CS2750: Project Proposal

Haoran Zhang, Keren Ye

2016-02-16

Abstract:

In this project we want to predict Yelps star rating by topic modeling technology, sentiment analysis, and common machine learning methods. While there are lot of works that predict the relationship between star ratings and reviews, our project is more focus on topic modeling processes, and analysis the sentiment among the topics. Then, we use the result as a new feature to predict star ratings for business. We may need to modify few traditional methods to apply our application. For example, the traditional LDA only can discover hidden topics, but cannot figure out the sentiment among each topic. Thus we have to modify whatever the method itself or features of data to make it adapt our requirement. We hope our project could have an acceptable result compare with other methods.

Introduction:

There are several datasets provided by the industrial companies for the purpose of research. In this project, we shall mainly focus on the Yelp dataset. The Yelp dataset challenge provides academics a good way for practicing machine learning algorithms. It provides not only the rating data, but also the users reviews. It makes it possible to leverage the abundant text information from users review. However, predicting the ratings by using the traditional recsys technologies does not apply the informative textual corpus. Also, simple model such as bag-of-words makes strict assumption of the conditional independent, meanwhile, it suffers the problem of synonyms which means different words may indicate the same topic. Moreover, it is hard to incorporate all of the features into a comprehensive framework.

In our work, we are planning to project both the users and the local business into the topics feature space. There are two advantages: 1) comparing with the bag-of-words representation, the topic-vector representation is more compact, which means it will not suffer the problem of data sparsity; 2) the topic model is easy for human to understand, thus in the future we can apply some semi-supervised labeling to label each of the topic classes. Then, we will schedule to incorporate this representation with extra features to predict the star rating of a new local business given a specific user.

The contribution of our works (although not start yet) may include: 1) create a modified topic modeling method so that it can analysis sentiment among each topic; 2) apply a comprehensive and concrete framework for non-linear regression/classification problem.

Objectives:

The final goal of this project is predict the user rating for a business. Although there are to many ways to reach this target, such as linear regression, collaborative filtering, and so on. While none of them focus on the hidden topics and sentiment among these topics as features. In this project, we divide the final goal into two tasks. The first task is find out the hidden topic and preform sentiment analysis. The second task is predicting the star ratings.

The first task is using topic modeling processes to learn the hidden topic from users reviews from the Yelp dataset. For example, if a users review talks about low price feature of a restaurant, this task aim to find out this feature of the restaurant. In addition, this task also focuses on find features for users, with the same example, that user will have a feature that have more interested in low price restaurant, so that we can use this information to cluster users and businesses and predict user rating in the next task. Current traditional topic modeling processes could figure out the topics such as ”food”, ”price”, ”service”, but cannot figure out the users sentiment expressed in the review for each topics, such as ”good food”, ”low price”. ”nice service”.

In our project, we may use a topic modeling method called Latent Dirichlet Allocation to training our data. However, traditional LDA still have the same problem that only can figure out the topics but not the sentiment among the topic [1]. Since LDA is a kind of statistic machine learning method, so we may modify feature space so that help LDA to figure out the sentiment.

In our second task, the goal is to predict the star ratings of the Yelp dataset. There are several related works for predicting scores. K-nearest neighbor averages the ratings from similar users, and User-based collaborative filtering is a wider form of the KNN approach. However, this kind of approach suffer from the curse of dimensionality. The SVD approach tries to decompose the user-item matrix while it does not apply the informative review information.

In our project, we shall use the KNN and SVD approach as a baseline, comparing with our approach applying the representations extracted from the users review. So the problem becomes prediction of a real value star rating given a representation of the user and a that of the local business. It can be described as:

$$(userrepresentation, localbusinessrepresentation) \rightarrow star_rating$$

It seems that Euclidean distance is a good choice if the dimensions of the user and local business representation are the same. However, we do want a more flexible framework that can incorporate new features. For example, the local business representation may incorporate a feature called the stars_count which is the average star ratings among all users. Also, the review_count seems a good feature that measures the popularity of the local business. Therefore, what we want to model is a star rating given all extracted features.

$$P(rating|userrepresentation, localbusinessrepresentation, averagerating, review_count, ...)$$

Our goal in task II is to model this probability, or instead, approximate this probability. To our knowledge, the linear model family can hardly model the interconnections between features. For example, linear model cannot predict for a large score if both the representation in user and local business says good service. Thus we shall investigate a bunch of approaches such as neural network for seeking the solutions.

References

[1] Blei, D. M., A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” 2003.