Andres Santana A99420007

CSE 190 – Data Mining

Assignment 2

Abstract

For this project, we train a simple linear regression model to predict the taste rating of a beer. Dataset are beer reviews from *RateBeer* and the attempt is to find the best features of the data to predict accurately and measure performance by MSE.

1. Dataset

Dataset consists of beer reviews from beer review website *RateBeer*. Reviews contain product and user information, along with plaintext review and ratings of five different aspects of the beer: appearance, aroma, palate, taste, and overall rating. A sample of the data is presented here:

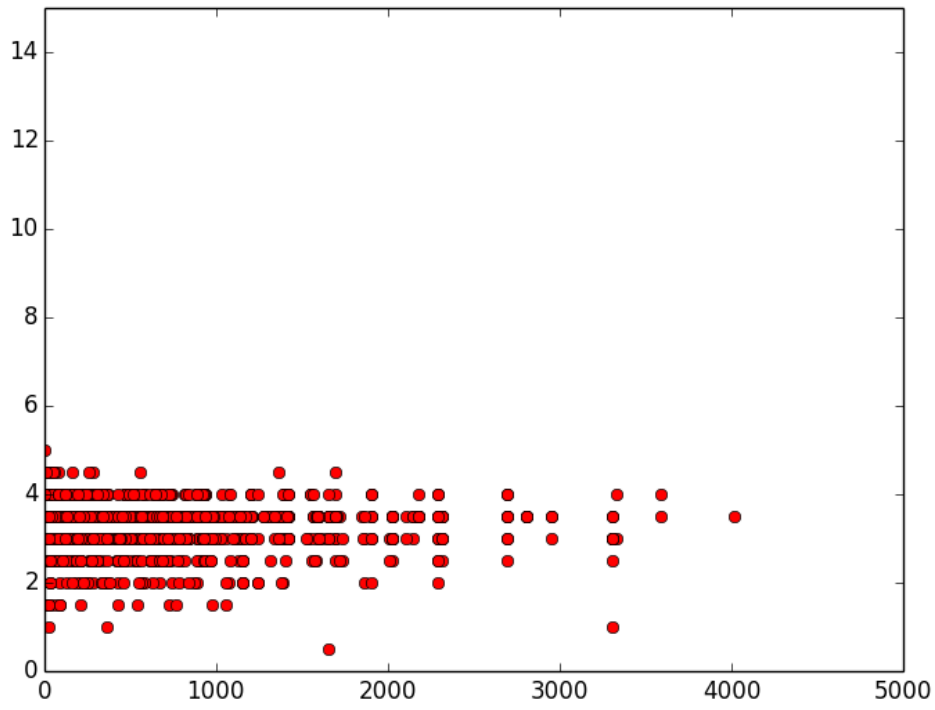
```
beer/name: John Harvards Cristal Pilsner
beer/beerId: 71716
beer/brewerId: 8481
beer/ABV: 5
beer/style: Bohemian Pilsener
review/appearance: 4/5
review/aroma: 5/10
review/palate: 3/5
review/taste: 6/10
review/overall: 14/20
review/time: 958694400
review/profileName: PhillyBeer2112
review/text: UPDATED: FEB 19, 2003 Springfield, PA. I've never had the
        Budvar Cristal but this is exactly what I imagined it to be.  A clean
        and refreshing, hoppy beer, med bodied with plenty of flavor.  This
        beer's only downfall is an unpleasant bitterness in the aftertaste.
```

Dataset includes reviews up to November 2011 and contains 2,924,163 data entries from 29,265 different users reviewing 110,369 different beers from 7,547 different brewers.

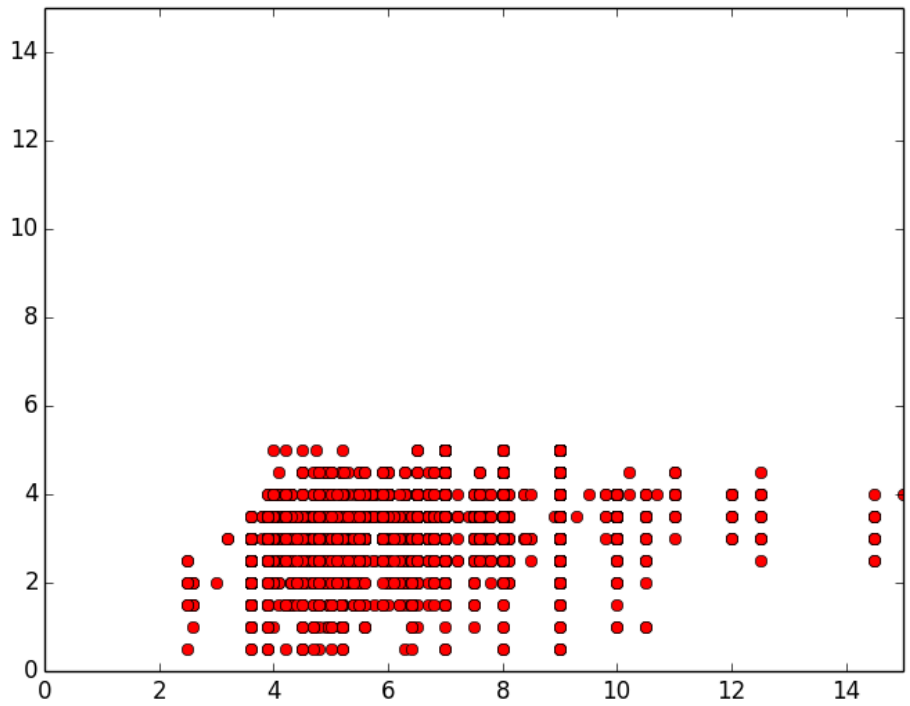
Now, since aspects have different rating range (out of 5, 10, or 20), we normalized data to be all ratings out 5. Additionally, since data set is quite large, we used 900k data entries for training and 100k for validation chosen at random. Moreover, we noticed out that many of the reviews are missing the “beer/ABV” number, so we also made sure the training and validation data included data entries that contained it. Next statistics and properties are made on the training data. Amongst this data there are 19,262 users, 35,953 beers, 2,810 brewers, and 88 beer styles.

We have mean for 'review/appearance' : 3.47224, mean for 'review/palate' : 3.29604, mean for 'review/overall' : 3.35215, mean for 'review/aroma' : 3.23025, mean for 'review/taste' : 3.27559, and mean ABV : 6.64663. We can see that palate and taste reviews are the closest to overall, so we choose to try to predict taste, since it seems to be intuitively the most important aspect of a beer.

Here we show a plot of review/taste versus user popularity, where user popularity is the amount of reviews a user has made (less data was used to see a better distribution):



Here we can see that the less experienced a user is, the more the rating varies, and while they gain more experience they seem to be more concise. Now we show a plot of rating versus beer/ABV (less data was used to see a better distribution):



The distribution here is more spread, but it still seems to have some correlation.

2. Predictive Task

The predictive task will be to predict the review/taste rating based on certain features of the data entry. As a baseline, we will have a very simple model where it just predicts the average of taste review seen in training data. The features we consider are the beer/ABV value, beer/style, brewer popularity, beer popularity, and user experience. The beer/style we form a categorical set of features like [1,0,0,0,…,0], [0,1,0,0,…,0], and so on depending on the beer style (like the example of building features based on months seen in lecture). We consider beer/style since we think the taste is highly correlated with the style, and hence to the review/taste. The brewer popularity is the amount of times a brewer appears in the dataset. We consider the brewer since we think how popular a brand is and some kind of loyalty to a brewery may affect the user's

rating. Beer popularity is simply the amount of times a beer is reviewed, and the user experience, as mentioned, is the amount of reviews a user has made, which is somewhat related to the amount of beers they've had, hence the experience. The baseline will be a very naive predictor that simply output the average overall rating from all the training data for all test data.

3. Model

The model will be a simple linear regression of the form:

$$review/taste = \theta_0 + \sum_{f \in \{features\}} \theta_f * f$$

Basically, θ_0 is a constant and the rest is θ for that feature * the feature of that data entry, where {features} = {beer/ABV, beer/style, brewer popularity, beer popularity, user experience}.

We will measure the validity and performance of the model by MSE. Other models considered are basically using a subset of the mentioned features to see how they work best for our prediction. Also, we attempted to use brewers the same was as beer/style, but since there ware quite a large amount of brewers, the feature vector was almost 3,000 in size and it was quite power consuming to calculate a linear regression. Therefore, we decided to use the popularity of a brewer instead.

4. Related Literature

Dataset used is from the provided datasets in http://snap.stanford.edu/data/ (in Lecture slides) and was gathered from beer review website *RateBeer*. This data set was used in some previous research [1, 2] in the contexts of sentiment analysis [1] and product recommendation [2]. This dataset has been used together with similar data from beer review website *BeerAdvocate* [1] to try to understand how multi-aspect ratings data (having different aspects of a product that influence a user's opinion) can help learn about the use of language to describe those aspects, how the aspects are connected to each other, how certain words can be positive or negative depending on the aspect, and so on [1]. Additionally, this dataset has been used with similar datasets of product online reviews from *BeerAdvocate*, wine reviews

from *CellarTracker*, and reviews from *Fine Foods* and *Movies* categories from *Amazon* [2] in an attempt to better recommend products based on user experience, under the hypothesis that users acquire a "better" taste and experience as they consume more products [2].

Though the dataset is not used for the same end goal as we are using it, the existing work and its conclusions [2] suggest that accounting for user experience (as to how they will enjoy a certain product), as we did, is a valid approach, since we found that indeed user experience seems to be related to the rating given.

5. Results and Conclusions

For the baseline, the result for the MSE is:

$MSE_{train} = 0.6026736995$

$MSE_{valid} = 0.5948392787$

On a model with {features} = {beer/ABV, beer/style, brewer popularity, user experience, beer popularity}, recalling that beer/style are features like [1,0,0,…,.0] for each style, the result for MSE is:

$MSE_{train} = 0.4277513353$

$MSE_{valid} = 0.4081067291$

On a model with {features} = {beer/ABV, beer/style, user experience, beer popularity}, the result for MSE is:

$MSE_{train} = 0.4327220467$

$MSE_{valid} = 0.4128721811$

For the above models, parameter vector is too big to show here, but we can analyze the parameters on the next simpler models.

On a model with {features} = {beer/ABV, user experience, beer popularity} the result for MSE is:

$MSE_{train} = 0.5042208555$

$MSE_{valid} = 0.4809986938$

Here the parameter vector looks like [2.39990747e+00, 1.20188847e-01, -1.62740634e-05, 1.97876614e-04]. Seems like beer popularity has a little more weight than the user experience, which seems intuitive since popularity can be a strong bias for a user's impression. Additionally, parameter for user experience is negative, which means the more experienced a user, the lower the rating they give. This kind of follows the plot we saw earlier, and it also seems intuitive that the more experienced a user is, the stricter they are on their preferences.

On a model with {features} = {beer/ABV, user experience} the result for MSE is:

$MSE_{train} = 0.5189229540$

$MSE_{valid} = 0.4896890272$

Here parameter vector looks like [2.47428858e+00, 1.26396358e-01, -4.53375452e-05]. So, we can see that the beer/ABV has a bigger weight on result than user experience per unit. However, the beer/ABV range is way smaller than the user experience, which can be thousands of reviews per user. Then, with this in mind, the user experience has some heavy weight on the result.

As we can see, our model and choice of features had the best performance by MSE on the train and validation set. Additionally, there was no problem with over-fitting. We beat the baseline by a considerable amount and we show how a simple regression model can be a powerful predictor. We could keep adding more and more features like how much beer a certain brewer produces, or even add some features based on the review text, but as we add more features we might get to a point where we over-fit.

6. References

[1] J. McAuley, J. Leskovec, and D. Jurafsky. Learning attitudes and attributes from multi-aspect reviews. ICDM, 2012.

[2] J. McAuley and J. Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. WWW, 2013.