10% or more improvement in predicting the ratings in hold-out dataset over Cinematch, which is evaluated by root mean squared error (RMSE) (Salganik, 2018; Bell, Koren, & Volinsky, 2010). Though the specific criterion function was not given in the paper, it implies that the decrease of RMSE indicates the improvement of the predicting model and thus the criterion function uses the RMSE criterion:

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

Where y_i is the observed value for the its observation and \hat{y}_i is the predicted value (<http://statweb.stanford.edu/~susan/courses/s60/split/node60.html>). In addition, the winning team should have an RMSE at least 10% lower than the RMSE using Cinematch:

$$\frac{RMSE_{Cinematch} - RMSE_{competitor}}{RMSE_{Cinematch}} > 10\% \quad (2)$$

Therefore, the evaluation procedure implies that there is a cutoff beyond which a submission may not be judged at all. Based on equation (2), if the $RMSE_{competitor}$ is larger than $RMSE_{Cinematch}$, it means that the proposed algorithm is even worse than Cinematch and it is not really necessary to consider such a model for competition.

At the beginning of the Netflix Prize contest, the most commonly used method is a collaborative filtering method called “nearest neighbors” (Bell, Koren, & Volinsky, 2010, p25). The nearest neighbors method uses “a weighted average rating of similar items” (Bell, Koren, & Volinsky, 2010, p25) by a user to predict the rating for an item by the same user. The authors mentioned a typical nearest neighbors model (Bell, Koren, & Volinsky, 2010, p25) estimates the rating r_{ui} of the item i by the user u to be:

$$\hat{r}_{ui} = \frac{\sum_{j \in N(i;u)} S_{ij} r_{uj}}{\sum_{j \in N(i;u)} S_{ij}} \quad (3)$$

However, the nearest neighbor method has its own weaknesses, so did other single models. Soon researchers realized that combining multiple models would make better predictions. If a model is not highly correlated with other models, blending it with other models can improve prediction. This still holds even if the RMSE for one set was much worse than the other (Bell, Koren, & Volinsky, 2010, p28).

References

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

Salganik, M. J. (2018). *Bit by bit social research in the digital age*. Princeton, NJ: Princeton University Press.

2. Collaborative problem solving: Project Euler.

(a) Username: hanjx

Friend key: 1409836_dvRvI2PhU6aASUDp8T8UqoFHIWKc7fxd

(b) I chose to solve problem 1: *Multiples of 3 and 5*. Here is the description and the answer for problem 1:

Multiples of 3 and 5

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.

Answer: **233168**

Completed on Fri, 16 Nov 2018, 20:48

Here is the Python code that I wrote:

```
[In [1]: ls = []

[In [2]: for i in range(1000):
...:     if i % 3 == 0 or i % 5 == 0:
...:         ls.append(i)
...:

[In [3]: sum(ls)
Out[3]: 233168
```

(c) I would like to achieve *The Journey Begins* (progress to level 1 by solving twenty-five problems), *C for Commitment* (solve the first one-hundred problems), and *Big Game Hunter* (solve twenty-five of the fifty hardest problems) awards.

Since I just registered an account on Project Euler, I would like to start with just solving 25 problems and get *The Journey Begins* reward. Though this may not be very hard to achieve, I believe this award would motivate me to solve harder problems. Then I would like to solve the first one-hundred problems to get *C for Commitment* award because I just started to learn Python and I would like to have more practice in this language. Finally, once I am more comfortable with Python, I would like to challenge myself a bit more by solving 25 hardest problems and get *Big Game Hunter* award.

3. Human computation projects on Amazon Mechanical Turk.

- (a) I chose a human computation project entitled as “Classify the text into one of the given categories.”
- (b) The most I can see is the reward column in which is written as “\$0.01”. There is no other information about payment structure of this HIT beyond that.
- (c) There are two qualifications that are required for this task: (1) “HCJCPTJJQPHFGURVYMZ has not been granted”; (2) and “ebf7f7a1-60a7-49e7-b06e-26a3c4bf10d8 has not been granted”.
- (d) The allotted time for this task is 25 minutes. I can probably do at least 10 tasks in one hour if allowed. However, the implied hourly rate for this task is $\frac{60}{25} \times \$0.01 = \0.04 .
- (e) The job will expire on Nov. 16th, 2018 at around 7:18 pm.
- (f) The project would cost at most $0.01 \times 1000000 = 10000$ dollars if 1 million people participate in the task.

4. Kaggle open calls.

- (a) I have registered a Kaggle account and this is my profile: <https://www.kaggle.com/hanjiaxu>
- (b)

I am interested in a competition called “Quora Insincere Questions Classification” (<https://www.kaggle.com/c/quora-insincere-questions-classification#evaluation>) which is initiated by Quora - an online platform where users can ask, answer, edit questions, and

share knowledge in almost any fields. The sponsor of the competition is a private company — Quora, Inc., the publisher of the Quora website. The company was founded in June 2009 by two former Facebook employees and is based in Mountain View, California (<https://en.wikipedia.org/wiki/Quora>).

The competition asks competitors to predict whether a question asked on Quora is insincere which is defined as intending to “make a statement” as opposed to looking for helpful answers. Quora provides a training data that includes questions and the corresponding ratings: 0 if the question is a sincere question and 1 if it is insincere. Competitors are expected to submit models that predict each question in the testing set is either sincere (0) or insincere (1).

Each model is evaluated using the F1 score between the predicted and the observed value. The higher the F1 score, the higher the precision level of the model. The F1 score is the “harmonic average of the precision and recall” (https://en.wikipedia.org/wiki/F1_score):

$$F_1 = \left(\frac{recall^{-1} + precision^{-1}}{2} \right)^{-1}$$

A total of \$25,000 will be awarded to the first three people (or teams) on the Leaderboard. The first place will be awarded \$12,000. The second place will be awarded \$8,000, and the third place will be awarded \$5,000. There are several honor code issues associated with this competition. First, each participant is only allowed to sign up to Kaggle from one account. Second, privately sharing code or data outside of teams is prohibited, unless the code is made public to all teams. Third, team mergers are allowed, but the total submission count should be under the limit of maximum submission numbers. Fourth, each team should have less than or equal to 8 members. Finally, every participant may submit at most 5 times per day and select up to 2 final submissions for judging.

Three deadlines are important to note for people who want to join this competition. First of all, each participant must accept the competition rules before January 29, 2019. Second, participants may join or merge teams before January 29, 2019. Finally, February 5, 2019, is the final submission deadline.

Participants should submit files directly from Kaggle Kernels. The submission file should contain a column of “qid” which is the question id in the test data set and another column of “prediction” which is the corresponding prediction value for the question (‘0’ if the question is sincere, ‘1’ if the question is insincere).

(c) Quora has already been using machine learning and manual review to identify and flag insincere questions. It is possible that Quora will integrate the winning team’s method into the existing protocol and develop a better method to weed out toxic content on its website.