Status Versus Utility Classification and Prediction Algorithm for Amazon Products

Linda Wogulis, Yijun Zhang

August 6, 2015

1 Questions

The overall goal of our research project is to be able to identify and label a product as either a status good or a utility good. In order to do this, we must answer two main questions. The first is whether there even is a universal understanding of a difference between the two. That is, does everyone categorize products similarly between these two options? This is a necessary precursor to the overall project because we must have a stable definition for what is is were trying to categorize. The second question, which is the heart of the project, is whether we can train a model to predict the categorization of each product into status or utility. We will be looking for linguistic cues to do so, by looking at peoples reviews of the products.

2 Experimental Design

2.1 Hypothesis 1

To test Hypothesis 1, we will use Amazons Mechanical Turk (AMT) to ask our subjects to collectively label (how many) Amazon product data. AMT is an online labor market that allows one to outsource small tasks (HITs) to a large pool of workers for relatively low pay. The system has been used extensively under different research topics. According to previous study discussing the validity of doing research on Mechanical Turk [1], AMT workers are more demographically diverse compared to standard Internet samples. One concern while using AMT is how to obtain high quality data, especially to avoid spammers who attempts to make as much money as possible that they do not care about the instructions. There are two things we will do to reduce the noise:

Qualification A HIT approval rate indicates how many of the workers responses are accepted by reuqesters in percentage. We will assign our task to the workers only if their HIT approval rate is above a specific threshold. In addition, we will include a captcha question

and state in the instruction that not having the correct answer to this question will result in the HIT being rejected.

Repeated Labeling Previous work has shown that repeated-labeling, the process of having multiple individuals labeling some or all the data points, directly improves the quality of labeled data [2]. We will test on a small product sample to find out how many labelers shall we allow to produce the most cost-efficient result.

The correlation coefficient between the result from different labelers will be computed. The result will be a value between -1 and 1 indicating the strength of association between two variables, where -1 means two variables are perfectly contrast to each other and 1 means perfectly agreed with each other. Our hypothesis is falsified if the correlation indicates weak or non-existent relationships.

2.2 Hypothesis 2

In order to test our second hypothesis, we will mainly be employing linear regression. We will use the labeled data obtained from the first hypothesis in order to accomplish this. The basic idea of linear regression is to use the equation Ax=b, with matrices. A is a matrix of feature vectors, x is the multiplication factor to be solved for, and b is the resulting value to be predicted. For our project, b will be the classification as either status or utility, and A will be feature vectors from product reviews. We are trying to find x by training on a dataset, using different feature vectors in A. For example, we will start with the bag of words model, which will attempt to find the associations between the presence of certain words and whether they indicate a status or utility good. There are plenty of other potential indicators of what we want: punctuation, length of the review, the bag of words using n-grams, etc. There are many other options to explore, and we will be testing each for falsehood.

In order to attempt to disprove our hypothesis, we will use a very similar method to that described above in the first hypothesis. We will have trained the model on a subset of the labeled data, and then we will have to test it on a different set. In order to see whether the produced output is accurate, we will use measurements such as the Pearsons correlation coefficient, which will compare the labels to our predicted values. In order to ascertain how effective our predictor was, we will compare the values to those obtained from a random labeling. That way, we should be able to tell how much of an improvement, if any, weve made. The randomly labeled data will serve as our null hypothesis. Therefore, if our results dont achieve a much better Pearsons value than the random data, our hypothesis will have been proven false.

3 Expected Results

If both hypotheses are supported, our experiment will result in a model that, when given a product and its reviews, is able to predict whether the product is status good or utility good. Since we will be performing sentiment analysis on an Amazon product review dataset to train the model, we also expect to have a lexicon containing the words people tend use to describe each kind of words.

References

[1] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5, 2011.

[2] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622. ACM, 2008.