MTHM503: Applications of Data Science and Statistics

Victoria Volodina

December 2024

Introduction

This assignment consists of three sections. In all sections, you must use `Python`. All answers should be submitted in a rendered version of a Jupyter notebook. The rendered files must be in PDF format. Your Jupyter notebooks should be clearly organised and annotated numerically according to the corresponding questions. In these exercises, please focus on writing clear, understandable code (with comments where necessary). The code you submit will be evaluated not only for its accuracy but also for its clarity. Make sure to explain your approach and illustrate key decisions taken in your solutions.
This is an individual assignment and you must not collaborate with others when working on this assessment. Please refer to the Faculty guidelines on plagiarism: `https://www.exeter.ac.uk/students/facultycases/academicconductandpractice/academicmisconduct/`.

This assessment is AI-supported and permits ethical and responsible use of GenAI tools. You may use GenAI tools to help with your coding, improve the structure of your work or quality of English language. You MUST NOT use GenAI tools to help with your modelling or data science and statistical analysis.

If you have used GenAI tools in you assessment, you are required to acknowledge that you have used GenAI tools in your work. In addition to your statements which acknowledge your use of GenAI in your academic work, you should also reference the use of GenAI where appropriate. Please refer to the Faculty guidelines on the use of GenAI: `https://libguides.exeter.ac.uk/referencing/generativeai`

A. Analysing the gene expression data set [25 marks]

Data science methods are commonly used in the analysis of genomic data. In this part of the assessment, you will apply Principal Component Analysis (PCA) and logistic regression to cancer cell line microarray data.

Data

The file `gene_file.csv` contains brain cancer gene expression from Curated Microarray Database (CuMiDa) with the following information:

- 4 cancer types
- 54675 gene expression measurements
- 117 cancer cell lines

In this dataset, each cell line is identified to its associated cancer type. The data has 117 rows and 54,677 columns.

(i) Split the dataset into training and testing sets retaining 80% for training and 20% for testing. Scale the variables (gene expression measurements) to have mean zero and standard deviation one. Explain why the data should be scaled.

(ii) Perform dimensionality reduction using PCA on the scaled data set and choose the optimal number of retained principal components. Explain your choice.

(iii) Visualise the results of dimensionality reduction and comment on your findings. Consider using scatter plots or other relevant visualisations to showcase the reduced dimensionality.

(iv) Construct a logistic regression using the PCA-transformed dataset to accurately predict the cancer type. Choose a performance metric to evaluate the effectiveness of your logistic regression model.

B. Sea level change modelling [35 marks]

The data within `sealevel.csv` contains the following information:

- Change in Mean Sea Level (mm) recorded monthly from 01/01/2000 until 01/01/2024;
- Location corresponds to the recorded change in mean sea level across the World, the Bering Sea and the Mediterranean;

You are asked to build a regression model to explain the changes in mean sea level in these three locations over time. You should aim to present a parsimonious model: one that is just complex enough to explain the data. Using your regression model, comment on the variation in the changes in mean sea level over time at these three locations: the World, the Bering Sea, and the Mediterranean.

Your answer should include graphs (three maximum) and text (three short paragraphs).

C. Power demand clustering [40 marks]

Aim

The aim of this part of the assignment is to perform clustering on power data recordings at substations in order to see whether there are groups that have similar demand profiles.

Data

The dataset `January_2013.csv` contains variable data for 535 substations. Each row of a dataset represents a day's recording at a given substation, on a particular date. The first two columns indicate the date and Substation ID. The remaining 144 columns indicate the raw power data recorded by monitors at 10 minute intervals throughout the day.

1. [5 marks] Data-processing

(i) Each row of the data represents a day of observation. Divide the power recorded in each 10 minute interval by its corresponding daily maximum (i.e. the maximum value in the row). Print the first two rows of this processed data.

(ii) Remove all the observations (power data recordings) obtained during the weekends.

(iii) Each substation has a number of days when power data were collected. In this part, we are going to calculate the average daily power demand profiles, which represent the average power produced by the substation in each 10 minute interval, across the many days of collection. In other words, for each substation, obtain a single row of length 144 where each element of it represents the average power produced within that 10 minute interval. Note, for this question, you are to use the normalised data obtained from the previous part. Print the first two rows of this processed data.

2. [20 marks] Hierarchical clustering.

In this question, we are going to use the average daily power demand profiles

(i) Using your preferred choice of a distance function (i.e. Euclidean distance, Manhattan distance etc.), create a distance matrix for these data (i.e. the averaged data created in the previous part) and produce a dendrogram by performing hierarchical clustering. Explain your choice of the linkage in hierarchical clustering. Hint: to visualise distance matrix, you can use the following `seaborn` function `clustermap`.

(ii) Using your dendrogram visualisation as a guide, choose an appropriate number of clusters and label each substation according to its cluster membership. How many substations are in each of your clusters?

(iii) For each of your clusters, visualise the daily average demand profiles. Note a demand profile is the power produced in each of the 10 minute intervals across the day.

3. [10 marks] Cluster evaluations.

Perform cluster evaluation using silhouette values and silhouette plots to validate your clustering analysis results.

4. [5 marks] Putting data back into context.

Based on your clustering analysis, explain the patterns you observe in the power usage. Describe the defining characteristics of daily average demand profiles of each cluster.