Data Modelling and Analysis

Lecture 6: Data Analysis and Visualsiation Excercise

Dr Mercedes Torres Torres

Contents

1 Data Set **1**

2 Data Analysis and Visualisation **2**

3 Data Pre-processing **2**

1 Data Set

The data set for this exercise is a slightly modified version of a reference data set available from the *UCI Machine Learning Repository* [1]. The data concerns the modelling of wine quality based on physicochemical tests [2].

Each record consists of over 10 attribute (input) columns, and one class (output) column corresponding to the quality of wine, rated on a ten-point scale. The attributes record various physical and chemical properties of the wine. The entire data set consists of over 1,600 instances.

Here is are the first 10 rows:

Table 1: The first 10 rows of the modified wine dataset

sample	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	residual.alcohol	alcohol	region.of.origin	quality
1	8.5	0.28	0.56	1.8	0.092	35	103	0.9969	3.30	0.75	0.06	10.5	5	7
2	8.1	0.56	0.28	1.7	0.368	16	56	0.9968	3.11	NA	0.03	9.3	4	5
3	7.4	0.59	0.08	4.4	0.086	6	29	NA	3.38	0.50	0.06	9.0	4	4
4	7.9	0.32	0.51	1.8	NA	17	56	0.9969	3.04	1.08	0.06	9.2	3	6
5	8.9	0.22	0.48	1.8	0.077	29	60	0.9968	3.39	0.53	0.09	9.4	3	6
6	7.6	0.39	0.31	2.3	0.082	23	71	0.9982	3.52	0.65	0.07	9.7	3	5
7	7.9	0.43	0.21	1.6	0.106	10	37	0.9966	3.17	0.91	0.09	9.5	1	5
8	8.5	0.49	0.11	2.3	0.084	9	67	0.9968	3.17	0.53	0.02	9.4	1	5
9	6.9	0.40	0.14	2.4	0.085	21	40	0.9968	3.43	NA	0.08	9.7	2	6
10	6.3	0.39	0.16	1.4	0.080	11	23	NA	3.34	0.56	0.09	9.3	4	5

[1] http://archive.ics.uci.edu/ml

[2] http://archive.ics.uci.edu/ml/datasets/Wine+Quality

2 Data Analysis and Visualisation

1. Explore the data

   i. Provide a table for all the appropriate attributes of the dataset including type, measures of centrality, dispersion. Provide evidence of duplicates in the data and how many missing values each attribute has.

   ii. Produce the same table is in the first exercise, but grouping according to quality.

   iii. Produce relevant visualisations to study the distributions of the appropriate attributes within the data. You may also use additional statistics to help you characterise the shape of the distribution.

2. Explore the relationships between the attributes, and between the class and the attributes

   i. Calculate the correlations and produce scatterplots for the following three pairs of variables: free sulfur dioxide and volatile acidity, pH and fixed acidity, and density and fixed acidity (three correlations, three scatterplots). What do these tell you about the relationships between these variables?

   ii. What are the highest correlated attributes? And the lowest correlated attributes?

   iii. Produce boxplots for all of the appropriate attributes in the dataset. Group each variable according to the class attribute.

3 Data Pre-processing

1. Dealing with missing values in R: Replace missing values in the dataset using two strategies: replacement with 0 and median and contrast these approaches and its effects on the data.

2. Attribute transformation: Using the datasets generated previously, explore the use of mean centering and normalisation to scale the attributes. Contrast these approaches and its effects on the data.

3. Attribute / instance selection:

   i. Starting again from the raw data, consider attribute and instance deletion strategies to deal with missing values. Choose a number of missing values per instance or per attribute and delete instances or attributes accordingly. Explain your choices and its effects on the dataset.

   ii. Starting from an appropriate version of the dataset, use Principal Component Analysis to create a data set with 8 attributes. Explain the process and the result obtained.