Extracting Rating Dimensions with Text Reviews

Abstract

The same rating from different users usually stands for different meanings. Individual users may assign such aspects different weights 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 the reviews that leads to the final rating. By extracting the information, we can get where a restaurant shines and where it needs to improve. If we have k topics in reviews then we extract k (as the same 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.

1. Research Context

Most recommendation systems with a focus on recommendation algorithms generally require the user to complete complicated and time consuming surveys and rarely consider the user's current context.

The problem is that Yelp provides same rank for everyone. Especially in a diversified country like United States, every people have different taste for food. Some people like Mexican food while others like Asian food. Some people care about taste only while others care about decor and services. Yelp is not addressing this problem right now, but if we can build a system that can identify a user's preferences and provide customized rankings for each individual users, people will benefit more from the service.

Also, the current recommendation system rank the restaurants based on distance ratings and price (low to high). It does not take advantage of reviews. Instead, the system just recommends reliable reviews to the top but does not extract information from them and then recommend the restaurant directly to users.

2. Data/Design

The dataset of yelp reviews and rating stars for restaurants for a specific city. I modified LDA algorithm in the situation where some of the topics of review is already known and then learn modified the method of linking rating and topics in reviews.

3. Methods

Modified LDA, instead of assign topic to the words randomly, I add to some of the reviews a response variable: categories that exactly appear in the review.

1. Draw topic proportions $\theta \mid \alpha \sim \text{Dir}(\alpha)$.
2. For each word
   (a) Draw topic assignment $z_n \mid \theta \sim \text{Mult}(\theta)$.
   (b) Draw word $w_n \mid z_n, \beta_{1:K} \sim \text{Mult}(\beta_{z_n})$.
3. Draw response variable $y \mid z_{1:N}, \eta, \sigma^2 \sim \text{N}(\eta^\top \bar{z}, \sigma^2)$.

Finally, the topic distributions themselves (θd) are assumed to be drawn from a Dirichlet distribution.

Then, we create a transformation from the topic distributions (θd) to the k dimensional rating γ_d.

The recommendation rating is computed by

$$rec(u,i) = \alpha + \beta_u + \beta_i + \gamma_u \cdot \gamma_i$$

where α is an offset parameter, β_u and β_i are user and item biases, and γ_u and γ_i are K-dimensional user and item factors (respectively).

4. Significance

We can create a better recommendation system from the results. The recommendation system can recommend restaurant directly to a specific individual if he or she leaves preference information on the website.

5. Reference

[1] J Linshi. "Personalizing Yelp Star Ratings: a Semantic Topic Modeling Approach." Yale University.

[2] J. Leskovec and J. McAuley. "Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text." Department of Computer Science, Stanford University. 2013.