Assignment 1: Predicting Amazon Review Ratings

Richard Park
r2park@acsmail.ucsd.edu

February 23, 2015

1 Dataset Analysis

The dataset selected for this assignment comes from the set of Amazon reviews for automotive products that can be found at the Stanford University SNAP website [1]. The dataset is formatted as a JSON document where each example is a user review for a product. The format and information represented by each example is shown in Table 1. The dataset consists of 188,728 user reviews of automotive products that are sold by Amazon. The reviews were written by 133,256 unique users for 47,577 unique products. In order to facilitate training a predictive model for user ratings, the dataset was partitioned into training, validation, and testing subsets using a 60/20/20 split of the overall dataset.

An analysis of the training dataset revealed that the mean rating is 4.14, the variance is 1.7747, and the median is 5.0 which is the maximum possible rating that a user can give a product. The average, variance and median rating show that there is an optimistic bias towards higher ratings in this particular dataset. Figure 1 shows the proportion of each rating within the training data.

Due to the low variance of user ratings in this dataset, I believe that developing a model with significantly high performance will be a challenge.

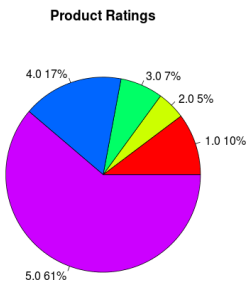

Figure 1: Proportion of each rating in training data

2 Prediction Task

A least squares linear regression model was trained in order to predict the user rating given a set of features that were derived from the fields of the user reviews. The predictor takes the following form:

$$X\theta = y$$

Where X is a feature matrix, θ is a column matrix of learned parameter, and y is a column matrix of predicted labels.

User ratings are represented as integers between 5.0 and 1.0. A predictor based upon a linear regression model was selected because the weighted sum of features including the pseudo-feature would output a real number that represents a predicted user rating. The

Table 1: Fields in the Dataset

product/productID	a unique identifier assigned to each product
product/title	the title of product as it appears to users on Amazon
product/price	the products price if it is known, otherwise the entry shows 'unknown'
review/userID	a unique identifier for the user who created this review
review/profileName	the Amazon user name of the review author or 'anonymous' if the author chose not to divulge their user name
review/helpfulness	a rating assigned to this review by other users that reflects its helpfulness. This is expressed as a ratio of people who found the review helpful to the total number of people who entered a helpfulness response
review/score	a rating given by the user for this product that ranges from 1.0 to 5.0
review/time	time and date that this review was submitted as expressed in Unix time format
review/summary	a user written summary of their review
review/text	a user written review of the product

real valued output can be mapped to an actual user rating by rounding to the nearest integer value. Due to the metric that we chose to evaluate the models the discrepancy between integer labels and real valued output does not affect the efficacy of our experimental method or results. The metric that was used to evaluate the models is the mean squared error (MSE) of actual labels and predicted labels.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - X_i\dot{\theta})^2$$

The baseline against which we evaluated a models performance was the variance of ratings within the training data. This metric was selected because a trivial predictor where the predicted rating was equal to the mean rating of the training data performs with an MSE equal to the variance. It follows that a model which returns a lower MSE than the variance is capable of outperforming the trivial predictor.

For each set of features, we trained a standard linear regression model and a linear regression model with regularization. The results of our experiments motivated the inclusion of a regularised model due to overfitting of the training data when a large number of features were used.

3 Relevant Literature

The automotive reviews dataset that was used for this assignment is a subset of a larger dataset that comprises Amazon reviews from many different product categories (e.g. books, movies, pet supplies, etc)[1]. The dataset was compiled by McAuley and Leskovec to assist with their research into product recommendation models that combine latent review topics and latent rating dimensions [2].

The Amazon review corpus has been widely used in a wide range of research fields due

to it's accessibility, size, and quality. Business and management researchers have used the corpus to analyze what factors determine the helpfulness of reviews [4] and the effect of online reviews on sales of physical product [5]. Natural language parsing techniques have been applied to extract useful information from textual reviews by determining the relevant product features that were reviewed and performing a sentiment analysis to summarize them[6].

Current state of the art methods for predicting ratings based upon review data incorporate latent feature dimensions and textual analysis of the reviews in order to discover the sentiments, topics, and opinions expressed within them.[2][3][7]. The current research informed the selection of features for the model that was trained for the rating prediction task. Due to the scope of this project, implementing a model that incorporated abstract information extracted from the review text was not a realistic goal. However, as will be explained in the following section on feature identification, all models utilized the text in an attempt to improve the performance of the models.

4 Feature Identification

The dataset provides 10 raw fields including the rating of the product. This leaves 9 data fields which can be transformed into a feature or set of features that is compatible with a linear regression model. Analysis of the model helped to eliminate some of these fields as useful features. The userID field and productID field were not considered for incorporation into the model due to the relatively large proportion of unique ID's in the training data. The inclusion of those fields within the logistic regression model would have vastly increased the dimensionality of our feature set. This would result in severe overfitting and a heavy load upon the limited computing resources that were used to fit the models.

Features that were created from the review data fields using feature functions which map the data into a format usable by the linear regression module include the following:

1. anonymity - The userName was used to create a binary feature indicating if this review was submitted anonymously.
2. price - The value was used directly as a feature unless the price was indicated as 'unknown'. In that case the price was set to 0.
3. month - The previous lecture material on Beer review analysis showed that the month of year was a useful feature. The unix time field was converted to a 12 element vector where each element corresponds to a month.
4. review length - We count the number of characters in the users full review of the product.

4.1 Bag of Words

The belief that information extracted from the actual user review text would provide the most gains in performance motivated the inclusion of a bag-of-words feature. A bag-of-words model was created by parsing all of the review text in the training data and identifying the N most commonly used words. The words were then mapped to a binary vector of length N. This value of N became a parameter during the optimization of our model. The feature function for bag-of-words extracts the feature from an example by parsing the review text and determining which words are also in the set of N most common words.

To eliminate sequences of characters that are very common across all types of ratings, an intermediate data set consisting of review text was created where punctuation characters and stop words were removed. This intermediate data set is used to fit the bag-of-words model and create a feature for every example in the data set.

5 Model

As explained in the preceding section on the prediction task, a linear regression model was selected due to the relative closeness with which a rating predicted as a real value could be mapped to the actual ratings which correspond to between 1 and 5 stars. Initially a standard linear regression model was used to explore the performance of the trained models in predicting the labels of the validation set using only features which did not include the bag-of-words. This approach was fine due to the limited number of features that the model consisted of and overfitting was not a problem that was observed. We based our optimization decisions upon the MSE between training/validation labels and predicted labels of those respective sets.

Once we incorporated the bag-of-words feature into our model, problems with scalability and overfitting were observed. We parameterized N, the number of words to include into our bag-of-words model, so that we could use it to guide the optimization process. As N grew larger the time to fit a model increased exponentially and the degree of overfitting increased. At this point it became clear that a regularization term would be required to create a model that performed well on the validation set. A new round of models were fit and validated with L2 regularization. The optimization of these models included a parameter for regularization strength, λ. The optimal parameters of N and λ were discovered by way of grid search and assessing the performance of each model against the baseline variance of the training set.

6 Experimental Results

Table 2 shows the training performance, expressed as mean squared error, of predictors based upon a single feature. Each feature, except the anonymous submission feature, performs better on the training set than the baseline predictor which has an MSE equal to 1.7747. An interesting result is that the validation error is lower than the training error for all of the features. This runs counter to the result we expect where validation error is equal to or greater than training error. The performance gained over the baseline predictor is quite low. The review length in character achieved the best performance, however it's coefficient of determination shows that it is slightly better than the baseline $R^2 = 1 - \frac{1.7386}{1.7747} = 0.0203$.

Table 2: MSE of models incorporating a single feature

Feature	**Training**	**Validation**
Price	1.7728	1.7445
Month	1.7713	1.7447
Review Length	1.7658	1.7386
Anonymous Submission	1.7747	1.7478

Our next set of experiments involved training models which combine the single features into a model as shown in table 3. The figure shows that the anonymous submission feature, which was the poorest performing in the previous experiment, does not increase the performance of the predictor.

Table 3: MSE of models with multiple features

Features	**Training**	**Validation**
Price, Month	1.7693	1.7412
Price, Month, Review Length	1.7603	1.7318
Price, Month, Review Length, Anonymous	1.7603	1.7318

The bag-of-words models were fit and tested without additional features in order to facilitate a search for the optimal set of parameters. Table 4 shows that that regularization has a significant effect on preventing overfitting of the model to the training set, but different strengths of regularization have no discernable effect. The figure also suggests that the optimal N is 4.

Table 4: MSE of validation set for different combinations of λ and N

λ/N	**2**	**4**	**8**	**16**	**32**
0.0	3.35	3.034	2.596	2.249	1.995
0.1	1.756	1.748	1.753	1.757	1.759
1.0	1.756	1.748	1.753	1.757	1.759
10.0	1.756	1.748	1.753	1.757	1.759

Once the optimal value for λ and N were determined through experiment, three models were created and evaluated against the test set. Table 5 presents the results which show that the model that uses only the bag-of-words featured performed the best.

Table 5: MSE of different models on the Test set. $\lambda = 1.0$ and $N = 4$

Features	**Test**
Bag-of-Words	1.7482
Price, Month, Review Length	1.7632
Price, Month, Review Length, Bag-of-Words	1.7542

7 Conclusion

The coefficient of determination (R^2) for the top performing Bag-of-Words model is $R^2 = 1 - \frac{1.7482}{1.7747} = 0.015$. This is far less than what I would have liked to achieve. I believe the main challenge in predicting the rating for the automotive dataset is that the variance of ratings is already low and that the median rating is 5.0. This is similar to the unbalanced binary classification problem where there is an overabundance of one type of label. A multiclass SVM that treats each rating as a separate class would most likely outperform the logistic regression model because it could be tuned so that the imbalance among classes could be compensated for by assigning different weights. This is something that I would pursue given another round of experiments.

Another promising direction to take is trying to extract more useful information from the review text by using more powerful natural language processing techniques. Surprisingly, the bag-of-words model outperformed the model which incorporated features from the review meta data, but it is still a very naive way of modeling text because much information regarding structure and semantics is lost. An approach that incorporates N-grams, sentiment analysis, or topic modeling would be interesting and likely perform better according to the performance of the bag-of-words model.

References

[1] http://snap.stanford.edu/data/web-Amazon-links.html 2015

[2] J. McAuley and J. Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. RecSys, 2013.

[3] G. Ganu, N. Elhadad, and A. Marian. Beyond the stars: Improving rating predictions using review text content. In WebDB, 2009.

[4] SM Mudambi, D Schuff. What makes a helpful review? A study of customer reviews on Amazon.com. MIS quarterly, 2010

[5] JA Chevalier, D Mayzlin. The effect of word of mouth on sales: Online book reviews. Journal of marketing research, 2006

[6] M Hu, B Liu. Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international, 2004

[7] H Wang, Y Lu, C, Zhai. Latent aspect rating analysis on review text data: a rating regression approach. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining Pages 783-792, 2010