Extracting Rating Dimensions from Hidden Topics in Text Reviews: a Better Recommendation System

Xuwen Shen, Xinzhe Yang

November 10, 2017

Abstract

The ultimate goal for the project is to create a recommendation system which recommends restaurants to a specific user given the user's preference and the restaurants' rating with respect to the user's preference. To achieve the goal, the most important part is to extract dimensions for both overall rating for restaurants and user preference using information from reviews. Information we care about includes what a specific user cares about, food or view, price or service and what factors lead to the overall rating for a specific business.

1 Introduction

The same rating from different users usually stands for different meanings. Individual users may assign different weights to such aspects when determining their overall score.For example, a spendthrift hotel reviewer might assign a low weight to "price" but a high weight to "service", thus explaining why their overall rating differs from a miserly reviewer who is only interested in price. There exists hidden information in reviews that leads to the final rating. By extracting the information, we can get where a restaurant shines and what it needs to improve. If we have k topics in reviews then we extract k (the same number as the number of topics in Yelp reviews) dimensions in rating for each topic respectively and then use the k dimensional rating to compute the recommendation score for an individual.

The importance of customer reviews has been proven in terms of giving insights to businesses and also providing customers personalized services.

However, few of previous researches have been studying both the topic models of a specific user and those of a specific restaurant with rating break-downs.

Currently, Yelp has a recommendation that does not take the user's preference and potential matching with restaurants into account. Our recommendation will benefit customers by providing a more personalized user experience.

2 Background

Discuss past research done on recommendation system, especially analyze the methods in the previous winner paper from Yale and paper “Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text”

3 Research Design

Our objective is to learn hidden dimensions of behind a overall rating for a specific restaurant by combining latent rating dimensions, such as the topic learned by topic modeling methods like LDA. First we run topic model to get topics from reviews written by a user and topics from reviews on a restaurant. Then we combine topics we learned from the user and the restaurants as topic factors. Notice that the user preference and topics of restaurants is the same number, as well as topic factors. Finally, we learn the hidden dimension of rating by minimizing the mean squared error function defined by difference between the overall rating calculated by scores of hidden dimensions

and the real rating of a review from one user for a restaurant.
In order to do that, we collected all the reviews from one user to get topics from the text as well as the probability of the topics. In effect, the probability of the topic in all the reviews from each user can be the "preference" of the user. After we got the topics of the user, we subsequently used the same model on all the reviews for one specific restaurant. However, the results we get from LDA lack the adjectives that can distinct the positive or negative quality of the topic. So our next goal is to learn the positive or negative quality of topics.
After achieving the goal of finding the distinct quality of topics, we can then learn the hidden dimension of rating based on the topic and the overall rating.

3.1 Method 1

We will talk about the method that predicts rating subscores for restaurants and then minimizes the loss function.
The $n \times p$ matrix W in the above figure represents the rating score of each topic for each restaurant. Each row represents a restaurant and in that row, each column represents the rating score of a specific topic. The $p \times 1$ matrix M on the right represents the user's preference for each topic. Multiplying the two matrices gives the personalized score of each restaurant for each user.
First of all, to find the column of each user's preference, run LDA on all of text reviews of all restaurants and generate k topics. Then for each user, we extract all the reviews as a corpus and calculate the probability that each word belongs to each of the k topics. Finally we add up and normalize to get the preference column of each user where there is a 0.0-1.0 weight for each of the k topics.
Then to find the rating scores for subtopics for each restaurant in the matrix W, we use the following formula to train a model to predict the score of subtopics. The loss function is essentially the least squared summation of recommended score minus real rating score.

3.2 Method 2

We will talk about the method that is similar to the one discussed in “Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text” which treats both restaurant rating and user preferences as variables.

4 Results

Perform evaluation by predicting the rating score of any restaurant for any user and compares the result with the actual rating. Compare the results of two different methods and explain why one outweighs the other.

5 Further Studies

Explore possible improvements to our current method. For example, better bias and parameter choices; how to apply the codeword method to it without manual labeling everything.