STA571: Advanced machine learning

Search Relevance Improvement by Topic Modeling and Ranking Algorithm

March 17, 2016
Authors: Shijia Bian

1 Motivation

The searching engine is widely implemented on various types of websites. The customers expect the searching engine to give them the most relevant result in terms of their inputs. The business also keep improving the accuracy of the searching result in order to provide the customers with good user experiences, and also compete with the their competitors. However, it is difficult to have the searching engine to give the accurate searching results all the time, especially when the inputs are not informative. The traditional way of improving the searching experiences is to manually label those inputs that are not able to result the accurate results. This is very expensive and time-consuming. In this project, I am going to summarize the topic modeling and ranking algorithm. Then, I am going to create a new algorithm by combining these two techniques. The topic model is applied to be an initial filter and return the specific topic the inputs might belong to, and then the ranking algorithm can give the relative relevance for each results under this topic. Other possible ML algorithm will be possibly tried in the final paper, but these two algorithms are the main algorithms used to develop the learning method in this paper. The experiment will be carried on the data provided by the "Home Depot Product Search Relevance" on Kaggle.

2 Problem definition

We are assuming that the training data contains the customer inputs, the returned result, the manually defined relevance score and the detailed description for the returned result. The customer inputs are strings of words. The returned results are item names that are also strings of words. The relevance score is a numeric score, but it only has 13 options. It ranges from 1 to 3 with 0.25 differences. The descriptions are a string of texts describing the returned items. In this supervised learning, we are going to build up a model that can classify these data set in terms of the relevance scores that are like labels. Thus, the input variables are the customers' inputs and the extracted features from the item descriptions. According to the exploratory analysis, the most difficult part is for all the the searching inputs can be found in the item name or item description, but their relevance scores are still not full score and have difference. It is very difficult to build up a sensitive model to detect this difference. Therefore, the combination of the topic model and ranking algorithm can help. The returned results are considered to be topics. The topic model can preliminarily define if the returned results are topics for the returned results. Then the ranking algorithm can give a relevance score. We are expecting that this model can evaluate the the test set and give the accurate relevance score for the customer inputs.

3 Models and methods

The two main algorithms we are going to adopt are the topic modeling and ranking algorithm.

The topic model serves as a preliminary filter before running the ranking algorithm. The way the filter works is providing a score that can be used for the next step of learning. So we want first to conduct the feature extraction. The feature extraction needs to be conducted for three parts: the item description, the input query and the item name. This will first eliminate the stopping words in these three parts. The key words can indeed be extracted, such as the product name, usage, size, special instructions and so on. Then we want to calculate the frequency of the key words appearing in the description and item name. Therefore, we have these situations below: 1). the searching key words do not appear in the keywords; 2) the searching key words partially appear in the keywords, such as $1/2, 1/3, 1/4, \ldots$ of the searching keywords appear in the keywords; 3) all of the searching key words appear in the keywords; 4). The frequency of the corresponding keywords appearing in the description and the product name. Therefore, we can build a model to calculate the probability that the returned result falls into the same category in terms of the item name and the description. This probability will be considered as a score for the next step of ranking modeling.

The ranking algorithm will take the relevance score as a response variable. We will regress this response variable on the extracted features from the last step. So we can see that the probability of getting a specific score in terms of the available features. This can be done through a GLM algorithm.

After training this model, we can test the model in the test data set. The model will output the probability of getting each score. So we can choose the relevance score with the highest probability.

4 Results and validation

The outputs of the model are numeric relevance scores. These scores will be uploaded to Kaggle. And Kaggle will compare the score with the true score and rate it based on the MSE. Because this is the only data available, I expect to carry out this ML methods on more data from different type of websites. The reasoning behind the model should be similar.