ORIE 4999 End of Semester Report

--Yujia Liu

**Introduction**

This is a summary for my ORIE 4999 Research project this semester. It’s a brief overview of what I did and things I learned in this independent research class.

**Project Scope**

* At the beginning of this semester, I finished up the Navigate-Informs website, which is a project I started last summer.
* Later on, I started to participate in the Yelp Data Mining project under the guidance of Jian. I have regular weekly meeting with Jian and he assigned me some readings about data mining to let me prepare for this project. After finished those readings, I gained some background information about data mining, which helped me better understand this project. Since I didn’t have the chance to sign the contract for the Yelp project in order to get the access to the actual data, I learned the algorithm developed by Jian and ran some experiments on it.
* Besides this, I also compiled two datasets for the experiments of Massey’s project. One of these from the Yelp Academic Challenge Dataset. This dataset contains Yelp reviews near many universities. We filter the whole dataset on only MIT and Harvard because there are more interactions between these two places. I parsed this dataset into two dictionaries, one is the user product interaction Yij, where I define Yij =1 if user i rates business j a score above 4. The other one is a vector representation of each business j. I used bag of words approach to do this. Among all text descriptions of the business reviews, I selected the most frequent 1000 words as the word clusters. For each business j, I counted the frequencies of reviews of j containing the 1000 words and record it as a vector with dimension 1000 for each business j. The other dataset is from Amazon reviews for fine foods. I used similar approach to parse the data as above.

**Description of Yelp Data Mining Project Participation**

The algorithm models a simpler version of the real project by simplifying some assumptions. We assume that there are two categories of user class and k categories of advertisers. When a user arrives, it belongs to either one of the user classes. We want to show each arrived user one of the k advertisers while satisfying the constraint that each advertiser j has a demand d, which means that advertiser j must display to exactly users. For each advertiser j, there is a click through rate cij models the likelihood user I will click j. We want to get an assignment of advertisers to users in order to maximize the total clicks.

Let Cij denote the click through rate, uij denote the probability of assigning user I advertiser j, λi denote the total arrivals of users from class I, k is the number of advertisers. We model the arrival of users using Poisson distribution and let the rate changes over t. we assume that the total time period is 1 month, so t is from 0 to1. We can use linear programming to solve the probability vector u. The algorithm solves 10 linear programming under 10 different k values ranging from 100 to 1000. Under each linear programing solution u, we did 50 simulations to compare the result of simulation against the linear programming solution. During each simulation, if one advertiser is fulfilled, meaning that it meets the demand, we just exclude it from the list and solve the LP again.

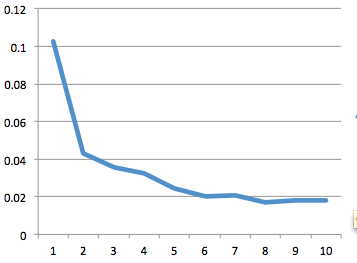
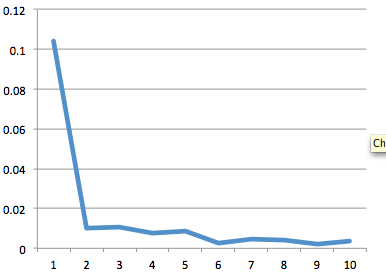
In this way, we got the average gap between LP solution and simulation for each k, as well as the variance of it. Below is a plot showing the result.

Chart 1: Average Gap VS k

Chart 2: Average Variance VS k

Then I did several other experiments with some variations on some of the parameters and assumptions.

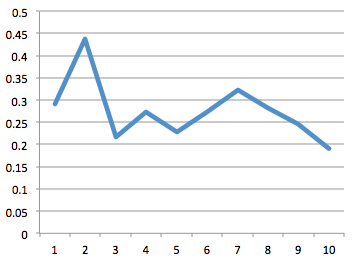
1. Instead of generating the click through cij using uniform distribution, I changed it to normal distribution. Below is the plots of Gap and variance:

Chart 3: gap vs k

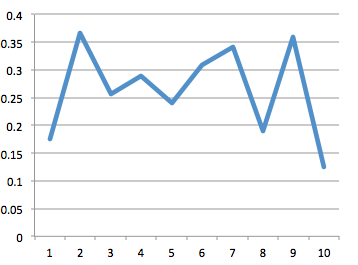
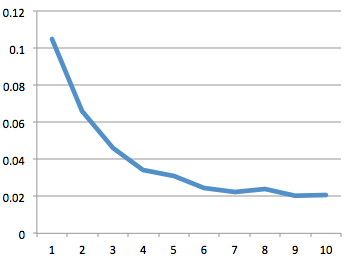


Chart 4: variance vs k

As we can see from the above graphs, as k increases, the gap and variance first fluctuates and then go down. The average gap is slightly higher than the uniformly distributed cij but not quite a lot. It shows that uniform distribution might be a better choice. It also shows that the gap is going down as k increases no matter what distribution cij is, which suggests that our algorithm models well if there is a large population of users.

1. In the original model, we model the distribution of λi, the arrival rate, using 0.5\*t for class1 and 1- 0.5\*t for class2. Now I changed this distribution to t for class1 and 1-t for class2 to see if there is affect of this. Below is the corresponding plots for the new case:

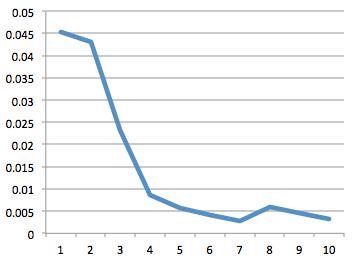
Chart 5: gap vs k

Chart 6: Variance vs k

In this case, the two plots look similar to the previous 0.5\*t distribution that the gap and variance both go down quickly as k increases. It also shows that the distribution of rate doesn’t affect too much on the overall quality and either would be a good prediction.

**Project Summary and Take-aways**

As we can see from the above plots, the algorithm predicts well for this model as k goes larger and larger. Overall, this is a good prediction of advertiser assignment that satisfying all the constraints. By doing this project, I learned how to set up a linear programming problem and write code to solve it. I learned some basic techniques in data mining, as well as how to do simulation to test a model.

Above all, it’s a summary of things I did this semester and my learning achievement in the ORIE4999 class. I learned a lot of theories about data mining and its application in online advertising. I also learned many techniques in linear programming and simulation. I really appreciate to have this opportunity and will continue exploring this field by self-learning in the future.