CITS5508 Machine Learning
Semester 1, 2024
Lab sheet 8
(Not assessed)

You will develop Python code to perform clustering using principal components analysis (PCA). Certify that the presentation of your Python notebook is good and that you used the Markdown cells well. Make sure you properly format your plots and results. For instance, all your diagrams/plots should have proper axis labels and titles to help the reader understand what you are plotting. Use the lab sheets to learn how to improve the presentation of your notebook, as you will need this in the assessments.

A clustering analysis on airlines safety records

The website *FiveThirtyEight* provides discussions based on data-driven views of selected topics. In 2014, Nate Silver, the editor-in-chief, wrote an article about people's reactions to high-profile airline incidents and why they would avoid travelling with particular airlines. For interested readers, the article is available at https://fivethirtyeight.com/features/should-travelers-avoid-flying-airlines-that-have-had-crashes-in-the-past/. In this lab sheet, we will use the provided data set to investigate which airlines are similar based on their past safety records.

The provided `airline-safety.csv`[1] data set contains:

1. airline (asterisk indicates that regional subsidiaries are included)
2. avail_seat_km_per_week: available seat kilometres flown every week
3. incidents_85_99: total number of incidents, 1985-1999
4. fatal_accidents_85_99: total number of fatal accidents, 1985-1999
5. fatalities_85_99: total number of fatalities, 1985-1999
6. incidents_00_14: total number of incidents, 2000-2014
7. fatal_accidents_00_14: total number of fatal accidents, 2000-2014
8. fatalities_00_14: total number of fatalities, 2000-2014

Tasks

1. Considering the k-means clustering, plot the silhouette score for values of k varying from 2 to 8. Discuss the results and comment on what would be a good choice(s) for k. For the k-means clustering, you should use the Euclidean distance and set `random_state` to "5508".
2. Apply k-means clustering with the value of k obtained in Task 1. Describe the main characteristic of each group, that is, provide the interpretation of the groups in terms of safety records. For the k-means clustering, you should use the Euclidean distance and set `random_state` to "5508".
3. Explain your decision about scaling or not the data before running k-means (on Tasks 1 and 2), and explain your decision about using or not all variables in the analysis.

[1] More information about the data set can be found at https://www.kaggle.com/datasets/fivethirtyeight/fivethirtyeight-airline-safety-dataset

4. Perform a k-means cluster analysis, considering the value of k from Task 1, and: (a) the three variables from the years 1985-1999; (b) the three variables from the years 2000-2014. Did the clusters change? Explain the results. For the k-means clustering, you should use the Euclidean distance and set `random_state` to "5508".

5. Consider three new features as the difference of the variables from 2000-2014 divided by the respective variables from 1985-1999. Now, perform a k-means cluster analysis, considering the value of k from Task 1. Present the results of this cluster analysis, and compare them with the results from Task 2 and Task 4. For the k-means clustering, you should use the Euclidean distance and set `random_state` to "5508".