Sarah Krentz
Data Science Semester 3
Assignment 1, due April 11, 2018

Tasks:

- Read WineQuality.pdf (done)
- Use RedWhiteWine.csv that is provided (downloaded)
    - Note: If needed, remove the objective quality attribute (done)
- Build an experiment using Naïve Bayes Classifier (completed code posted at https://github.com/SarahKrentz/Semester-3_Assignment-1)

Determine:

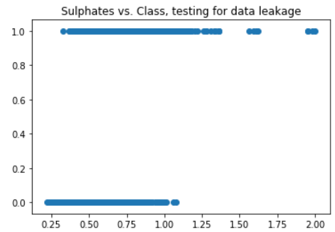
- What is the percentage of correct classification results (using all attributes)?
- What is the percentage of correct classification results (using a subset of the attributes)?
- What is the AUC of your model?
- What is the best AUC that you can achieve?
- Which are the minimum number of attributes that allow you to properly classify 95% of the samples? Why?

The dataset RedWhiteWine.csv contains 6497 rows of measurements of variants of red and white wine. 11 different numeric, physical, attributes (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol) were measured and the final column of data classifies the wine class as Red or White. An experiment was run using a Naïve Bayes Classifier model to attempt to predict the wine class (Red or White) based on these 11 numeric attributes. In each run, the dataset was split into 75% training data and 25% testing.

The model was first run using all 11 attributes, and the most recent run gave an accuracy of 96.91 and mislabeled 50 out of 1619 points. The ROC curve has an excellent AUC of .9634.

Naïve Bayes operates on the assumption that data are normally distributed. For each of the 11 attributes, a histogram was plotted to roughly assess the distribution of the data. From visual estimation, all attributes were normally distributed except residual sugar, total sulfur dioxide, and alcohol. The model was then run again using only the 8 normally-distributed attributes, with the assumption that this would likely improve the accuracy and AUC. Interestingly, this wasn't the case. The most recent run gave a lower accuracy of 94.85 and mislabeled 85 out of 1651 points. The ROC curve has a good, albeit lower, AUC of .9416. This means that at least one of the 3 attributes not used contribute positively to a more accurate wine classification, despite imperfect normal distribution.

Each of the 11 attributes were also plotted against class, to confirm none of the the attributes could be used on their own to determine the wine's class, which would skew the model significantly. The example below is representative of all attributes, none clearly or cleanly clustered into Red and White groups on their own.



To compare, the Naïve Bayes model was run using only sulphates as an input attribute. Predicably, it performed the lowest, with an accuracy of 79.57% and mislabeling 333 out of 1630 points. Its ROC curve also had higher false positive rate at higher thresholds, with an AUC of .6400.

In my attempts, the only run that produced >95% accuracy was using all 11 attributes. This also produced the highest AUC I was able to achieve, .9634. There may be a lower number of attributes that can produce a higher accuracy and AUC, however I was not able to find it by my manual testing. Definitively finding the maximum accuracy would require an attempt with each combination of at least 2 and at most 11 different attributes. This would be ${}^{11}_{2}C + {}^{11}_{3}C + \cdots + {}^{11}_{11}C$ = 2036 different attempts.