University of Birmingham
School of Mathematics
3AS/3AS4: Applied Statistics

Data Analysis Project
2019-2020

Wine industry shows a recent growth spurt as social drinking is on the rise. The price of wine depends on a rather abstract concept of wine appreciation by wine tasters. Pricing of wine depends on such a volatile factor to some extent. Another key factor in wine certification and quality assessment is physicochemical tests which are laboratory-based and takes into account factors like acidity, pH level, presence of sugar and other chemical properties. For the wine market, it would be of interest if human quality of tasting can be related to the chemical properties of wine so that certification and quality assessment and assurance process is more controlled. Two data sets are available from `https://archive.ics.uci.edu/ml/datasets/Wine+Quality` of which one data set is on red wine and have 1599 different varieties and the other is on white wine and has 4898 varieties. All wines are produced in a particular area of Portugal. Data are collected on 11 different properties of the wines based on chemical including density, acidity, alcohol content etc. All chemical properties of wines are continuous variables. The last column is quality, which is an ordinal variable with possible ranking from 1 (worst) to 10 (best). Each variety of wine is tasted by three independent tasters and the final rank assigned is the median rank given by the tasters. Details of the variables involving chemical properties can be obtained from the data website.

In this project, you may consider both red and white wines or only red or only white wines. Main objective is to build a model to predict the quality of the wine based on its physiochemical properties. Some suggested guidelines are as follows:

1. Download the data sets from the website. It is a big data set and so physical checking of errors is nearly impossible.

2. If you are using both types of wines, merge the data sets into one with a column indicating the wine type (red or white).

3. As a preliminary step, check if the properties of these wines differ for red and white wines or for good quality and poor quality wines. You may compare the means using proper tests and/or you may use visualization tools.

4. Create training and test data sets, by making some random partitions.

5. You may use linear regression to predict the quality of the wines. Remember to use model selection techniques to choose the best model and check for the assumptions of the linear regression.

6. You may create a dummy variable indicating good wines and poor wines based on the quality and then use classification techniques to predict the quality of the wines.

In a preprocessing step, you may like to scale all variables. Report how do you choose parameters of your classification technique. What are the training and test error rates?

7. You may also use some unsupervised learning methods like cluster analysis to see if you can cluster the wines based on their chemical properties. Do these clusters correspond with their quality or the wine type?

You need to submit a report not exceeding 10 pages in a single pdf file, which should have the following structure:

- Introduction
- Pre-processing
- Data Analysis
- Conclusion
- References

All suggestions for data analysis are suggestions only. You are free to use whatever you like to explore the main problem. Extra credits will be given for comparing more than one method. All results should be presented in proper tables or plots. Copy/pasting R output is considered as a poor presentation. Proper references should be cited in the text. You don't need to provide any R code.

Marking of the projects will be based on the following:

- **Presentation**. Presentation of the report should be in a technical writing form. You do not need to include mathematical details, but the methods used should be properly mentioned. You must also have a proper structure of a presentation. Tables and plots should be as scientific looking as possible. For examples, look into research papers on data analysis.
- **Data Analysis.** Correctness and logical development of the analysis is the most important part of the project. Innovative approaches will get extra credit. Everything should be clearly documented so that one can reproduce the analysis.
- **Conclusion.** You must have concluding remarks about your project, stating what is the final model, whether there are any limitations to your analysis, any further analysis, difficulties faced etc.
- **References.** There must be some proper references on the methods, data source, or any part of your project you think it is necessary.

All projects need to be submitted in Canvas and they will be checked for **plagiarism**. Suspected cases will be investigated further and may lead to serious penalties including a fail in the whole module.

Please feel free to consult me for any aspect of the project.