C S 488/508 Introduction to Data Mining

Homework 2: Data exploration

1 Objective

This homework requires you to explore datasets to understand their basic statistics. This is an *individual* homework.

2 Requirements

Use a data set called Breast Cancer Coimbra from the UCI machine learning repository. Its csv file can be downloaded from Canvas (BreastCancerCoimbra.csv). Read the data set description at `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra#` to know about its attribute information and write programs to perform the following tasks.

(1) (20 points) For the attribute BMI, calculate its mean, standard deviation, minimum, and maximum values.

(2) (20 points) Compute Pearson's correlation between pairs of attributes.

(3) (20 points) Display the histogram for each of the quantitative attributes by discretizing it into 10 separate bins and counting the frequency for each bin.

(4) (20 points) Display a boxplot to show the distribution of values for each attribute. Which attribute has outliers?

(5) (20 points) Consider the first four attributes: Age, BMI, Glucose, Insulin. For each pair of those four attributes, display a scatter plot. Based on the scatter plot, what are possible correlations that you can observe?

(6) (20 points, **CS 508 only**) Use parallel coordinates to visualize the dataset. The visualization should have a legend that shows different labels. Please explain what you get is useful for any analysis. You may want to consider normalize the values of the columns and compare the different parallel coordinates plot.

You can write Python or Java code. Please put the code for these question to one file. Please properly organize the code to make grading easy.

3 Submission instructions

A zipped file `hw-lastname.zip` consisting of all the code.

4 Grading criteria

(1) CS 508 students need to answer all the questions. Your scores will be scaled to 100.

(2) CS 488 students do not need to answer questions marked with *(CS 508 only)* although you have the freedom to work on them. Your scores will be scaled to 100. If CS 488 students answer the questions marked with *(CS 508 only)* , you will not have any points deducted if your answers are wrong; you will not get any extra points either if your answers are correct.

(3) The score allocation has already been put beside the questions.

(4) Please make sure that you test your code **thoroughly**.
FIVE points will be deducted if files are not submitted in the required format.