Home Assignment #1

Machine Learning and Deep Learning (CDSCO2041C)

Somnath Mazumdar
sma.digi@cbs.dk
Department of Digitalization, Copenhagen Business School

Deadline: 28-02-2025

Instructions

1. Write your complete name and student ID on the report.
2. The project report should be in pdf format of a maximum of 5 pages. The report should conform to the general formatting guidelines and academic standards that are expected for written projects at CBS.
3. If you have more content to present, feel free to include them in the Appendix to the report. But, not the complete code.
4. APA seventh edition reference style is ONLY acceptable.
5. Complete solution code should be submitted as **one single *jupyter* notebook** and attach separately.

Assignments

Question 1

Exploratory Data Analysis (EDA)

The 2024 summer Olympics was held from July 26 to August 11, 2024, in France, with several events starting from July 24. Perform EDA on the given dataset to extract the primary factors that increase Airbnb booking prices.

1. Write a Python code to perform covariance and correlation-based analysis to detect price variability (if any). **Do not** use any built-in covariance and correlation method. Write your **own** code for the calculation.
2. Write another Python code to visualise your findings from the previous step and explain your findings briefly. *Hints: You might look into* histograms, boxplots, scatterplots *few to name*.

Question 2

Cluster Analysis

Cluster analysis helps to group data points based on high/low intra- or inter-cluster similarity. Choose one of the clustering algorithms that were covered during the lectures.

1. Then choose the desired data that you think are suitable for reasoning the factors of increasing Airbnb rental prices. Explain the results, and you are free to use graphs/plots and any other sort of visualization.

Question 3

Principal Component Analysis (PCA)

1. Now, apply PCA further to identify the main drivers of Airbnb rental inflation.
2. The dataset has multiple features. Suppose that you decided to retain the top 'n' principal components that explain at least 95% of the variance. How can we determine the optimal 'n' without computing PCA for all features?
3. Now, you have performed PCA on the dataset with 'n' features. Here, you want to add a new feature that is a linear combination of existing features. Do you think it is a good idea? How will this impact the principal components and explained variance? [**Answer should be in textual form**]
4. Finally, now compare the knowledge obtained from the cluster and PCA analyses.

Data

Use the Filename "listings.csv" and "reviews.csv.gz" data of "Paris, Île-de-France, France" for **all** questions. If needed for analysis, then feel free to use other related data files, such as neighbourhoods.csv or similar.

- Data Source: `http://insideairbnb.com/get-the-data/`