Homework Assignment 1, January 22, 2025

Read Chapter 1 (all) and Chapter 2 (only sections 2.1, 2.2, 2.3 and 2.4) of the first textbook. Follow the Anaconda install jupyter, numpy, pandas, and matplotlib, and the install (use github) the SciKit-Learn codes of the second textbook https://github.com/ageron/handson-ml2 .

If you do not have the textbook, you can access the first edition at the linked shared folder at https://drive.google.com/open?id=1N_KKjNcFZae4arEtXfAxnXGlJugwwmG8) .

1. (2 points) When dealing with real-world data, you usually find that your data set many missing entries either by the nature of data such as medical data for patients or human errors, device fault. It would help if you had a tidy data frame for any data mining project. Please suggest three methods to deal with missing data entries for the following three types of hypothetical data set characteristics. Remember, the fitness of your suggested methods depends on the characteristics of the dataset. Please rank the methods and justify your idea with pros and cons under your data characteristics.
   (a) You merged two financial data sets, one is a stock trade data set that were traded only on weekdays (5 days maximum), and the other is bitcoin dataset that were traded also in weekends (7 days maximum) at the same year. You need to fill the data features of the stock trade datasets in the weekends so that you can compare the return of investments. Suggest how would you like to fill the missing data of the stock data set in all weekends.
   (b) A very large data set with 500,000 records and 3000 attributes, but most records missed a few data entries spread in different attributes.
   (c) A small data sets with 200 of records and 4 numerical data features of the spatiotemporal data. The dataset missing 10% of data values crossing the 4 different variables. How would you treat the missing values of the data set.

2. (1 points) What is your strategy to use the OpenAI tools such as ChatGPT and CoPilot to improve your learning and working efficiency. Please feel free to discuss this problem with your classmate and provide answer here.

3. (2 points) Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied.

4. (2 points) Problem #18 on Page 109 of the second edition of the textbook (page 92 of first edition,

5. (2 points) Problem 19 on page 109 of the second edition, (page 93 of the first edition)

6. (2 points) Problem 20 on Page 110 of the second edition, (page 93 of the first edition)

7. (2 points) Follow the steps to test your Weka (see weka walk though). If everything works,

report the visualization result in item D. If you have problem to complete the steps, post the problem in the course blogs under the Weka Tutorial page and welcome anybody else to follow your blog and post answers and discuss relevant issues.
A: Install the Weka under your programming directory, copy the data files to a convenient work directory that you can access easily.
B: Launch the GUI and Exploration Manu. Then Open the file, Weather.arff.
C: Follow the tutorials to work through the pipeline
D: Try to visualize some result.

8. (2 points) Install all software above and follow example 1.1 of textbook 2 by Geron from page 21-29, but replace the dataset by using the Housing dataset Housing.csv with 13 variables, (variable name file) find the regression relation between the median value (MEDV, dependent variable) and TAX (independent Variable) and make a similar graph as Figure 1-19 and Figure 1-23 based on the two different categories of houses, CAT. MEDV = 0 and CAT. MEDV = 1.

9. (2 points) Read the Example 2.1 of the textbook 1 of Tan et al, and answer the questions
   (a) Have you found anything wrong in either the statistician or data miner? If yes, who is it, and what is wrong?
   (b) What does the dialog implies about the relationship between Statistics and Data Mining?

10. (3 point) (a) Describe a data mining project that you are interested to work in the last five weeks
    (a) Please list the mathematics and computer science or data science courses you have taken.
    (b) Describe a project (it can be either different or the same depends on the scope of the problem) that you may lead after you complete this course and
    (c) Describe a wish list of expertise of your team other than data mining, e.g. programming skills, mathematics background, domain knowledge in a particular domain, etc. from your teammates.