Exploratory Data Analysis

COSC5557 – Practical Machine Learning

William Baumchen

11/27/2023

1 - Introduction

The behavior and quality of machine learning models is dependent on the data being used to train those models. The type of model and the hyperparameters chosen are also important, but without useful data inputs the model will not be useful. As a result, data preprocessing is a critical step in machine learning. Exploring the characteristics of the given data set and finding those qualities that need to be modified in some way can massively impact the performance of the resulting machine learning model. In this exercise the white wine quality data set is examined in a number of ways.

2 – Analysis Results

2.1 – Dataset

The dataset used is the white wine quality dataset, containing 4,898 observations, with 11 features, all of which are real numbers taken from a continuous range. There is one 'target' feature, with seven possible classes. There are no missing values in the observations and target feature.

2.2 – Feature Correlation & Principal Component Analysis

Feature correlation is an important quality of a dataset that can have potentially negative effects on some types of supervised learning. For example, some types of linear machine learning models are constructed with the expectation that the dataset used to train those models does not exhibit strong feature correlation. If such is present, training will result in wildly inaccurate models. In addition, through principal component analysis and other similar methods, the removal of heavily correlated features from the model will reduce the effective dimensionality of the dataset, increasing the speed and efficiency of the machine learning model. An effective way of visualizing feature correlation is through a heatmap of the correlation coefficients, as seen in Fig. 1 below. In this exercise, the correlation between features was found using the corrcoef function in MATLAB. As can be seen, some of the variables are relatively heavily correlated with one another, for example the free and total sulfur dioxide features, and while less heavily correlated, the chlorides and density features.

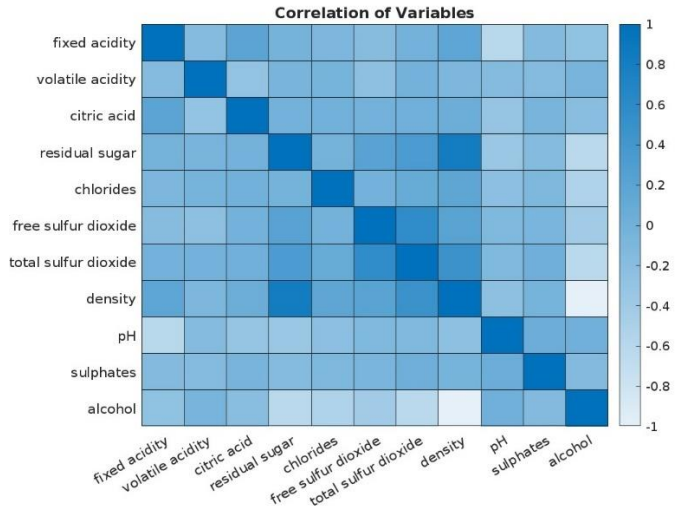


Figure 1 – Variable Correlation Heatmap

Conducting Principal Component Analysis (PCA) on a dataset results in a transformed set of observations, with features that now correspond in decreasing effective variation in the target feature space of the model. This, more properly defined, is given as an orthogonal linear transformation that transforms the given data to a new coordinate system, such that the greatest variance by some scalar projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on [1].

Here, PCA was conducted using the pca function in MATLAB [2]. It was found that of the resulting eleven features, the first and primary was responsible for 90.97% of variation in the target space, the second responsible for 7.93%, and the third responsible for 1.02%. The rest are shown in Table 2.1. As can be seen, to capture around 98.9% of the variation in the feature only the first two transformed features are necessary, allowing for a large increase in efficiency of future optimization and learning.

Table 2.1 – Component Variation from PCA

Feature	% Variation
1	90.965734397450902
2	7.933386311638994
3	1.015427419571099
4	0.050600445035768
5	0.032340939450079
6	0.000872769740299
7	0.000672986618088
8	0.000539060917788
9	0.000407002123016
10	0.000018652532245
11	0.000000014921728

2.3 – Normalization

Normalization in data pre-processing is a term with a number of definitions. Here, the process refers to scaling the data such that it has a mean of zero and standard deviation of one, while retaining the shape properties of the original data set. Normalization is especially important in some machine learning models, as some models are constructed with the assumption that the data used for training is drawn from a uniform distribution with the mean of zero and standard deviation of one. In this exercise, the normalization of the white wine quality dataset was completed using the normalize function in MATLAB. As can be seen below, Figs. 2-3 show the histograms of the data before and after normalization. The shape of the distribution of the data does not change, but the distributions are scaled. And, in Fig. 4, box plots of the un-normalized and normalized data are shown. This allows for easy visualization of the transformation results.

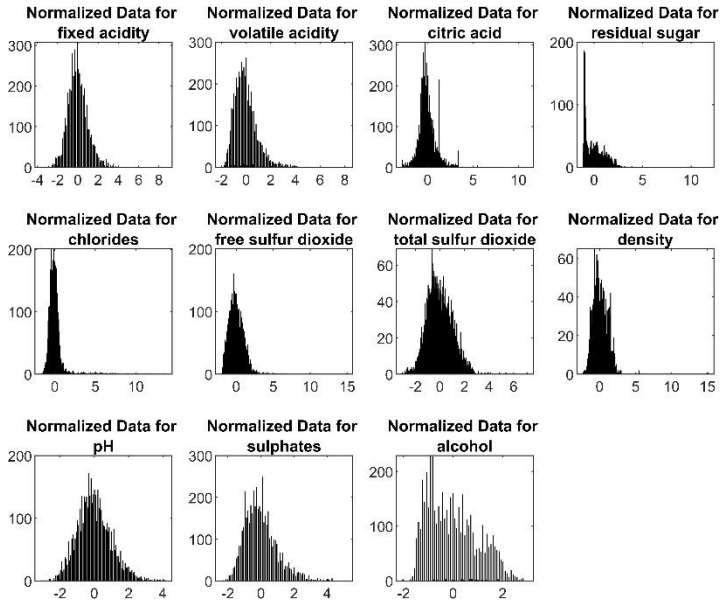


Figure 2 – Normalized Distribution

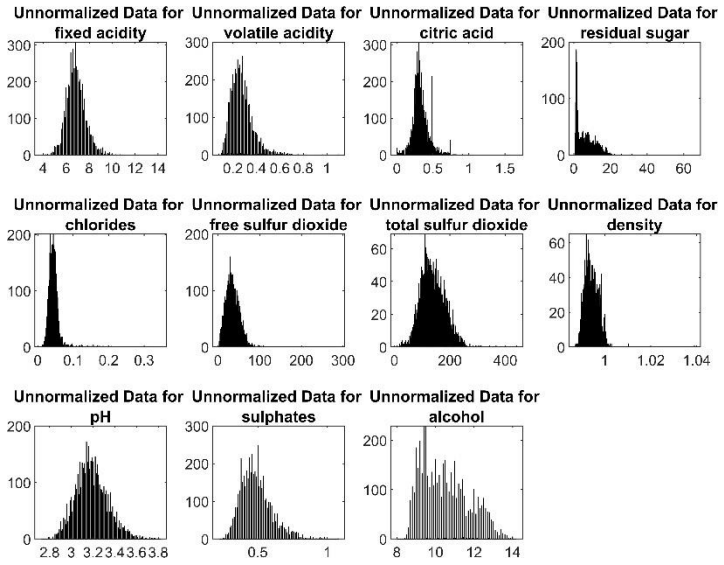
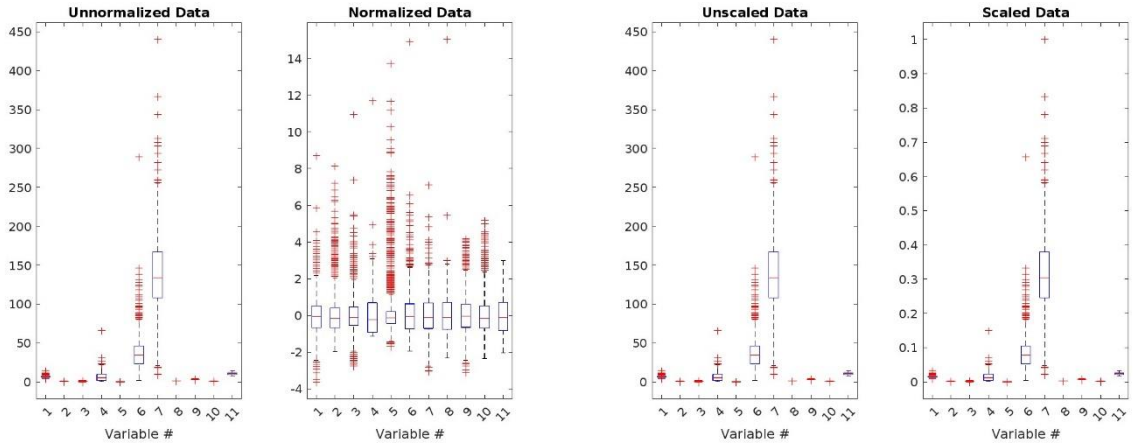


Figure 3 – Unnormalized Distribution

2.4 – Scaling

The process of rescaling in data pre-processing is a phenomenon with a number of labels. Here, the process known as rescaling refers to the simple scaling of data such that each feature has some set range. Common ranges include [0,1] and [-1,1]. Scaling is especially important in some machine learning models, for example support vector machines. In this exercise, the scaling of the white wine quality dataset was completed using the rescale function in MATLAB. As can be seen below in Figs. 5-6, the shape of the distribution of the data does not change, but the distributions are scaled. In addition, Fig. 7 shows box plots of the un-scaled and scaled data are shown. This allows for easy visualization of the transformation results, where the behavior of the data is the same after the transformation, and only the scale is affected.



Figures 4 & 7 – Boxplots of Normalization and Scaling

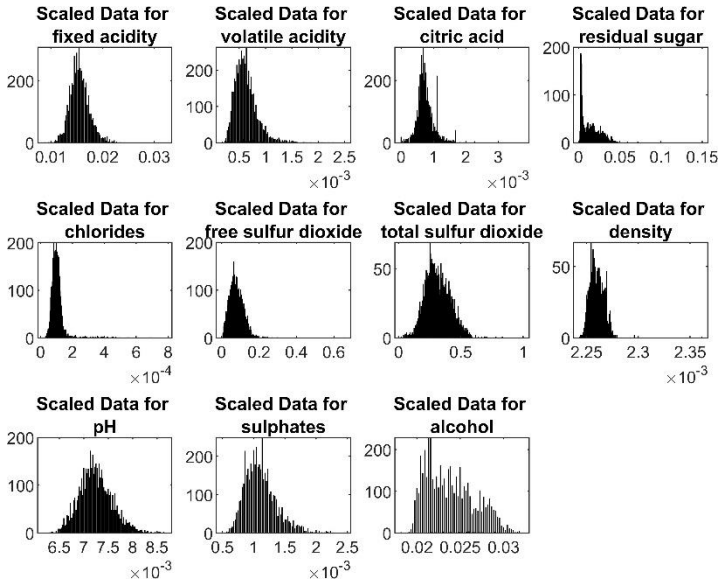


Figure 5 – Scaled Distribution

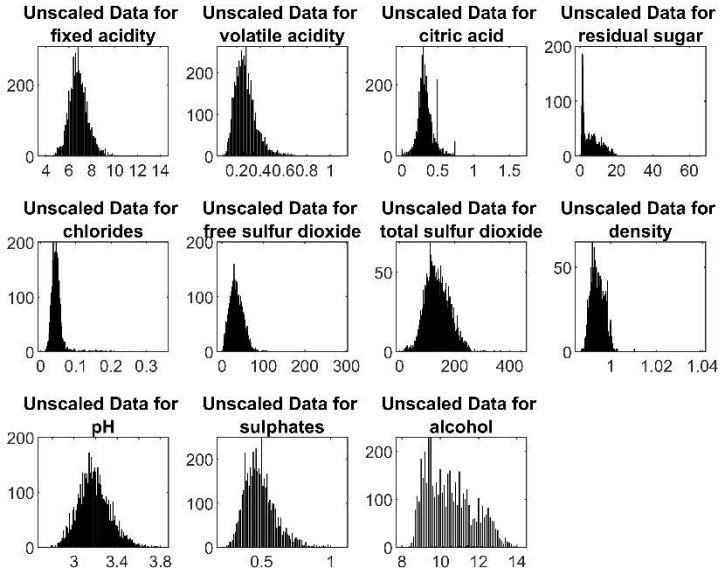


Figure 6 – Unscaled Distribution

References

[1] Jolliffe, I. T. *Principal Component Analysis*. 2nd ed., Springer, 2002.

[2] The MathWorks, Inc. (n.d.). *pca*. Principal component analysis of raw data - MATLAB. https://www.mathworks.com/help/stats/pca.html?s_tid=doc_ta