HW#6
Angela Zorro Medina

1. Netflix Prize and Bell, Koren, and Volinsky (2010) (3 points). Read the paper Bell et al. (2010).

(a) Describe how submissions to the Netflix Prize open call contest would be judged? That is, what was the criterion function? Were there any cutoffs beyond which a submission would not be judged (i.e., the fit was so poor that it would be called a zero)?

Before defining how the submissions were judged, we need to define what was the contest. Netflix opens a contest in which the task consisted in using the data provided by Netflix and build a model to predict a rating for a set of 3 million ratings. According to Bell et al. (2010), the model that Netflix was looking for in this contest was a collaborative filter in which collective information is used to make individualized predictions.

Bell et al. (2010) explain that the complex part of these models is that missing some signals can cause an inaccurate prediction, but models overfitting the data can overestimate the weight of small details and miss the big picture of what really drives the process.

A team had to achieve an improvement of 10% over Cinematch (Netflix system) to win. The improvement was measured by the improvement in the root mean squared error (RMSE) of Cinematch. The RMSE is a measure of the accuracy of a model on predicting the values of a population based on a sample. Netflix posted a quiz for which they reveal how well were the teams doing and a test for which Netflix did not provide any feedback before the competition was closed. At the end, two teams were tied, and the final decision was made based on which team submitted first its answer. The winning team won because they submitted their answer 20 minutes before the second team.

By looking at Figure 2 of Bell et al. (2010) and the discussion around the improvements, we can see that Netflix was asking for an improvement of at least 1% of their Cinematch system using the measure of the root mean square error. The submissions that had an improvement below that 1% were not considered.

(b) At the beginning of the Netflix Prize contest, what was the most commonly used method for predicting ratings (stars) on movies?

According to Bell et al. (2010), at the beginning of the competition, the most common method used was a collaborative filtering method based on “neighbor” information. Bell et al. (2010) used the example of the movie “Saving Private Ryan” to explain these type of filtering methods. The “neighbors” of this movie would be other war movies or other movies directed by Steven Spielberg, or the actors of this movie (e.g., Tom Hanks). The filter would use previous ratings of users of those “neighbors” in the prediction of the raiting.

Bell et al. (2010) present this equation to explain these methods:

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

Where, N (i, u) is the set of neighbor characteristics of the movie i that were rated by user u, while s is a measure of the level of similarity between movie i and movie j based on those neighbors (Bell et al., 2010).

(c) The best predictive models in the Netflix Prize open call were hybrids of multiple models (ensemble methods). What characteristic of one model relative to other models made it improve the overall prediction when blended with the other models?

Bell et al. (2010) explain that they merged with other teams to win. The first merge was with a team with a model that used nonlinear blends via neural networks. And the second merge was with a team with a model that incorporated rating frequency to model temporal behavior. Bell et al. (2010) highlight that the reason why they were able to win the competition because their team included people from many different academic backgrounds (engineers, statisticians, and computer scientist). Based on this, they were capable of designing a blend model that differs from others by using individual components that did not need to share anything. This model was defined by the authors as a blend of nearest neighbors models, latent factor models and models that fit nearest neighbors to residuals from matrix factorization.

2. Collaborative problem solving: Project Euler (3 points). On its face, Project Euler is simply a collection of math problems that require clever solutions that make use of understanding of theory and understanding of computation. Project Euler is also a great way to improve your coding, modeling, and problem-solving ability. In addition, these exercises also have some of the characteristics of human computation projects as well as open call projects. The answers are always easy to check (just a number or a vector), and discussion boards arise to discuss the best of all the correct solution methods. Project Euler also incentivizes attention by giving awards for various achievements.

(a) Register as a user of Project Euler. I put my Project Euler friend key (below) on the last page of my CV. Report your Project Euler user name and friend key.

username: ap.zorro263

friend key: 1410280_AeWRe4wWa9PMzZLujSvemB4LGcmpg7Rk

(b) Look through the Project Euler archives of problems. The earlier problems are easier problems. Choose one of the problems and complete it using either Python or R programming languages. Report both your code and your answer.

I answered question 7: 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 10 001st prime number?

The answer I got was: 10001st Primer Number: 104743

Code:[1]

```
#First, we set up the number until which we want R to calculate prime numbers
n <- 10001
#Then we tell R to calculate a sequence of numbers until n=10001
#Basically, we define prime.seq as a matrix of 1 row and 10001 columns.
#We are going to replace each column with the pattern of the prime numbers
prime.seq <- numeric(n)
```

[1] To solve this question, I used these two sources: https://projecteuler.net/overview=007 https://www.r-bloggers.com/project-euler-problem-7/

```
#Then we tell R to set number 1 and 2 of the sequence as 2 and 3. These are going to be first 2 numbers of our sequence. We pick them because the questions start with those 2 prime numbers, that happened to be consecutive numbers

#What this is doing is just replacing our two first columns for the values 2 and 3. Which means that now we have

#[2,3,0,0,.....,0], we have 2,3 followed by 9999 zeros.

prime.seq[1:2] <- c(2, 3)

#Now we define our prime number in terms of our sequence. What we are going to do is basically define our prime_number in

#terms of the column we are looking. Because we defined n as the number we want here we define our prime_number in terms

#of our number that is the second column, which is the last column that we have with a number (3)

prime_number <- prime.seq[2]

#Now we need to create the rest of our prime numbers. We are starting with 3, so now we want to create a sequence of numbers

#But we want to exclude the ones that are not prime, so we tell R to not count them that skip them

#The way I decided to do it was to exclude every even number, starting with 3. I'm only going to use odd number starting with 3, which I program following this logic 3+2=5 is a prime number. So, I program take the previous number and add 2 and that would be the next number in the sequence.

#Because I have to repeat this process n times to find the n prime number I do a loop from column 2 to column 10001.

#In each column R should take the new odd number generated and run the sequence.

#Because we know that not all odd numbers are prime (for example 15), I need to exclude the odd not prime numbers. To do this what I do is to tell R that excludes the values those numbers that when divided by the smaller prime number have a remainder of 0, because that means that it can be divided by other numbers beside itself, which is the definition of a prime number.

for (i in 2: n) {
  prime.seq[i] <- prime_number
  prime_number <- prime_number + 2
  while (any(prime_number %% prime.seq[2:(i-1)] == 0)) {
    prime_number <- prime_number + 2
  }
}

#After running the loop the last prime number generated which means the 10001th prime number

cat("10001st Primer Number:", prime.seq[n], "\n")
```

(c) Look through the Project Euler Progress page. List the three awards that you would most aspire to achieving and describe what you like about those awards.

I selected three awards that are related. I am interested in solving problems first that other

person. Based on this, I choose the awards thinking that to achieve the maximum award, I can try to win the other two first. First, I want to be one of the first hundred persons solving a problem (One in a Hundred Award). Second, after achieving the "One in a Hundred Award," I want to be among the first ten persons to solve a problem (Chart Topper Award). Finally, after recognized with the "Chart Topper Award," I want to be the first person to solve a problem, which will give me the "Golden Medal Award," which I consider to be the maximum award.

3. Human computation projects on Amazon Mechanical Turk (2 points). Sign in to Amazon Mechanical Turk (MTurk) as a worker.

(a) Select an MTurk human intelligence task (HITs) that is a human computation project and IS NOT a survey or an experiment. Most HITs on MTurk are human computation projects.

I selected a projected created by Flying Fish and consisted on rating a set of questions generated from the source sentence based on given criteria. The task did not allow me to see the criteria before being selected to do it. It required a time allocation of 60 minutes.

(b) Describe the full payment structure of this HIT. That is, the reward column says an amount, but there is a lot more information available as to what that amount means.

This HIT has a reward of $0.05 and offers 3,338 HITs. After reviewing the description provided by Amazon Turk with the qualification of "Master has been granted," I interpreted this as in the following way: all workers must be "Mater" workers, which means that 5% extra is charged by Amazon Turk.

Total cost per HIT=reward+20%fee+5%perMasterWorker=0.05+0.01+0.0025=0.0625

Total cost of 3,338HITs=$208.625

(c) Describe any qualifications, eligibility requirements, or restrictions (or lack thereof).

It only has one requirement: Masters has been granted. The master qualification is an award that Amazon Turk gives to workers who have demonstrated excellence in the jobs that she has done.

(d) What is the allotted time for this task? How many items do you think you could do in an hour? What is the implied hourly rate (dollars per hour)?

Because the time allocated to this HIT is 1 hour, the hourly rate is the reward for the HIT, which is $0.05. In this project it was not possible to preview the questions, but based on the description we know they are going to give us a criterion to rate some questions, and that the questions are created using one sentence. This

(e) When does this job expire?

I logged in on Tuesday 13rd Nov 2018, and at that moment the HIT expired in 8 days. That means that it expires on the 21st Nov 2018.

(f) What is the most this project would cost the HIT creator if 1 million people participated in the task?

Total cost of 1'000.000HITs=$62,500

4. Kaggle open calls (2 points). Kaggle is an open call project and dataset platform.

(a) Register for a Kaggle account from the Kaggle home page.
(b) Describe one of the open competitions. Make sure that your description is paraphrased (in your own words) and not just copied and pasted from the text in the open call project.

Include in this description the following information:

- the title of the competition

Quora Insincere Questions Classification-Detect toxic content to improve online conversations

- the sponsor of the competition

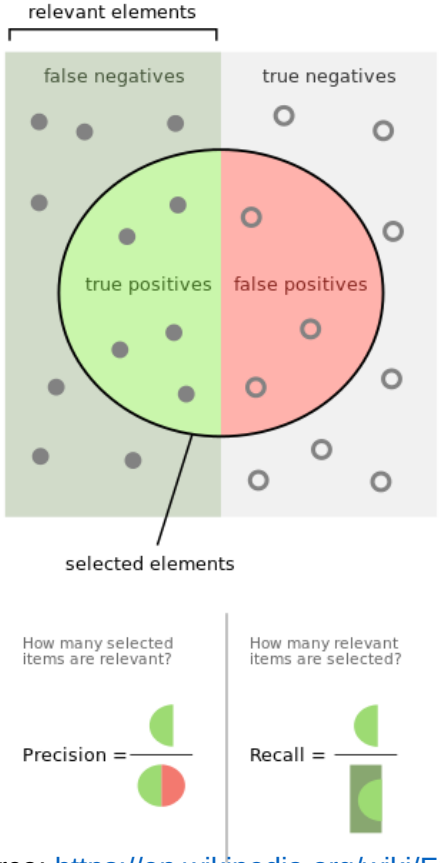
Quora (https://www.quora.com/)

- a description of what type of company or what type of person the sponsor of the project is

Quora is a platform where people ask questions and gets answers. According to Quora web page, their mission is to connect people "who have the knowledge to the people who need it."

- How submissions will be evaluated

According to the evaluation description posted at Kaggle, the submissions are going to be evaluated using the F1 score. The F1 score is based on two things: precision and recall, where the perfect score is 1, and the worst is 0. At Kaggle webpage they refer people to Wikipedia webpage regarding the F Score. At this webpage, I found this graph:



Source: https://en.wikipedia.org/wiki/F1_score

In this particular competition each team needs to predict if a question is insincere (1) or not (0), and there cannot be other outcomes of the model different to 1 or 0. The F-score will depend on the number of correct predictions of the insincere questions. By looking at the graph presented above, we know that precision will depend on the number of correct insincere questions detected divided by the total number of outcomes predicted by the model to be insincere (true insincere + false insincere detected), On the other hand, the recall will depend on the number of true insincere questions detected divided by the total number of insincere questions that existed in the sample.

The team with the highest F-1 scores wins. The second and the third highest scores also win a prize.

- Prize structure for winning submissions

First place wins $12,000, Second place wins $8,000, and Third place wins $5,000.

- Any honor code issues of importance

- Participants can only have one account and cannot submit multiple codes from different accounts.
- Private sharing of code or data can only happen inside the teams, if someone is going to shared data or codes with members of other teams, then it has to be done in the forums of the competition and must be available to all participants.
- Teams can only submit five entries per day and select up to two final submissions to be judged.
- Team mergers are allowed but the maximum size of a team is 8, and the total submissions of the team merged cannot be larger than the daily submissions allowed multiplied by the total number of days of the competition.

- Timeline description

To participate in this competition, teams must accept the rules and join the competition before January 29, 2019. Additionally, this date is the last day teams can merge or add new members to their group. And Finally, teams must submit their final code by February 5, 2019, at 11:59 pm UTC.

The competition started on November 6th, 2018. And the timeline description has a note specifying that the organizers reserve the right to change the deadlines if they find it necessary.

- Submission instructions

According to the terms posted at the evaluation section of this competition, the participants have to submit their code using the following header and format:

```
qid,prediction
0000163e3ea7c7a74cd7,0
00002bd4fb5d505b9161,0
00007756b4a147d2b0b3,0
...
```

The submissions must be made directly to the Kaggle Kernels platform.

(c) Given your answer about what the sponsoring entity does and your description of this project, what do you think the sponsoring entity will do with the winning submission answer? How will they use it?

In this case, it is clear why Quora is interested in sponsoring this competition. The product of this competition will give Quora an algorithm to filter the questions that are posted in their forums. The algorithm will increase the precision of Quora identifying which questions are actually from people looking for answers and not some people trying to hide other types of discourses as questions to use this platform to get an audience of their ideas.

References

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