ST 557: Applied Multivariate Analysis

Final Project

The files ‘winequality-red.csv’ and ‘winequality-white.csv’ contain data on chemical analysis for 11 attributes and wine expert ratings for 1599 red and 4989 white wines. More information on this data is given in the file ‘winequality-info.txt’.

[*Source:* P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. “Modeling wine preferences by data mining from physicochemical properties.” *Decision Support System>, Elsevier, 47(4):547-553. ISSN: 0167-9236.*]

Analyze the data, with the two goals described below. Write a detailed report (*no more than 5 pages, not including R code*) of the analyses you performed to address each goal, and be sure to include your interpretations and conclusions. ***Include all R code as an appendix at the end of your report.***

Goal 1: Distinguish white wine from red wine

a) Is there a difference in mean vectors between red and white wines for these 11 chemical attributes? Which attributes seem to differ most between red and white wine?

b) Classification: come up with a classification rule based on the 11 chemical attributes to distinguish between the white wines and red wines. Explain your approach, and why you chose it. Evaluate the performance of your classifier: what is the probability that you would correctly classify a new red wine, if it were drawn from the same population of red wines?

c) Clustering: Cluster these wines into two clusters using the 11 chemical attributes. Think carefully about the distance measure and the clustering method you chose, and justify your choices in your report. If you split these data into two clusters, how well do these clusters reflect the red/white classification?

Goal 2: Better understand which of these variables is most important to wine quality. (Focus on red wine)

a) Is there a difference in mean vectors between wines with different quality scores? What if you group the wines into Low (Quality 3 – 4), Medium (Quality 5 – 6), and High (Quality 7 - 8)?

b) Classification/Prediction: come up with a rule for predicting wine quality based on the 11 chemical attributes. You could consider methods like linear regression, nearest neighbors, classification and regression trees, etc.

c) Perform PCA on the 11 chemical attributes, and plot the wines color-coded by quality score. Did you standardize the data before performing PCA? Why or why not? Come up with another rule for predicting wine quality, but now use only the scores for the first two principal components. How does the performance of this new classifier compare to the classifier that used all of the predictor variables? Explain how you compared performance.