Queen Mary
University of London

ECS766P DATA MINING

Dr Emmanouil Benetos, Dr Dimitrios Kollias

Assignment 1 (worth 20% of total mark)

DUE DATE: FRIDAY 7 NOVEMBER 2025 (16:00)

A general note:

Most of the below tasks will require code in order to be addressed. Please **do make sure to show your workings** - i.e., how did you derive the result by showing the code that was used to generate the result that addresses the question and by writing down your thinking. The code ***should not*** be a screenshot/image; it should be copied and pasted (e.g. from your notebook) to the pdf/doc/docx that you will submit (see deliverables below). The code should be in a correct & readable format, e.g. with correct indentation for python.

Tasks

In the following, tasks 2, 6 & 8 are only pen-and-paper exercises and do not require code.

1) The function "*load_wine*" from "*sklearn.datasets*" can be used to load the wine dataset into a "*DataFrame*" by using the below commands:

```
data = load_wine()
df = pd.DataFrame(data.data, columns=data.feature_names)
df['target'] = pd.Series(data.target)
```

a) Load the wine dataset. Which feature is categorical and why? Compute the frequency (not the occurrence) of each value of the categorical feature. Include the code in your report. [8 marks]

b) Compute two different univariate and two different multivariate summaries for all numerical features. Include the code in your report. [8 marks]

c) Group observations by the categorical feature & compute the corresponding median for each remaining numerical feature. Include the code in your report. [8 marks]

d) Create a scatter plot for the pair of distinct numerical features with the highest correlation. Include the code in your report. [4 marks]

2) Consider the following sales data:

[5, 20, 1, 6, 13, 8, 9, 11, 17, 7, 2, 12]

Apply the following binning techniques on the data, assuming 3 bins in each case:

- Equal-frequency binning
- Smoothing by bin boundaries

[8 marks]

3) Load the file *country-income.csv* which includes numerical and categorical features. Perform data cleaning to replace any NaN values with the mean value of that particular feature. Then replace any categorical features with numerical features. Display the resulting dataset. You can use the sklearn.impute and sklearn.preprocessing packages to assist you. Include the code in your report. [8 marks]

4) Load the file *shoesize.csv,* which includes measurements of shoe size and height (in inches) for 408 subjects, both female and male. Plot the scatterplots of shoe size versus height for female and male subjects separately. Compute the Pearson's correlation coefficient of shoe size versus height for female and male subjects separately. What can be inferred by the scatterplots and computed correlation coefficients? Include the code in your report. [8 marks]

5) Using the *wine dataset* from question 1, perform Principal Component Analysis (PCA) with 2 components. Transform the data and plot the scatterplot of all samples along the two principal components, color-coded according to the "target" column (this column is the class and should not be used in the PCA analysis). What insights can you obtain by viewing the scatterplot of the principal components? Can you easily distinguish the samples that belong to one class from the samples that belong to another class and so on? In other words, are the different classes (quite) distinctive one from the other, or is there a lot of overlap? If it is the latter, then why is this happening? What can be done to the data prior to performing PCA in order to alleviate this issue? Do this action first and then perform PCA with 2 components, transform the data and plot the scatterplot of all samples along these two principal components, color-coded according to the "target" column. Now are the different classes (quite) distinctive one from the other? Include the code in your report.

[20 marks]

6) In Lab session 3 (Data Exploration and Data Visualisation), in subsection 1.9 you had created and visualised a heatmap for the distance matrix for the *graduation_rate.csv*. You may have noticed that the distance matrix visualisation is not very informative. However, it is still possible to infer that the average distance between students whose parents only have some high school education and students whose parents have a master's degree is larger than the average distance between students whose parents only have some high school education. Explain how this inference is possible from the visualisation.

[4 marks]

7) Use the file *country-income.csv* and perform the following:

a) Load the CSV file using Cubes, create a JSON file for the data cube model, and create a data cube for the data. Use as dimensions the region, age, and online shopper fields. Use as measure the income. Define aggregate functions in the data cube model for the total, average, minimum, and maximum income. Include the code in your report (and show the files created). [8 marks]

b) Using the created data cube and data cube model, produce aggregate results for: i) the whole data cube; ii) results per region; iii) results per online shopping activity; and iv) results for all people aged between 40 and 50. [8 marks]

8) Consider a dataset that contains only two observations $\mathbf{x}_1 = (1,2)$ and $\mathbf{x}_2 = (-1,0)$. Suppose that the class of the first observation is $y_1 = 1$ and that the class of the second observation is $y_2 = 0$. How would a 1-nearest neighbour classifier based on the Euclidean distance classify the observation $\mathbf{x}_3 = (3,2)$ and why? How would the same classifier classify the observation $\mathbf{x}_4 = (0,1)$ and why? [8 marks]

Submission Requirements - Deliverables

You are asked to submit a **report** that should answer the above questions, show the obtained results and include code (not as screenshot/image). The report should be in **PDF, .doc or .docx format** (so *not notebook* etc), named as:

Assignment1-StudentName_StudentSurname-StudentNumber.pdf (or .docx/.doc extension)

Marking Criteria

- Correct and sufficient explanations, plots and written code that answer the questions and show understanding.
- Clearly and succinctly written report

Important notes about the assignment:

- This is an individual assignment.
  Plagiarism is an irreversible non-negotiable failure in the course (if in doubt of what constitutes plagiarism, please ask).

- The submission cut-off date will be 7 days after the deadline. Late submissions will receive late penalties in line with the late penalty policy, see EECS handbook and QMUL assessment handbook.

- Cases of **Extenuating Circumstances (ECs)** have to go through the proper procedure of the School in due time. Only cases approved by the School in due time can be considered.

- No other means of submission other than submitting your assignment through the appropriate QM+ link are acceptable at any time. Submissions sent via email will **not** be considered.

Submission Checklist

- Has your file been saved in **PDF or .doc/.docx format**?
- Have you clicked **[Submit]** after uploading?
- Have you checked that the file you uploaded is the correct version?

The first time you submit, you will be required to accept the Turnitin End User Licence Agreement.

After uploading, it is your responsibility to check that your file is in the correct format and that it is readable.