Assignment_6_Answer

Question 1:

(a)

The winning team will be decided upon who has made "the greatest improve in root mean squared error (RMSE) over Netflix internal algorithm, Cinematch" (Bell et al., 2010, p24).. This measure is computed by taking the square root of SSR (sum of squared residuals), which goes by the formula below:

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

There is also a threshold to be judged: one's model has to improve the system by at least a magnitude of 10%.

(b)

At the beginning of the competition, Nearest Neighbors is the most used collaborative filtering method for predicting ratings on movies.

(c)

When one model has achieved similar RMSE but is meanwhile highly uncorrelated with other models, blending this one with other components can substantially improve the overall prediction.

Reference

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

Question 2:

(a)

Username: `cken` Friend key: `1407559_Ic4WStyxcZp5F0FdbqoxidQdB0oEXjbo`

(b)

Problem: By listing the first six prime numbers: 2, 3, 5, 7, 11, and 13, we can see that the 6th prime is 13. What is the 10001st prime number?

```
def nthPrime(n):
    '''n is an integer greater than 1'''
    def dividable(x, l):
        flag = False
        for i in l:
            if x % i==0:
                flag = True
                break
        return flag

    primeList = [2]
    idx = 1
    curr = 3
    while idx<n:
        if dividable(curr, primeList)==False:
            primeList.append(curr)
            idx +=1
        curr +=2

    return primeList[-1]

nthPrime(10001)
```

```
104743
```

My answer for the question is: `104743` and it is checked correct.

(c)

Awards I want to aspire to achieve most:

- **Flawless fifty**: This requires me to solve fifty consecutive problems. It is quite a bit of work, but can greatly enhance my problem solving skills and avoids me from shying away from hard problems.
- **Ten out of ten**: To acquire the award, I have to solve the ten most recent problems. New problems are always challenging and I have to work them out as soon as possible, so this demands both creativity and efficiency.
- **Hello World!**: This simply requests me to make my first permanent post. I want to attain the award because being open and communicative on such intellectual platforms to critical to improve my capability of solving hard computational questions.

Question 3:

(a)

The HIT I select: Medical Article Classification from Abstracts. This task asks its participants to classify articles using abstract, title, and other metadata.

(b)

Any participants that complete the task will earn $0.04.

(c)

There are four qualification requirements:

- HIT approval rate (%) has to be greater than 90
- The participant lives in US or GB
- Total approved HITs is greater than 0
- One has to take the Medical Article Classification Test, and earn no less than 90

(d)

The time allotted for this task is 20 minutes. I can presumably do 3 items in an hour, which implies an hourly rate of $0.12.

(e)

The project will expire on 2/21/2019.

(f)

This project will cost at most $40,000 if one million people participated.

Question 4:

(a)

My Kaggle username: cken

(b)

The competition that appeals to me most is one initiated by Quaro: "**Quora Insincere Questions Classification**". The competition is sponsored by **Quora Inc.** itself.

Let's first talk about Quora, as noted on the competition **Quora Insincere Questions Classification**'s overview page, it is a well-known online platform that encourages and empowers knowledge exchange. On Quora, people can ask questions and look for unique insights and quality answers in a style similar to blog posts. They can also connect with other contributors, and subscribe to interesting topics. The project aims to find effective solutions to weed out insincere questions, which they identify as questions founded upon false premises, or those intend to make a statement rather than look for helpful answers.

The sponsor has a held-out test set that will be used to evaluate the prediction accuracy on whether a corresponding `question_text` is an insincere one. They measure they use is F1 score between predicted and observed targets. The F1 takes both recall and precision into consideration when evaluating the classifier's performance. It has a generic formula:

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

There are three layers of prizes: the first place winning prize is $12,000, the 2nd place winning $8,000, while the 3rd place winning $5,000. Only one team will earn the prize for each layer.

To fulfill the project, each team has to keep in conformity with certain honor codes: (a) There are strict policies upon data access and use: one can only use the data for non-commercial purposes, and has to take effective acts preventing any skeptical third party from accessing the competition data. External data is not accepted unless otherwise stated. (b) Besides, any source or executable codes developed upon the competition data are prohibited from private sharing. But one can share the code publicly if this does not infringe intellectual property rights of any third parties. Any use of open source codes has to be licensed under an Open Source Initiative-approved license.

It is vital to follow the timeline: the competition has started on November 6, 2018; the deadline for entry is January 29, 2019 and any team merge must be prior to this; the competition will end on February 5, 2019 11:59 PM UTC.

Successful submission requires the following steps: (1) Compile the final predictions into a csv file, where the format complies with the sample_submission file on the competition's Data page. For this competition, the submission must be named submission.csv. (2) Commit the Kernel, which also has time limits ("GPU Kernel no more than 2 hours run-time"), forbids internet access, multiple data sources and use of custom packages. (3) "Then navigate to the Output tab of the Kernel and Submit to Competition".

(c)

As described above, the project is expected to find effective algorithms to weed out insincere questions. Once there came out a satisfactory solution, the sponsor may immediately put it into practice. They can take advantage of the model to screen out existential insincere questions, and also reject new posts of the kind. They may also collaborate with the winning team to develop the model further, extend its functionality, such as being able to detect unhelpful or hostile answers. All these practices will improve the user experience of this platform and incentivize a friendlier community.

Reference

Quora Insincere Questions Classification, retrived from **https://www.kaggle.com/c/quora-insincere-questions-classification**