Problem Set #6

MACS 30000, Dr. Evans
Keertana V. Chidambaram
(UCID: 12211266)

Problem 1: Netflix Prize and Bell, Koren, and Volinsky (2010)

(a). The competition challenge was to predict the movie ratings of each individual given a massive dataset of many individuals. Each model was judged based on the RMSE value of its prediction w.r.t a hold-out sample. A team is eligible for a prize only if they beat Netflix's internal recommendation system Cinematch by a margin of at least 10%. That is, the RMSE values for the winning model had to be at least 10% lower than the RMSE from Cinematch's model. The team passing this threshold and with the lowest RMSE values (to the fourth decimal place) was to be the winner. However, two teams ended up with the same RMSE value (to the forth decimal place). The tie was broken by awarding the prize to the team that made the earlier submission.

(b). The most popular method for predicting the rating of a movie during the Netflix challenge was called the nearest neighbor method. In this model, the rating of a movie would be the weighted average of ratings of other similar movies, with weights being determined based on the similarity of the predictor movie with the given movie. The more the similarity, higher weight was given to the predictor movie's average rating in the model.

(c). Even blending with inferior models made the solution better off when the inferior model had a lower correlation with the other components of the master model. This meant that an extra model that had low similarity to the other models in the master model had significant impact on lowering the RMSE values.

Problem 2: Collaborative Problem Solving: Project Euler

(a). Project Euler User Name: keertana
Project Euler Friend Key: 1410213_MnK1edG6v0wMkfTD7OAqUoDigvffa1k1

(b). Chosen problem: Problem #15, Lattice Paths [1]
Solution Code:

```
>>> #Project Euler, Problem #15: Lattice Paths
>>> import math
>>> print(math.factorial(40)/(math.factorial(20) * math.factorial(20)))
```

Output:

[1] https://projecteuler.net/problem=15

```
137846528820
```

(c). The three awards that I would love to set as my goals are:

- Big game hunter: Solve twenty-five of the fifty hardest problems.
  My learning curve is usually the steepest when I tackle hard problems. So working towards this award should be a very gratifying experience.
- One percenter: Better than 99% = One badass problem solver.
  This award would be useful because it would help me (and anyone looking through my Euler page) evaluate my position w.r.t other Euler problem solvers.
- C for Commitment: Solve the first one-hundered problems.
  I would prefer this to be my first milestone. This is a more achievable milestone compared to the other two. Hence it would make a great short-term goal to start the ball rolling.

Problem 3: Human computation projects on Amazon Mechanical Turk

(a). Experiment: "AI research - add incorrect answers to commonsense questions". See Figure 1 (Appendix) for the screenshot of the experiment.

(b). The reward is $0.06 persumably for adding incorrect answers for each question (that's the best I could gather from the description and the preview page, although not stated directly).

(c). Eligibility: The user should not have participated in earlier MTurk tasks for creating and verifying the AI questions that are used for this study. Missing eligibility: education levels/ fluency in English of the user, location of the user, a single user can take the survey multiple times.

(d). Total allotted time for the task = 60 mins. Esimated time per task = 1 min. Therefore, a total of 60 questions can be answered in a period of 1 hour. Thus, implied hourly rate = $0.06 * 60 = $3.60/hour.

(e). The job expires on 11/18/2018.

(f). Cost for completing the task 1 million times = 1,000,000 * $0.06 = $60,000 (note that a user can participate multiple times, so it makes more sense to find the cost for completion of the task 1 million times).

Problem 4: Kaggle Open Calls

(a). My Kaggle homepage: https://www.kaggle.com/keertanavc0

(b). The competition that I have chosen to describe is Google Analytics Customer Revenue Prediction [2].

The goal of the challenge is to predict the revenue for customers at a given Google Merchandise Store (GStore). RStudio is the main sponsorer for the challenge. RStudio develops both open source and free tools for the programming language R and they also develop commercial software products (also for R) for organizations.

Solutions are evaluated by using the RMSE value between the logs of the actual (from the given dataset) and predicted revenues (future data from 12/1/2018 to 1/31/2019). The team with the lowest RMSE value will be the winner. There are 2 sets of prizes for winner: (1) The top three teams on the leaderboard will win $12k, $8k, and $5k respectively. (2) The top three teams on the leaderboard who have also used R in formulating the model would win $10k, $7k, and $3k respectively. (Note: a team can win prizes in both the categories).

Honor codes given were that codes could not be shared privately outside a team. However, codes could be shared in public, given the condition that they would be accessible to all participating teams in the competition. The following is the timeline given for the competition:

- Nov 23: final date for entry into the competition by acceptance of competition rules and for team merges
- Nov 30: deadline for the final submission
- Dec 1 to Jan 31: data that will be used as test dataset for model evaluation

The two instructions given for submission were: (1) each team could only make at most 5 submissions per day and (2) for the final evaluation, each team can submit at most only 2 entries.

(c). The commercial product from RStudio are enterprise-level software tools. Predictive algorithms from the winning team, if good enough, might be inncorporated by RStudio in their commercial products. It is worthwhile to note that all winning submissions will have to license their code under an "Open Source Initiative-approved license", so that RStudio gains permission to use the solution. Apart from access to winning algorithms, competitions like these can help convince businesses about the power of predictive analytics, which in the long term would result in a larger customer base for RStudio.

[2] https://www.kaggle.com/c/ga-customer-revenue-prediction

Appendix

Figure 1: Mechanical Turk Experiment