2IV35 Visualization Set 2

Jeroen van Oorschot Student number: 0721913
j.v.oorschot@student.tue.nl

Mart Pluijmaekers Student number: 0753117
m.h.l.pluijmaekers@student.tue.nl

February 1, 2015

Contents

1 Dataset 3

2 Analysis of the data 3

3 Visualizations 3

3.1 Scatter plot 3

3.2 Interval tree 3

3.3 Median tree 4

3.4 Average tree 4

4 Results 4

5 Conclusion 4

1 Dataset

Many people in the world take joy in drinking wine although many of them do not have a clue about what they are actually drinking and what makes their wine taste better then others. Many people go to wine-connaisseurs for advice, or enter liquor stores where usually someone has some knowledge about wines, to help give advice on the purchase. However, it has many properties which can be measured other than the year. For example, sweetness, acidity and even chlorides play a part in the taste ánd quality of the wine. Not even the best and most experienced connaisseurs can possibly have tasted and formed an opinion on every wine in existance. Which means that it can be bennificial to have a way to objectively analyze wines to help determine which wines are expected to be good and which are not. Ofcourse the same principle also applies to produces of wine, who can use the conclusions to see whether their new wine stands any change of being succesfull.

2 Analysis of the data

TODO

3 Visualizations

The visualizations listed in this section are implmented to help analyze the dataset

3.1 Scatter plot

Since we are interested in correlation of data, it makes sense to use a scatter plot since it can visualize this correlation. As the dataset already has a quality parameter available, we want to (possibly) correlate all other parameters to the listed quality. Using this analysis it will become possible to see which parameters influence the quality more and which influence it less.

Usefullness of this visualization would be limited if all differences in measurements are very small and therefor the datapoints would lie close to each other. On the other hand, if the differences are big, but no corrolation is found, scatter plots are useless too.

3.2 Interval tree

The interval tree is a specialized visualization specific to this dataset. It allows for specifying of intervals to which all values are subdevided for easy visualization.

The root of the tree has children for all parameters and each of these children have children for every existing quality. When creating the interval tree, input of the user is required in the form of the number of required intervals (n). The program then creates n equal devisions represented as children (leaves). Those children then get colored and scaled. The smallest value for a interval gets colored red, the highest green and all intermittent values on a linear scale from red to yelloy to green.

The size of the children depicts the number of elements which fall into that interval. Whenever a leave is big, more measurements were made which fell into its interval. This also points us to the problem with the interval tree. Whenever values follow a distribution which resembles a normal distribution, the middle value will be very big, since almost all measurements will fall in the middle interval. This problem worsens with a low number of intervals.

3.3 Median tree

Median trees can be used to determine the average quality for a specific parameter. It again takes a parameter but instead of using it to subdevide the total interval into equal subintervals, instead it selects n medians and uses those to subdevide the interval.

3.4 Average tree

The average tree does the same as the interval tree, except it uses the average to devide the total interval then deviding it equally. Therefor we can see what the impact of some value is, since we now get The average tree is very similar to the interval tree, except that it only has two devisions for every quality. The first for the data bigger then the mean, while the latter is for all data smaller then the mean.

This is usefull since when the measurements are very unbalanced, assume one measurement is ten times bigger then all other measurements. If the program has to devide the total interval in 3 sections, we get 1 green leaf, while all other leaves are red and have a size of almost 100%, which obviously does not help.

4 Results

TODO

5 Conclusion

Using our visualizations we see that a top wine usually has a relatively high alcohol and sulfur dioxide concentration while the concentration of chlorides, total and residual sulfur dioxide.