Wine classification

Aleksandar Savić, Filip Savić, Drago Stevanović

1. Motivation

Our motivation comes from the two of our team members who are wine lowers. Therefore, we would like to create a classifier which would help determine the quality of a certain wine. This would help us (and other people) learn a bit more about wines and what affects their quality the most.

Also, our classifier could help other people who are not that eager to learn about wines to find a quality wine for themselves.

Finally, our classifier could help new sommeliers who have just started training their taste buds to quickly test their assumptions and estimations about a certain wine's quality.

2. Research questions

We want to classify a certain wine into a quality category (from 0 to 10, 10 being the best quality) based on some of it's given properties, such as fixed and volatile acidity, citric acid and alcohol percentage.

The data set contains information about 12 properties of 1600 Portuguese "Vinho Verde" red wines. Properties are:

- fixed acidity - most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
- volatile acidity - the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
- citric acid - found in small quantities, citric acid can add 'freshness' and flavor to wines
- residual sugar - the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
- chlorides - the amount of salt in the wine
- free sulfur dioxide - the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
- total sulfur dioxide - amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
- density - the density of water is close to that of water depending on the percent alcohol and sugar content
- pH - describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
- sulphates - a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant
- alcohol - the percent alcohol content of the wine
- quality - output variable (based on sensory data, score between 0 and 10)

The data set is clear (i.e. it contains no null values). Also, it is misbalanced as it contains mostly medium quality wines (with quality grades 5 and 6). Hence, we will use the f1 score measurement.

3. Related work

Since the dataset is relatively new, there are only a few of the solutions available. We found most of them incorrect because they made predictions based on "quality" column. They made additional features such as 'category = {"bad", "good"}' from quality column and then split the dataset into training and test. Also, another big issue was that classifiers were tested on the same data which they were trained on.

4. Methodology

First, we divided the original dataset into training and test with 75:25 (training : test) ratio. Then, we wanted to do some data preprocessing, i.e. feature selection and normalization. We tried using the PCA (Principal Component Analysis) which helps in finding the most important variances in features and taking those variances into principal components. When we ran it on our data, we got the scree plot which can be seen on Figure 1:

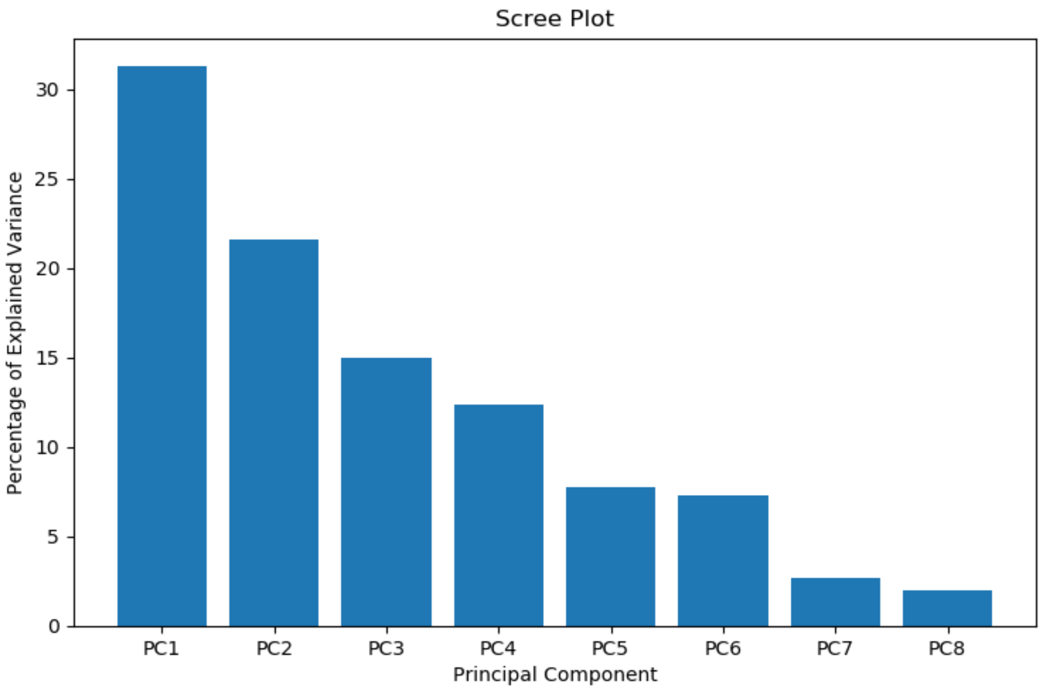

Figure 1 – PCA Scree Plot

On the Figure 1 we see that our variance is well distributed, i.e. it is not good enough to use only 2 or 3 principal components to truly describe the variance in the original data set.

After PCA we did feature selection. We checked the correlation matrix and saw that the PH, residual sugar and free Sulphur dioxide did not have much impact on the quality of the wine. We removed them from the original 11 features (plus the ‘quality’ feature) and were left with 8 features. We finished the data preprocessing with the normalization of the data.

We divided the problem of classifying the wines into three subcategories:

- low quality – wines with the score of 3 and 4
- mid quality – wines with the score of 5 and 6
- high quality – wines with the score of 7 and 8

Then, we created low, mid and high classifiers and predictors. Classifiers are responsible for deciding if a certain wine is in the classifier's category. For example: low classifier will give a True for the wines with the quality of 3 and 4, and False for the wines with quality of 5 or more. Predictors are responsible for determining which exact quality grade does the wine have. For example: low predictor will take a low quality wine and give it a quality grade of 3 or 4. Classifiers were trained on the entire training set while predictors were trained on the corresponding subset of the training data set. As a final classifier we used a modified stacking. Our main classifier (decider) receives the values from classifiers and predictors and features from the original dataset.

5. Discussion

Testing was done using cross-validation, splitting the training data into (usually) 5 folds. As evaluation measure, we used the f1 micro-score. By analyzing the errors, we saw that chosen metric is not really the best. The problem was that most of the data samples falled into the middle categories (the quality grades 5 and 6). That's why most the first and third classifier were "thaught what doesn't fall into their cathegory", which resulted in having lots of falses. Since there really is a lot of true negatives for the first and third classifier, cross-validation gave pretty high scores (0.8 at least, 0.92 was the average).

We tried writing our own metrics, which turned out to be much more complicated than we thought. At the end we decided to keep the f1 micro as our evaluation score.

Our final score using f1 micro-score was 0.6825.