1/22/19

CSE 5243
Instructor: Jason Van Hulse
Homework 2

Due Date: 2/6/2019 11:59 pm (submitted through Carmen) *[there will be a 15% penalty for all homework handed in up to 24 hours late; homework will not be accepted more than 24 hours late]*

Objective: In this lab, you will work with the **Wine_quality** dataset. You should work through each of the sections listed below and answer the questions or perform the analysis as specified. The answers to these questions should be provided either in a report format or within a Jupyter Notebook. Your programming and analysis work should be done in R or Python. Unless specifically stated, you may use off-the-shelf packages.

Dataset Details: The wine_quality dataset can be found at the following URL:
https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv

This homework uses the *winequality-red.csv* dataset.

Data Preprocessing Tasks: Before you do any of the analysis below, execute the following tasks:

1. Create a binary class variable according to the following rule:
   quality <= 5 —> class = “Low”, quality > 5 —> class = “High”
2. You will need to remove the *quality* attribute after you create the class label. I will refer to this new variable as **class**.

Part 1: Business Understanding (5%): Similar to what you did in homework #1, in this section, you should create a section of your report that provides a brief overview of the data. Some questions you should consider are: Where did the data come from? What do the rows represent? Why and how was the data collected? What types of questions might you be able to analyze with this data?

Part 2: Data Understanding (25%): Perform exploratory data analysis of the dataset by looking at individual attributes and/or combinations of attributes. You should focus on identifying and describing interesting observations and insights that you might uncover from the data. Use the appropriate data visualization and statistical methods to analyze the data.

In addition, your report should answer the following questions (for each question, be sure to justify your response with data, visualizations or statistics as well as an explanation):

1) Are there any outliers in the data?
2) Are any of the independent variables redundant relative to each other?

3) Which attribute has the *strongest* relationship to the class label?
4) Which attribute has the *weakest* relationship to the class label?

Part 3: Data Transformations (20%): In this section, you will perform a number of attribute transformations and evaluate the impact of these transformations.

1) Perform *equal interval width discretization* (into 10 bins) for the attribute *total.sulfur.dioxide.* What do you notice about the result, and do you think this discretization is effective?
2) Perform *equal frequency discretization* (into 10 bins) for the attribute *total.sulfur.dioxide.* What do you notice, and how does this compare to the equal interval width discretization?
3) Perform a log transformation of the attribute *alcohol*. Compare the original variable with the transformation; what observations do you have?
4) Choose one *supervised discretization method* (e.g., the method makes use of the class label to more intelligently decide how to discretize a continuous attribute) to apply to the data. You can use an off-the-shelf R or Python package to do this (again, make sure that the functions you choose are appropriate for the given data), or you can code your own custom function. Describe the package that you used and any settings that you chose. Apply this discretization method to the *total.sulfur.dioxide* attribute, and compare this discretization to the equal interval width and equal frequency discretization methods performed above.
5) Perform both the *standardization* and *normalization* transformations for the attribute *total.sulfur.dioxide* as well as one other attribute of your choice. Analyze the outcome of these transformations.

Part 4: Principal Component Analysis (20%)

Using the 11 remaining continuous features (and excluding the class), use an off-the-shelf package (in Python or R) to perform a Principal Component Analysis of the data.

1) Briefly discuss and explain the package that you used and the parameters you set (if any).
2) Briefly present the results of the analysis and discuss your observations.
3) Provide the a table with the eigenvalues and the matrix U (with the eigenvectors).
4) How many principle components are needed to capture 95% of the variation in the original data?
5) Plot the scatterplot for the transformed data in two dimensions.

In addition to working through the analysis in each section outlined above, your **30% of your grade will be based on the overall readability and organization of your report.** Is your report well organized and does the presentation flow in a logical manner? Are there grammar and spelling mistakes? Do the charts/graphs relate to the text?

Collaboration: For this assignment, you should work as an individual. You may informally discuss ideas with classmates, but your work should be your own.

What you need to turn in to Carmen:

1) Program (if applicable)
2) Report

How to hand in your work:

Please do this work in either Python or R. All the related files (code and/or report) except for the data will be tarred in a *.zip file or *.tgz file, and submitted via Carmen. Use this naming convention: “Project2_Surnames_DotNumber.zip" or “Project2_Surnames_DotNumber.tgz.” The submitted file should be less than 5MB.