

# CITS3401 Data Exploration and Mining Project 2

Wine Classification

Mitchell Pomery  
21130887

June 2, 2014

# Introduction

The project specified that we are to develop several classifiers for wines using different classification methods to compare how machine learning performs compared to experts when rating different wines. The initial data is split into two groups, red wine and white wine, available from the UCI Machine Learning Repository[1]. Data analysis was done using Weka[2], data mining software created by Machine Learning Group at the University of Waikato, New Zealand. Specifically, we are using the classification and clustering tools in Weka Explorer for analysis.

## Data Preprocessing

The initial data provided was in two files, `winequality-red.csv` and `winequality-white.csv`, that were converted to Weka's ARFF file format using an online conversion tool[3]. This tool was used to output two datasets, dataset 1 (`ds1-red.arff` and `ds1-white.arff`) and dataset 2 (`ds2-red.arff` and `ds2-white.arff`). Dataset 1 contains all the information that was in the original data, and is used to create the classifier. The fields in this dataset are numeric, apart from the quality which is nominal, making it is possible to group wines that receive the same rankings in Weka. Dataset 2 contains all the numerical information from the original data and does not contain any information about the rankings from the wine tasters. The aim is to cluster these so that the wines fall into groups similar to the quality attribute of dataset 1.

## Clustering

Dataset 2 requires clustering before it can be classified as the quality attribute of each data point has been removed. Simple K means clustering was used and after experimenting, fixed acidity, volatile acidity, citric acid, and density were ignored for the red wine data. This gave a roughly similar distribution to the qualities found in the red wine dataset. In the white wine dataset, ignoring the different attributes has very little effect on the clustering. After ignoring different combinations of attributes, it does not seem possible to cluster the data points into a similar distribution to the initial data. Comparing the clustering with the red wines quality information shows a 35% accuracy in the groupings, while for the clustered white wines there is only 25% accuracy.

### Red Wine

```
Scheme:weka.clusterers.SimpleKMeans -N 6 -A  
"weka.core.EuclideanDistance -R first-last" -I 500 -S 10
```

#### Clustered Instances

0	632 ( 40%)
1	128 ( 8%)
2	196 ( 12%)
3	32 ( 2%)

4	339 ( 21%)
5	272 ( 17%)

White Wine

```
Scheme:weka.clusterers.SimpleKMeans -N 7 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Clustered Instances

0      1093 ( 22%)
1       536 ( 11%)
2       832 ( 17%)
3       569 ( 12%)
4       885 ( 18%)
5       658 ( 13%)
6       325 (  7%)
```

## Classification

### Naive Bayesian

Naive Bayes classifiers are simple to implement, fast, and are used in real world situations such as spam filters. They work by looking at the traits of an object, and using each individual trait to determine how likely it is that the object falls into a specific classification. Their downside is that they assume the presence or absence of particular traits has no affect on the classification.

In Weka, we used

#### Dataset 1

#### Dataset 2

### Support Vector Machine

Support Vector Machine's are learning models and algorithms that analyze data and find patterns, then use the patterns for classification of data. Unlike the Naive Bayesian classifier, the support vector machine is a non-probabilistic classifier, meaning that it will not provide uncertainty for the results. This means that each different category is separated by as large a gap as possible.

#### Dataset 1

## **Dataset 2**

## **Neural Network**

Neural Networks are based off of animal's central nervous systems, by using several input sensors that transform the data before handing it on to another neuron. The neurons are connected together in a network and work simultaneously , rather than sequentially, to process the data. Real world applications for neural networks include speech and handwriting recognition.

## **Dataset 1**

## **Dataset 2**

## **Results**

- Red wine dataset is not very spread out
- White wine distribution was. Looks like a bell curve

# Bibliography

- [1] UCI Machine Learning Repository: Wine Quality Data Set. 2014. UCI Machine Learning Repository: Wine Quality Data Set. [ONLINE] Available at: <http://archive.ics.uci.edu/ml/datasets/Wine+Quality>. [Accessed 01 June 2014].
- [2] Weka 3 - Data Mining with Open Source Machine Learning Software in Java . 2014. Weka 3 - Data Mining with Open Source Machine Learning Software in Java . [ONLINE] Available at: <http://www.cs.waikato.ac.nz/ml/weka/>. [Accessed 01 June 2014].
- [3] Online CSV to ARFF conversion tool. 2014. Online CSV to ARFF conversion tool. [ONLINE] Available at: <http://slavnik.fe.uni-lj.si/markot/csv2arff/csv2arff.php>. [Accessed 01 June 2014].
- [4] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.
- [5] Rennie, J.; Shih, L.; Teevan, J.; Karger, D. (2003). "Tackling the poor assumptions of Naive Bayes classifiers". ICML.
- [6] Non-Probabilistic Classification Methods. 2014. Non-Probabilistic Classification Methods. [ONLINE] Available at: [http://www.dcs.gla.ac.uk/~girolami/Machine\\_Learning\\_Module\\_2006/week\\_5/Lectures/wk\\_5.pdf](http://www.dcs.gla.ac.uk/~girolami/Machine_Learning_Module_2006/week_5/Lectures/wk_5.pdf). [Accessed 01 June 2014].