Improving Star Rating as Trend-Aware Performance Metrics

Kenta Takatsu, Caroline Chang

October 29, 2017

1 Research Context

Past research on the Yelp dataset has focused on extracting latent subtopics from review text using Latent Dirichlet Allocation (LDA) (Huang et al. 2014). Modified LDA algorithms have also been devised to better integrate semantic analysis into generating such subtopics (Linshi et al. 2014). However, what these past studies have not yet explored is a systematic way of assessing business performance — as indicated through star ratings — as well as the prediction accuracy of past predicted star ratings for different businesses. To establish such a metric, which captures temporal performance of business, we devise a fine-tuned classifier that gives an estimate of the average star rating for particular businesses over a given time interval by leveraging the latent information in review content for given categories.

2 Introduction

Our main objective is to develop a streamlined star rating prediction system that can ultimately give users a long-term idea whether particular businesses are undergoing positive and negative trends in the performance. A fine-tuned classifier for a given category would allow us to predict star ratings and provide performance trends over time for businesses. As an initial step to this project, we conducted an exploratory analysis to define the scope of the research; We hypothesized that by focusing on the data-rich category, our method will better capture the subtle trend in review context. Determining whether or not a category has sufficient data-richness depends on two main factors: abundance of reviews per restaurant (and category itself) and an adequate amount of spread of reviews over time. This week, we tackled extracting consistent topics across an entire category of reviews to use later for multi-feature testing and training. Last week, we were able to get a rough idea of how single-feature linear regression was performing, so we ran a comparison between single-feature and multi-feature linear regression using a case study business in the given category. Based on the performance of single-feature linear regression, we predicted that multi-feature linear regression would give better results.

3 Methods and Results

Our experimental design is broken up into several stages: 1) pre-processing, 2) testing-and-training classifiers, and 3) trend analysis. This week, we conducted the pre-processing and began training classifiers as well. We selected 'Chinese' category as a pilot run to demonstrate the proof of concept; we selected this category due to 1) fair data richness (containing hundreds of reviews per restaurant), 2) large data size to show the trend, yet attainable magnitude to compute without an access to a team server, 3) abundant existence of 'sub-categories'. We ran LDA topic modeling with the vocabulary of 100,000, and set the number of topic to be 128 as the previous study suggests (Tian et al. 2016). Generated topic distribution for each review is later used as input data for the classifier. We investigated how representative our LDA embedding is, and observed the cosine similarity of topic distribution across different businesses. Later, we implemented customized k-nearest-neighbor algorithm by using this cosine similarity as a similarity metric. In this method, we aggregated all reviews for each business and create 1-to-1 embedding for each business. The algorithm outputted over 99.9% similarity for all reviews in non-english reviews. This is due to the inherent skewness and over-abundance of English reviews which caused bag-of-words methods like LDA to cluster all non-english words as one topic. After filtering such reviews, the algorithm was able to identify some clusters among Canadian Chinese restaurants (92-95% similarity). To further analyze the performance of LDA topic model, we bootstrapped business that our kNN recognized as ‘similar’, and compare the confidence interval of topic distribution among reviews. As a result, we observed much lower similarity among reviews (.38 .55). We concluded that this is due to the fact that reviews can contain much larger variance even if they are about the same business, which raised the issue of our original implementation — the lack of generalization. Additionally, we used the generated LDA topic distributions to extract the most important topics pertaining to a particular category (in this case, “Chinese”). We wanted to find topics as features that are most relevant when testing and training our classifiers, so that our predicted star ratings are not based on topics that are not the most crucial and even are inconsistent metrics of business evaluation. For every review text for a particular business under the “Chinese” category, we keep track of the topic index of the maximum embedding value. We then count the frequencies of the topic indices and plot a histogram to determine which topics out of the 128 total topics are the most prevalent. We then select the top 3 indices and use them as the multiple features in both our linear regression classifiers.

4 Individual Work

Kenta

- Build an infrastructure that embeds each review text into k-dimensional vector that represents the distribution of LDA topic.
- Implemented the customized k nearest neighbors algorithm by using the cosine similarity of topic distribution as the similarity metric. The result demonstrated an interesting cluster in Canadian Chinese restaurant
- Validated the prediction accuracy by conducting confidence interval analysis. Discovered a large variance within in review topic, even when 2 given reviews are about the same business
- Explored the method to statistically verify “an appropriate time interval” by simulating the random distribution by Triangular Random Distribution and defined the range that generates p value that large enough that we fail to reject Kolmogorov Smirnov test

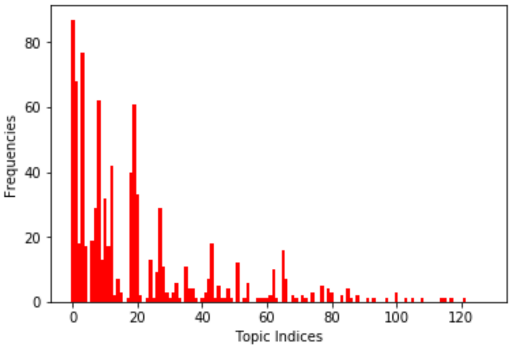
Caroline

- Determined most popular topics from review content for businesses across the entire ’Chinese’ category. The most frequent / popular topics across the entire category were visualized via a histogram and the top 3 topics were determined: topics 2, 1, and 0. Note: LDA does not generate topics in order of descending frequency–thus, we must conduct this check to make sure we do not operate under false pretenses when selecting appropriate topics
- Tested and trained a multi-feature linear regression model using the 3 most popular topics (chosen while examining the absolute peaks in the histogram) on a sample business in the “Chinese” category
- Cross-examined single feature linear regression and multi-feature linear regression and after calculating the mean squared error for both regression models that it is possible that three topics as features is too few in order to truly distinguish multi-feature linear regression and single-feature linear regression. (The mean squared errors were roughly the same.)

5 Further Studies

- The next steps for the following week are to determine how well multi-feature linear regression as well as linear SVMs perform using 3+ features. We will determine the threshold of number of features for there to be a considerable difference between the mean-squared errors of both multi-feature and single-feature regression.

Figure 1: Histogram of topic frequency



- Our testing sizes consistently make up of 60%. While this is appropriate for a baseline assessment of how our classifier is handling data, we would like to ascertain more specifically how big our interval size has to be in order to most optimally predict the next star rating in the subsequent time bin.

- We could potentially look into a possibly more robust metric of extracting most popular and relevant topics after running the LDA model on review text. Principal Component Analysis (PCA) might be a viable method that we could potentially try for feature extraction in the future (as suggested by Richert et al. 2013). However, a pitfall of PCA is that being a linear method, it might not be as flexible when used on non-linear datasets.

References

[1] Huang, J., Rogers, S., Joo, E. et al. *Improving Restaurants by Extracting Subtopics from Yelp Reviews* 2014.

[2] Linshi, J. et al. *Personalizing Yelp Star Ratings: a Semantic Modeling Approach* 2014.

[3] Richert, W., et al. *Building Machine Learning Systems with Python* 2013.

[4] Tian, F. et al. *Sentence Level Recurrent Topic Model: Letting Topics Speak for Themselves* 2016.