Homework 2 (25 points) Due on Friday, Feb 7, 2025

The datasets are available at https://drive.google.com/drive/folders/1R3aC0u9LCqNbwU6F9aqlTtI3kQXqgWVX?usp=drive_link .

1. (3 points) Not only read through, but also hands-on to work through the examples of chapter 6 of Geron book (page 175-187), and explore the iris dataset iris.csv and answer the following questions
   a. What are the names of the variables, which of them are nominal, and numerical data?
   b. Display the histograms of each numerical data type.
   c. Use the box diagram to display the summative statistics of each of the numerical data.
   d. How would you like to arrange the boxplots above so that you can compare the distributions of the data effectively? (hint, check how Weka arrange the comparison)
   e. Change the hyperparameter setting such as max_depth = 3 or min_samples_leaf = 50, print the tree in your answer sheet by using graphviz to transform the iris_tree.dot into iris_tree.png.

2. (2 points) Use Python matplotlib to exploring the dataset myFirstData.cvs
   a) Load data to your workspace. Note, you first need to specify your path. Please ask help either from me or the TA, or friends to set the path of the workspace. You can drag the file inside the quote marks in Windows, instead of typing the path/filename.
      For example, in R, usedcars <- read.csv("C:/*yourpath* …/myFirstData.csv")
      In MATLAB, >> mydata = importdata('*your_path*/myFirstData.cvs')
   b) Plot the histograms for both column of data,
   c) Use boxplot to display the summative statistics for each column of data.
   d) Plot the scatter point and regress line of the two columns of data.

3. (3 points) We will use the twomillion.csv
   a) use read_csv() to read the 10% of your data into the workspace
   b) For your sample, use the functions mean(), max(), var() and quantile(,.25) to compute the mean, maximum, variance and 1st quartile respectively. Show your codes and the resulting values.
   c) Compute the same quantities in part b on the entire data set and show your answers. How much do they differ from your answers in part b?
   d) Then open this file with Excel and compute the mean, maximum, variance, and 1st quartile. Provide the values and name the Excel functions you used to compute these.
   e) Exactly what happens if you try to open the full data set with Excel?

4. (4 points) Please use one to two sentences of your own words to answer the questions.
   (a) What is the pros and cons of the decision tree and Rule method based on Ripper Algorithm for an imbalanced dataset?
   (b) What is overfitting problem? Please give a definition in terms of resubstitution testing and generalization testing.
   (c) Use your own words to describe minimal description length (MDL) principle.
   (d) What the relationship between MDL Principle and overfitting problem of decision tree?
   (e) Among the three metaparameters, the number of instances, the number of classes, and the number of attributes, which one affect the complexity (i.e. the training time) of a decision tree most?
   (f) Assume that you need to take 1 hour to train a decision tree by using a reduced dataset of one million instances and 100 data attributes, how much time you have to spend to train the decision tree by using the original dataset with 10 million of instances and 200 data attributes?
5. (2 points) #3, p 186 - the 2nd edition of the book of Tan et. al (Page 198-199 of the 1st edition)
6. (2 points) #5, p 187 – the 2nd edition of the book of Tan et. al (Page 200 of the 1st edition),
7. (2 points) # 9 p 188 - the 2nd edition of the book of Tan et. al (#8 Page 202 of the 1st edition)
8. (2 points) # 10 p 189 - the 2nd edition of the book of Tan et. al (# 9 Page 202 of the 1st edition)
9. (3 points) Page 186, # 7 of Geron's book in Chapter 6
10. (2 points) Page (187) # 8 of Geron's book in Chapter 6.