

MACS 30000 Assignment 6

Shanglun Li

November 18, 2018

Problem 1.

- (a) The results produced by the submitted models were compared to Netflix's internal Cinematch algorithm. All submissions would be automatically judged by the system, but only the results with significant reduction of RMSE were likely to win the prize. The objective function or criterion function was “Root Mean Square Error” (RMSE), which is defined as the following:

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

where \hat{y}_i is the predicted rating and y is the actual rating in the test dataset (Bell et al., 2010). If a model improved Cinematch more than 10%, it is eligible to become a candidate for a winner of the open call (Bell et al., p. 24).

- (b) The commonly used method at the beginning of the contest was “nearest neighbors” method. This basically computed the “weighted average rating of similar items by the same user” (Bell et al., p. 25). By using this method, the similarities among movies were measured, and the degrees of similarities were used as weights to predict the ratings on movies.
- (c) The author’s “submission was a linear combination of 107 prediction sets, with weights determined by ridge regression” (Bell et al., p. 29). A model would improve the overall prediction of a blend when its correlation with other models is low (Bell et al., 2010).

Problem 2.

- (a) My user name is ShanglunLi. My friend key is 1409383_f3lUdiQgJencaaQ7T7gGEnoxmSSuEGD3.
- (b) I solved the problem 1, which is to find the sum of all the multiples of 3 or 5 below 1000. The answer is 233168. The Python code is the following:

```
def findSum(N):  
    res = 0  
    for i in range(1, N):  
        if i % 3 == 0 or i % 5 == 0:  
            res += i  
    return res  
print(findSum(1000))
```

- (c) The three awards that you would most aspire to achieving are Gold Medal, Master of Archives, and Big Game Hunter. For Gold Medal, I think it is very cool to be the first person to solve the problem. For Master of Archives, I think if I can finish all the problems in the archives, I should be able to handle every coding problem in the rest of my life and get my dream job. For Big Game Hunter, I think if I can finish many of the hardest problems in the archives, I am also able to finish the problems with lower difficulty, which means this award can actually prove my ability.

Problem 3.

- (a) I select the “Classify bird images!” by requester, Research Tasks.

- (b) The reward is \$0.15 per person and you can earn upto \$0.25.
- (c) The qualifications required are Location in US, HIT approval rate (%) is greater than 97, Auto-Granted Qualification For Bird Identification HITs is 1, and Total approved HITs is greater than 5000.
- (d) The allotted time for this task is 60 mins. Thus, the hourly rate is \$0.15 per hour upto \$0.25 per hour.
- (e) The job will expire in 7 days.
- (f) If one million people participate in this project and all of them got the \$0.25 award, the creator will cost \$250,000 in this project.

Problem 4.

- (a) I used my Google account to register the Kaggle account. My username is ShanglunLi, and displayed name is ShanglunLi.
- (b) I find a competition titled “PUBG Finish Placement”, where PUBG stands for a game called PlayerUnknown’s BattleGrounds. The sponsor of the competition is Kaggle, which is an online community of data scientists and machine learners, owned by Google, Inc. It allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges. The submission is evaluated on Mean Absolute Error (MAE) between the candidate’s prediction and observed value, where MAE is defined as the following:

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{n},$$

where y_i and x_i are the predicted value and observed value. The top five predictions on the placement leaderboard will receive Kaggle Swag at the conclusion of the competition. The competition started on Oct 4th, 2018 and will end on Jan 30th, 2019. Candidates can submit a maximum of 10 entries per day. They can use outside data resource. The submission file should contain the Id column and the prediction result column.

- (c) Kaggle provides this machine learning project opportunity for candidates to have a chance to encounter real world data set and practice their data science skills. Candidates can use this competition sponsored by Kaggle to get many useful experience on machine learning. In addition, candidates will have a basic idea about how the competition like so that they can be more prepared to participate the competitions hold by other companies.

References

Bell, R. M., Koren, Y., & Volinsky, C. (2010). All together now: A perspective on the NETFLIX PRIZE. *Chance*, 23(1), 24-24. doi:10.1007/s00144-010-0005-2