1 Introduction

One of the main concerns was developing some method to generate a departmental value that can be used to predict buyability. Using sources such as Netflix ratings[3] and customer preference[2], I created customer feature matrices and department proportion matrices to calculate a inner product score.

2 Concept

I want to identify a customer's inner product score or rating for a department. The steps for creating this department "rating" score R_D are listed, where $D = [ws = \text{Woven Shirts}, kt = \text{Knit Tops}, d = \text{Dresses}, p = \text{Pants}]$.

1. For each user/customer, calculate Customer Feature Matrix α_D
2. For each user/customer, calculate Customer Department Proportion Matrix θ_D
3. For each user/customer, calculate inner product score $theta_D^\top \hat{\alpha}_D$

Currently, I define the "rating" for a customer as "cust-dept interaction" attributes. These attributes are features associated with a customer's interaction with a specific department, and these features are described in the next section. But future development on the "rating" will try to include a "cust-dept. popularity" and "cust-dept price sensitivity." The current dataset also tries to take into account "seasonal effects" by observing orders within the months of May 1, 2017 to August 1, 2017.

$$R_D = \text{cust-dept interaction} \tag{1}$$

2.1 Cust-Dept Interaction

2.1.1 Customer Feature Matrix

For each user/customer, a department has attributes/features α_D. Currently, the features used are

$$\alpha_D = [N_D, n_D, S_D, s_D, si_D, D_D, d_D, di_D] \tag{2}$$

where features per department are respectively total number of purchased, average number purchased per order, total amount spent by customer, average amount spent by customer per order, average amount spent by customer per item, total discount amount, average discount amount per order, average discount amount per item. The idea was to track frequency of specific department purchases, how much a customer paid for items, and if sale items or discounts were important for customer to purchase.

In the future, features such as markdown, end use, pay type, item properties, and more will be included. These features were not included because I wanted to keep the initial model simple before adding more features to increase the complexity.

2.1.2 Customer Department Proporation Matrix

For each user/customer, a department has a specific customer department proportion value θ_D. This value is an estimate to the predictive distributions taken from [1][5][6]. To calculate the matrix, we define a few values: N_D being the total number a customer has purchased for a specific department, N_u being the total number of items purchased by a customer across all departments, $K = n$ being the number of departments we are considering $n = 4$, and α_i being a feature in α_D.

$$\begin{aligned} \theta_D &= [\theta_0, \theta_1, ..., \theta_i] \\ \alpha_D &= [\alpha_0, \alpha_1, ..., \alpha_i] \end{aligned} \tag{3}$$

$$\theta_i = \frac{N_D + \alpha_i}{N_u + K\alpha_i} \tag{4}$$

Using the customer department proportion matrix and customer feature matrix, we can define the "rating" as inner product of $\theta_u^\top \alpha_u$.

$$R_D = \theta_D^\top \hat{\alpha}_D \tag{5}$$

3 Future Development

Here are some initial observations that will need to be modified/changed for future development

1. The "rating" favors customers who only purchased in a single department vs. customers who bought across multiple departments.
2. The "rating" is more focused on "how much" and "how many" a customer has purchased

Ideally, I want the rating to not be biased towards purchasing in a singular or multiple departments. Also, I don't want to include customers who just spent alot to give higher ratings. I think this could be solved by including a "price-sensitivity" component on a per item level to determine a customer's "sensitivity" to a normalized priced for each customer, similar to [2]. As a result, a higher price from the normalized value it would produce a negative value that would impact negatively onto the rating. These are things to take into account once we move onto the second version of the modeling phase.

4 Future Modeling/Predicting

I have a few options:

Option A : To check predictablility of the rating value, I will create and run a initial regression models on the purchases using the ratings as a function of time. This will require exploring different timeframes.

References

[1] Feng-Tso Sun, Martin Griss, Ole Mengshoel, and Yi-Ting Yeh. *Latent Topic Analysis for Predicting Group Purchasing Behavior on the Social Web.*

[2] Francisco J.R. Ruiz, Susan Athey, and David M. Blei. *Shopper: A Probabilistic Model of Consumer Choice with Substitutes and Complements.*

[3] Stephen Gower. *Netflix Prize and SVD*

[4] Abhijit Raorane and R.V. Kulkarni. *Data Mining Techniques: A Source for Consumer Behavior Analysis*

[5] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling. *Fast Collapsed Gibbs Sampling for latent Dirichlet allocation*

[6] Thomas L. Griffiths and Mark Steyvers. *Finding Scientific Topics*