**36-350: Data Mining.**

**Tom Sullivan, Beer Analysis.**

## Executive Summary

There is a model behind beer consumer preferences. Drinkers dislike bitter beer and have mixed preferences regarding the other chemical components that make up beer. Based on a large set of data I have been able to build a model that predicts a consumers taste for beer given its chemical makeup. On average, this prediction model is accurate within five percentage points of the consumers preference rating.

## Introduction

The objective of our analysis is to understand the relationship between chemical properties of a beer and its mean consumer preference rating. Our data is from 9741 blind taste tests Molson Breweries conducted in major Canadian cities. A total of 91 beers were tested. The chemical properties of each beer were recorded, which breakdown into 18 analytical and 17 volatile variables variables. The consumer preference rating was measured on a nine point scale, with a rating of 1 representing "dislike extremely" and a rating of 9 representing "like extremely." A mean consumer preference rating of 5.76 tells us Canadians have an affinity towards beer, or at least the beer being tested. All calculations were performed on Matlab.

## Preliminary Analysis

The quality of data is high. Out of the 3158 observations, 91 beers with 35 properties for each beer, only 48 observations are missing. Those missing values represent less than 2% of the data recorded.

A simple boxplot (see attached) of the normalized beer properties reveals either a normal or exponential distribution for each property, which is to be expected for such a wide array of samples. While there are a significant number of outliers in several categories, most of these categories have a low correlation to the consumer's mean preference rating and consequently are not of high concern.

Preliminary analysis of correlation, after replacing missing values with zeros, failed to reveal any medium or large correlation between beer properties and mean preference.

This calculation was performed with the correlation coefficient tool, `corrcoef`. The strongest relationship was bitterness, with a negative correlation of `0.293`. While this relationship is small, the p-value of this statistic reveals a `0.0047` probability of getting a correlation this large by chance, hence the correlation is significant.

The top five correlations, all with p-values less than `0.06`, were: bitterness, 2-methyl-1-butanol, formazin turbidity, carbon dioxide, and Ethyl Octanoate, from largest to smallest significance.

## Missing Values

Before one can fully analyze the beer data and build models it is necessary to fill in the missing values. It is possible that these values are missing because they were outside of the instrumentation's measurable range. Unfortunately, it is quite difficult to see trends in data that does not exist, I can only assume that the missing values were within range. A peculiarity of our data is the pattern of the missing observations. The majority of missing data was concentrated among two beers, 28 and 31, each missing 17 out of the 35 properties. Using these beers in analysis and models, after filling in their missing values, could potentially dilute trends. Removing these beers from the training data should be explored, in an effort to improve the model.

I replaced missing values with an average of their neighboring beer's respective values. This neighbor was calculated using the euclidian distance from the beer with the missing value to every other beer, in a normalized matrix. Normalizing the data prevents properties with small ranges, most chemical volatile properties have a range in hundredths of parts per billion, from being dominated by properties with large ranges, such as formazin turbidity which has a range of 130 units.

## Analysis

With missing values accounted for I was able to perform the analysis necessary for building intelligent models. This meant recalculating the correlation coefficients, from which I will build my linear regression model, and calculating the principle component (PC) variance, from which I will build my PC model.

Replacing missing values had a greater effect on the certainty of correlations than the strengths of correlations themselves. While this change was slight, it had an impact on the ordering of the correlations, from largest to smallest. Formerly the fifth strongest correlation, Ethyl Octanoate, was demoted to the sixth position with N-Propanol taking its place.

The variance explained within a PC model impacts the number of principle components to be used in said model. See attached `pareto` plot for explained variance as a function of principle components. This plot shows that roughly 80% of the variance is explained by the first three principle components.

## Models

Throughout this project I built hundreds of models to predict the mean consumer preference based on a beer's chemical properties. Aside from simply guessing, these models fell into two categories: linear regression and principle component analysis.

**Linear regression** is a way to develop a best fit solution to an overconstrained set of equations. Performing least squares regression, a type of regression that minimizes the sum of the error from the prediction to the actual value, is such a common problem in Matlab that its function has been assigned to the backslash operator. Calculating a least squares regression model of the first property of our beer data against the mean consumer preference rating is as simple as:

```
a = X[ones(rows,1) beerData(:,property)]\beerData(:,end);
```

Where property could be a single property column of the beer or an array of column indices. Adding properties to the model increases its fit to the training data while increasing its complexity. One must find a balance between the number of variables used in regression and the quality of fit. Fortunately a few statistical criterion have been established to alleviate this dilemma. The Bayesian and Akaike information criterion are two of the most well known tools for model selection. I calculated these statistics as a function of the number of parameters used in my linear regression model, iteratively adding parameters according to their correlation with the mean preference rating. The parameters were added as follows:

```
Column, Correlation, PValue - Description
```

```
11, -0.2935, 0.0047 - BU: Bitterness units.
27,  0.2770, 0.0079 - 2-Methyl-1-Butanol
15, -0.2380, 0.0231 - FTU: Formazin turbidity units.
14, -0.2090, 0.0468 - C02: Carbon dioxide.
20, -0.2008, 0.0563 - N-Propanol
33,  0.1952, 0.0637 - Ethyl Octanoate
35,  0.1878, 0.0747 - Ethyl Decanoate
21,  0.1627, 0.1234 - Ethyl Acetate
```

Metrics of each model were recorded upon each iteration. MaxError is the largest difference between the predicted mean preference rating and the actual mean preference rating as given in the training data. The other metrics are self explanatory:

```
MaxError:
0.970, 0.805, 0.759, 0.764, 0.764, 0.706, 0.718, 0.728
Residual sum of squares:
9.366, 8.532, 8.268, 7.958, 7.957, 7.591, 7.388, 7.383
Bayesian information criterion:
6.748, 11.16, 15.64, 20.11, 24.62, 29.09, 33.57, 38.08
Akaike information criterion:
49.39, 48.15, 49.10, 49.85, 51.84, 52.34, 53.50, 55.48
```

While the Bayesian information criterion (BIC) suggests I build my linear regression model on only one beer property, the bitterness units, BIC is well known for over penalizing the number of variables used in a model. The Akaike information criterion suggesting I use a model with two beer properties, bitterness units and the amount of 2-Methyl-1-Butanol, and I agree. Adding additional properties to this model would improve its fit to the training data but it may not be improving the model's ability to predict mean preferences of other data. As such, my linear regression model is based upon the two properties most strongly correlated to mean consumer preference rating, which lie in column 11 and 27:

```
y = 5.9215 - 0.0496*C11 + 0.0367*C27
```

This model predicts the following mean consumer preference ratings for the beers in file 2.23.txt:

```
y = 5.8046; 5.8486; 5.9315; 5.8272
```

**Principle component** analysis (PCA) is another technique for building predictive models based upon multivariable data. PCA is best defined as "an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest

variance by any projection of the data comes to lie on the first coordinate." [1] The Matlab function `princomp` is helpful for PCA as it automatically computes the principle component coefficients, the scores (the data translated to the new coordinate system), and the variance of each principle component. I recorded the accuracy of each PC model as a function of the number of principle components used. While there is a negative linear relationship between the number of PC used and the RSS, the relationship is not nearly as strong as that of linear regression.

```
MaxError:
0.8696, 0.8092, 0.8290, 0.8070, 0.8295
Residual sum of squares:
9.7243, 9.6131, 9.5824, 9.4649, 9.3630
```

The first two principle components, which account for 70.8% of the data's variance produce a sloppy model compared to that of linear regression. The predictions for that model are as follows:

```
y = 5.7967; 5.7546; 5.7874; 5.8216
```

## Conclusion

Consumer preferences can be modeled with some degree of accuracy. While there were a few beers that debased the maximum error and RSS of my models, beer number 38 was an outlier with an error of 16.4%, the average prediction was off by 4.2%. On the mean consumer preference scale that represents an inaccuracy of 0.2421 units. These statistics are the result of my linear regression model:

```
y = 5.9215 - 0.0496*C11 + 0.0367*C27
```

My predictions for the mean preference rating based on the beers in file 2.23.txt are:

```
y = 5.8046; 5.8486; 5.9315; 5.8272
```

## Discussion

Were there time to delve deeper into this project I would recalculate my predictions with without beers number 28 and 31. These are the beers that were missing 17 properties each. It would also be great to build another model based on a different technique, perhaps clustering, but I feel that that model would only reiterate the limits of the correlations that exist in the data.

---

[1] http://en.wikipedia.org/wiki/Principal_components_analysis