Classifying Red and White Wine Quality based on Physicochemical Features

Project Update

Hien Le, Mahdiazhiari Austian, Vera Grosskop

Supervisor: Dr. Gosia Migut

So far, what we have achieved with the datasets from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/wine+quality) are:

- Load dataset: this includes transforming the data file type to numpy array, then splitting each dataset - white and red wine - into features (11) and labels (score from 0 to 10).
- Once this was done, we visualised the data with a histogram, and found out that the data is very skewed: most of the instances have label 5, 6, and 7, and there is a lack of instances for labels 0 to 4 and 8 to 10. This led us to relabel the targets, such that scores lower that 6 would be classified as “bad” wine and are expressed by 0s, while those higher than 6 would be “good” wine and expressed by 1s.
- We followed our original intention to classify using Logistic Regression, Random Forest, Decision Tree and SVM, and used the sklearn libraries for these. For each of these classifiers, 2 or 3 parameters will be tuned e.g tune regularisation parameter for Logistic Regression, tune ‘criterion’ for Random Forest.
- We obtained cross-validation (10-fold) scores of on average over 0.75 over all classifiers (Random Forest, Decision Tree and SVM).

- For Random Forest, the customised classifier (with ‘criterion’ and ‘max_features’ tuned) reduced a relatively high score on the prediction sets (average 0.80)
- For the SVM, it is found that the best kernel for this dataset is the RBF kernel (Radial Basis Function)

*What still needs to be done (quite a lot):

- Logistic Regression
- Tune more parameters and compete with other classifiers; beware of overfitting and underfitting
- Draw ROC curve and get AUC values
- Compare classifiers and compare these performances with the one in the article
- Final Report