Project 1: Beer Reviews Linear Regression
Due October 19

1 Description

The goal of this project is to build linear regression models to predict a user’s ratings of a beer on 5 dimensions: appearance, aroma, overall, palate, and taste. Since you have 5 different response variables, you will most likely want to fit 5 different linear regression models.

You will have the following predictors at your disposal: various features about the beer, various features about the user (e.g., gender, age), and the raw text of the review itself.

There are several required components of this project:

1. Feature Engineering: You should transform features and construct new features out of existing features. One particularly rich source of features is the raw text of the review. Maybe you want to create features based on whether certain words appear in this text or not. Or maybe you want to run a sentiment analysis program on the review. The sky’s the limit!

2. Model Selection: Since there are countless potential features, you will need some way to choose a good model. Carefully document the different models that you tried and how you decided between them.

3. Prediction Competition: You will be given both a training and a test set. You will submit the predictions on the test set to Kaggle (http://www.kaggle.com/c/beer-ratings/) and compete with the other teams in the class to achieve the lowest prediction error (RMSE). But remember: you are only allowed to use linear models!

4. Model Interpretation: Although linear models rarely yield the best predictive performance, they are the most interpretable. Interpret the final form of your models. Produce visualizations that help others understand your model. (Obviously you won’t be able to make one scatterplot that shows 10 variables at once, so be creative! For example, if you have a continuous predictor, you may want to make a scatterplot of just that predictor and the response variable.) You might also find it fruitful to compare your models. For example, you fit separate models for predicting the rating of the appearance and for predicting rating of the taste. Most likely, these two models will be different, with different coefficients and perhaps even different variables. What insights about the appearance and the taste of a beer can you glean by comparing the different models that you chose?
2 Getting Started

The data files are available on JupyterHub in the /data/beer/ directory and on Kaggle (http://www.kaggle.com/c/beer-ratings/).

3 Deliverables and Submission

• You must submit a hard copy of a 5-10 page write-up documenting your team’s attempts at feature engineering and model selection. It is recommended that you work on this write-up with your teammates in Google Docs, for easy collaboration. [10 points]

• In class on October 19, your team will give an in-class presentation that explains your final model to the rest of the class. You will be graded primarily on how well you interpreted the model and communicated it to us, so be sure to include lots of visualizations in your talk. It is recommended that you work on this presentation with your teammates in Google Slides, for easy collaboration. [10 points]

• You must make at least 5 submissions to Kaggle. You will be graded on your performance in this competition. [5 points]

Total: [25 points]