MACS 3000: Perspectives on Computational Analysis
Problem Set 6

Ma Adelaida Martinez Cabrera

November 18, 2018

1 Netflix Prize

(a) The submission to the Netflix Prize open call contest was judged based on the most significant improvement in root mean squared error (RMSE), as a measure of fit for the model, over Cinematch. To do this, Netflix released a training set consisting of data from approx 500,000 viewers and 18,000 movies that add to more than 100 million ratings. The task was to use these data to build a model to predict ratings for a hold-out set of 3 million. The cutoff for a submission to be judged was 10%, submissions with improvements lower than 10% were not judged.

(b) At the beginning of the Netflix Prize contest, the most commonly used method for predicting ratings on movies was nearest neighbors. With this method, the predicted rating of a movie for a specific viewer was a weighted average rating of similar movies rated by the same viewer.

(c) As is stated in the article, the blend of two methods even if the RMSE for one was much worse than the other, lead to a better result. In general, they said that a blend of $k + 1$ models is always better than a blend of k models. The added model improve the blend if it was not highly correlated to the other components.

2 Project Euler

(a) Project Euler user name and friend key:

madelaida: 1408197_22d7e3QpsNycVfZyzGI9I33QBiVA46ok

(b) **Even Fibonacci Numbers:**

Each new term in the Fibonacci sequence is generated by adding the previous two terms.

By considering the terms in the Fibonacci sequence whose values do not exceed four million, find the sum of the even-valued terms.

Answer: 4,613,732

Code: Fibonacci.R

(c) List the three awards that you would most aspire to achieving and describe what you like about those awards.

1) Baby Steps: Solve three problems. I like this award as a way to start in Project Euler. For me, is easier to commit to small tasks in the begging.
2) One In A Hundred: Be among the first hundred to solve a problem.
3) State of the Art: Solve the twenty-five more recent problems.

For me, the best strategy is to start with small goals and then make them more significant in a gradual way. If I set unattainable goals since the beginning, I lose interest in not seeing results. That is the reason why I would start with "Baby Steps." Then to add a competency component, I will like to achieve "One in A Hundred." To finish, I'll like to achieve "State of the Art," because the most recent problems are probably those of the highest level of difficulty and lower number of solutions.

3 Human computation projects

(a) I chose: Lookup data about software products. Given a company name and software product, lookup info on web.

(b) The reward column says $0.10

(c) Qualifications required: Masters has been granted

(d) The allotted time is 60 minutes. I could do 10 items in an hour. $1 dollar per hour.

(e) The job expires in 7 days

(f) Assuming a normal distribution of number of tasks per hour: $\$1,092,286$

Table 1: Project cost if 1 million people participated in the task

% People	No of people	Min per task	Tasks per hour	Price	Total
2%	20000	10	6.00	$ 0.60	$ 12,000
3%	30000	9	6.67	$ 0.67	$ 20,000
12%	120000	8	7.50	$ 0.75	$ 90,000
18%	180000	7	8.57	$ 0.86	$ 154,286
30%	300000	6	10.00	$ 1.00	$ 300,000
18%	180000	5	12.00	$ 1.20	$ 216,000
12%	120000	4	15.00	$ 1.50	$ 180,000
3%	30000	3	20.00	$ 2.00	$ 60,000
2%	20000	2	30.00	$ 3.00	$ 60,000
				Total	**$ 1,092,286**

4 Kaggle open calls

(a) username: madelaida

(b) **Title:**Human Protein Atlas Image Classification. Classify subcellular protein patterns in human cells

Description: The goal of this competition is to develop a model able to classify mixed patterns of proteins in microscope images. For this task, the teams must predict protein organelle localization labels for each sample in the dataset. There are 28 different labels present in the data set, each label is represented by an integer from 0 to 27. The data set includes 27 different cell types. Four filters represent each cell image, the protein of interest (green), the nucleus (blue), the microtubules (red), and the endoplasmic reticulum (yellow). The green filter should be used to predict the label and the other filters for reference. The submission will be evaluated based on the F1-score. This score is a measure of accuracy, it considers both precision and recall of the test, based on the number of correct labels.

The monetary prizes are structured as follows:

- 1st place - $14,000
- 2nd place - $10,000
- 3rd place - $8,000
- 4th place - $5,000
- Special Prize: NVIDIA Quadro GV100 GPU

The competition sponsor is KTH Royal Institute of Technology, is a university in Stockholm, Sweden specialized in Engineering and Technology. The sponsors of the monetary prices and the special prize are Leica Microsystems and NVIDIA respectively. Leica Microsystems is a manufacturer of microscopes and scientific instruments. NVIDIA is a technology company that designs and manufacture graphic processing units (GPUs).

The timeline is structured as follows:

- January 3, 2019 - Entry deadline.
- January 3, 2019 - Team Merger deadline.
- January 10, 2019 - Final submission deadline.

The most important rules are:

- One Kaggle account per participant. One submission per participant.
- Privately sharing code or data outside of teams is not permitted. It's okay to share code if made available to all participants on the
- Team mergers are allowed and can be performed by the team leader. In order to merge, the combined team must have a total submission count less than or equal to 8 members.

(c) The sponsoring entity will use winning submission answer to enrich The Human Protein Atlas. They will use this to answer questions in medicine, as they stated in the description: “Images visualizing proteins in cells are commonly used for biomedical research, and these cells could hold the key for the next breakthrough in medicine.” The sponsor of the monetary prizes will use these models to enhance the technology on their microscopic systems. For example, they say that they are going to build a tool integrated with their smart-microscopy system to identify a protein’s location from an image.