Predicting beer overall rating from all aspects

Yunli Wang - A53080269

February 24, 2015

Abstract

In this assignment, we try to expand the existing simple linear regression model seen in homework into a slightly complete one. Utilizing ratings from all the aspects of a beer, we are able to give a better prediction for the overall rating. Obviously, this model is very limited as it does not take in to account the review text which contained a lot more information than the numeric rating. Nevertheless, it still shows that even a simple model like the linear regression can achieve good results given enough data.

1 Introduction

The beer data used in this assignment is the "beerAdvocate" data from the assignment description. Originally it contained 1,586,259 from January 1998 to November 2011. A sample data entry is presented here.

```
beer/name: Sausa Weizen
beer/beerId: 47986
beer/brewerId: 10325
beer/ABV: 5.00
beer/style: Hefeweizen
review/appearance: 2.5
review/aroma: 2
review/palate: 1.5
review/taste: 1.5
review/overall: 1.5
review/time: 1234817823
review/profileName: stcules
review/text: A lot of foam. But a lot. In the smell some
    banana, and then lactic and tart. Not a good start.
      Quite dark orange in color, with a lively
   carbonation (now visible, under the foam). Again
   tending to lactic sourness. Same for the taste.
   With some yeast and banana.
```

We can see that most of the data are in the form of "free text review" which requires some text mining technique to learn. It is covered in [1]. Due to the time limitation of this assignment, it will not be explored here. Other than that, the most indicative information are in the numeric ratings in 5 different aspects: appearance, aroma, palate, taste and overall (impression). In this assignment, we will focus on predicting overall rating from the subjective ratings for all 4 aspects, as well as the objective ABV (alcohol by volume) number. We want to see that how much does each of these ratings affect users' rating on the overall impression. The intuition is that the overall rating should be very closely related to all 4 subjective ratings and loosely related to the ABV number.

2 Dataset

The dataset contains much more reviews than a single laptop can handle efficiently. So in order to reduce the workload, we randomly chose 100,000 reviews from the dataset, which represents about 6% of the original data. We consider this a reasonable amount for processing. Furthermore, we divide these reviews into two parts, 50,000 reviews in the training set and the other 50,000 for testing. When sampling data from the original dataset, we found out that many of the reviews are missing the "beer/ABV" number. Although it might not be a significant indicator for the final overall rating, we still want the data to be as complete as possible given the large number of raw data at our disposal. Therefore, during the preprocessing phase, we only chose those reviews with all the ratings and ABV number present. It was not a difficult task to retrieve 100,000 such reviews from the original dataset.

2.1 Exploratory analysis

We want to see how the distribution of the ratings look like. So we plot a simple histogram of each aspect, shown in Figure 1.

We can see that most of the ratings for the individual aspect mostly conforms with the overall rating, with taste rating highly similarly distributed as overall rating. While ABV distribution does not really show much correlation to the overall rating.

To further see how much the aspect ratings are related to overall rating, we plot the number of reviews according to the cross-rating of aspect rating and overall rating, as shown in Figure 2

We can see that "taset" and "palate" seem to have a narrower distribution over "overall" rating, which indicate stronger linearity. Meanwhile, other features show a wider range of rating distribution.

3 Predictive task

As mentioned earlier, the task of this assignment is to predict overall rating from the ratings for the individual aspect and the ABV number.

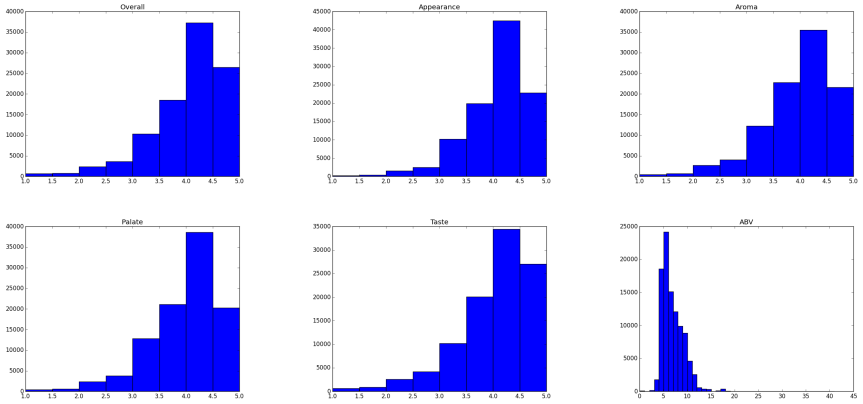

Figure 1: Histogram of ratings for each aspect and ABV

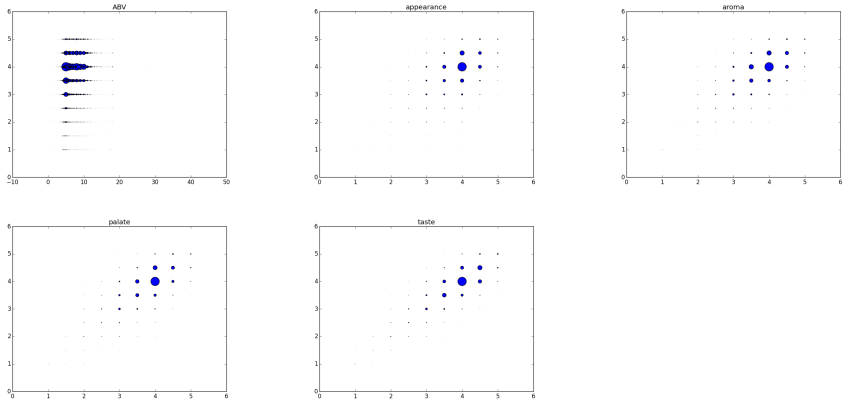

Figure 2: rating collision circles

4 Features

As mentioned earlier, the features selected for this task are the numerical ratings for the individual aspects. Note that we specifically chose those reviews with ABV number to make sure our features are complete for the model.

5 Model

As described earlier, the model will be a very simple linear model,

$$y = C + \sum_{f \in \{features\}} \theta_f \cdot rating_f, \tag{1}$$

where {features} = {ABV, appearance, aroma, palate, taste}

We will measure the performance by RMSE(root mean square error). The baseline will be a very naive predictor that simply output the average overall rating from all the training data for all test data.

For the baseline, the naive predictor achieved:
$RMSE_{training} = 0.713087652396$,
$RMSE_{testing} = 0.71708275673$.

To train the linear model with all the features, we modify the feature vector to contain a constant 1 to represent the constant in Equation 1.

The theta vector calculated for this linear model is [0.52654145 -0.04210266 0.04935083 0.07560512 0.27135139 0.55255899]. We can see that "taste" is indeed the most significant indicator for the overall rating and "ABV", "appearance" and "aroma" are only marginally indicative of the overall rating. The RMSE obtained by this linear model is:
$RMSE_{training} = 0.411611510241$,
$RMSE_{testing} = 0.411365643637$.

Out of curiosity, we decided to remove all the loosely related features and see what would happen if we only include palate and taste as indicator. The RMSE obtained by this simplified model is not much worse than the original model:
$RMSE_{training} = 0.422785449901$,
$RMSE_{testing} = 0.421756939338$.

This confirms our observation that taste is the most closely related to the overall rating and palate is the second. By looking at only these 2 features, we were able to very accurately predict the overall rating. This is also very intuitive as the most determinant feature of beer is its taste and palate. Even if the aroma and appearance is not that good, if the taste is good enough, the beer will still get a good overall rating.

6 Conclusion

Due to time limitation, this assignment only did very limited work on a very simple model. There are a lot more to be done with this dataset, especially those techniques that could have been used to discover the information stored in all the text reviews. Further into the course, we could explore deeper into that sort of problem.

References

[1] J. J. McAuley, J. Leskovec, and D. Jurafsky. Learning attitudes and attributes from multi-aspect reviews. In *International Conference on Data Mining*, 2012.