Introduction to Data Science Using Python (CSE 3054)

MAJOR ASSIGNMENT

1 Objective

The objective of the assignment is to provide programming practice regarding reading and exploring the dataset (using the mathematical and data visualization tools) in python.

2 Problem Statement

Suppose we are given a dataset, described in Section 2.1, about the quality of the wine. It is required to read and perform exploratory analysis of the dataset before it can be used to train a model for predicting the quality of the wine. The name of the dataset file is ”*winequality-white.csv*”. In this regard, perform the following exercises:

- Write a python program to read the dataset – Use “**with**” python module to open and “**csv.reader**” to read the delimited file.
- Perform following exploratory analysis about the dataset:
  1. Group each attributes into 30 discrete buckets and plot the histogram.
     (Hint: For creating the required numbers of buckets use following computation to determine the bucket size:)
     $$bucket\_size = \frac{max(attr) - min(attr)}{30} \tag{1}$$
     Use this bucket-size to bucketize each attribute’s values.
  2. Write a python program to compute and print the mean, median, mode and variance for each attribute.
  3. Write a python program to compute the “***covariance***” for each pair of (attribute, output label). For example; between the pairs (*fixed acidity*, *quality*), (*volatile acidity*, *quality*), and so on. Name the attributes sharing same directional relationship with the output label.
  4. Construct the “***correlation matrix***” for the dataset as follows:
     (a) Case I: Based on correlation between the pair of attributes. For example; between the pairs (*fixed acidity*, *volatile acidity*), (*fixed acidity*, *citric acid*), and so on.
     (b) Case II: Based on correlation between the pair (attribute, output label). For example; between the pairs (*fixed acidity*, *quality*), (*volatile acidity*, *quality*), and so on.
  5. Answer following based on “***correlation matrix***” obtained from previous steps:
     (a) From Case I, name the two attributes sharing maximum similarity and dis-similarity.
     (b) From Case II, name the attribute sharing maximum similarity and dis-similarity with the output label.
  6. Construct the “***scatter-plot matrix***” to show between the attribute’s relationships.
     (Hint: Use **plot.subplots** for plotting the scatter-plots in same figure.)

2.1 Dataset description

The dataset contains 11 attributes and one output label. The attributes indicate the value for different physicochemical factors, while the output label indicate the quality on the scale of 0 to 10. 0 denote the worst quality and 10 denote the best quality

Input Attribute Information:

- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol

Output variable (based on sensory data): *quality* (score between 0 and 10)

3 Mark Distribution

- Read dataset – [2 marks]
- Create histogram – [2 marks]
- Finding central tendencies and dispersion – [2 marks]
- Computing covariance – [2 marks]
- Constructing correlation-matrix – [2 marks]
- Solution to questions based on the correlation-matrix – [1 marks]
- Constructing scatter-plot matrix – [2 marks]