CITS5508 Machine Learning
Semester 1, 2023

Assignment 3

Assessed, worth 15%. Due: 5pm, Friday 26th May 2023

All work is to be done individually.

1 Instructions

Name your Jupyter Notebook file as **assignment3.ipynb** and submit it to LMS before the due date and time shown above. You can submit your file multiple times. Only the latest version will be marked. The data set mentioned in each part is available on LMS.

The assignment questions involve the analysis of some data. You should present a Jupyter notebook portfolio, with *Markdown* cells inserted at appropriate places to explain your code and describe your analysis. This portfolio should include the aims, methodology, results and discussion of your data analysis in a concise and readable fashion.

Marks for each question will be awarded for:

- Exposition: Your portfolio should be well organised, and you should aim to write concisely. For each question, you should explain the problem, the data and the aims of the analysis. You should also describe and discuss the machine learning techniques that you have used and the data preparation steps you did (when necessary).
- Data visualisations: Your portfolio should include appropriate and well-presented visualisations that are meaningful to the analysis. For instance, when presenting a plot, you should provide readable labels, axis values, etc.
- Application of techniques: Your portfolio should include the correct use of the machine learning techniques and the interpretation of the results obtained from these techniques.
- Presentation of results: Describe the results of your analysis and their interpretations. Software output is not a valid answer. You must format and present your answer appropriately (tables, graphs, important measures, etc.). You should not add irrelevant information when presenting the results.
- Discussion: Based on your results, describe the conclusions of your analysis.

Provide comments about your code. This is a data science project, and therefore it is also important that your code run properly and efficiently.

2 Plagiarism and Penalty on late submissions

See the URL below about late submission of assignments:

https://ipoint.uwa.edu.au/app/answers/detail/a_id/2711/~/consequences-for-late-assignment-submission

Plagiarism: In accordance with University Policy, you certify that all work submitted for this assignment is your own and that all material drawn from other sources has been fully acknowledged.

3 Questions

3.1 A model for diagnosing cancer (60 marks)

Determining whether a tumour is malignant or benign is one of the challenging aspects when treating cancer. Machine learning techniques can help identify cancer types by extracting the differences in cell nucleus features. In this part of the assignment, you will extend the analysis on the Breast Cancer Wisconsin (diagnostic)[1] dataset. The `breast-cancer.csv` data set provided contains:

1. Patient ID number
2. Diagnosis (M = malignant, B = benign)
3. Ten cell nucleus features, namely:
    - radius (mean of distances from centre to points on the perimeter)
    - texture (standard deviation of grayscale values)
    - perimeter
    - area
    - smoothness (local variation in radius lengths)
    - compactness (perimeter^2 / area - 1.0)
    - concavity (severity of concave portions of the contour)
    - concave points (number of concave portions of the contour)
    - symmetry
    - fractal dimension ('coastline approximation' - 1)

Task 1: Build a logistic regression and a decision tree model to predict the tumour status. The presentation should include a comparison of the two models and a recommendation regarding which would be more appropriate in a clinical setting. Note: You should use the fundamental steps of a machine learning project (e.g. hyperparameters fine-tuning, cross-validation, etc.).

Task 2: Describe the features that have a higher chance of impacting the prediction of the tumour status according to each of the two models. Discuss their similarities/differences.

Task 3: Using PCA, present the scatter plot of the data on the first two principal components. Add to your scatter plot different colours to represent the two classes in the data. What proportion of data variance is explained using the first two principal components?

Task 4: Considering the first two principal components from Task 3, present the biplot with the variables vectors and the observed data projected on the first two principal components (with the colours for the two categories). Give your interpretation of the results.

Task 5: Using the plot of Task 4, which variables are more related to the tumour status? Justify your answer. Compare the results obtained with the results obtained in Task 2.

Task 6: Using PCA, determine the number of components to retain 95% of the explained variance. Use as new features the resulting principal components scores and repeat task 1 on these new features. You can choose one of the models (logistic regression or the decision tree). What is the dimension of the new (projected) data set? Comment on the performance resulting from using the original and principal components features.

[1] More information about the data set can be found at https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data and https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

3.2 A clustering analysis on airlines safety records (50 marks)

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 part of the assignment, we will use the provided data set to investigate which airlines are similar based on their past safety records.

The provided `airline-safety.csv`[2] 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

Task 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".

Task 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".

Task 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.

Task 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".

Task 5: Consider three new features as the ratio 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".

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

3.3 A clustering analysis on the USArrests data (40 marks)

The `USArrests` data contains the statistics, in arrests per 100,000 residents, for three crime-related features (assault, murder and rape) for all 50 US states in 1973. An additional feature, UrbanPop, is also included and describes the percentage of the population living in urban areas.

The provided `USArrests.csv` data set contains 50 observations on 4 variables:

- Murder: murder arrests (per 100,000)
- Assault: assault arrests (per 100,000)
- Rape: rape arrests (per 100,000)
- UrbanPop: percent of the population living in urban areas

Task 1: Using the raw data, perform a hierarchical clustering with complete linkage and Euclidean distance to cluster the states. Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which cluster? Describe their characteristics.

Task 2: Repeat Task 1 after scaling the variables to have zero mean and unit standard deviation. What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled? Justify for your answer.

Task 3: Perform PCA on the data. Now perform hierarchical clustering with complete linkage and Euclidean distance on the first two principal component score vectors rather than the raw data. Cut the dendrogram at a height that results in three distinct clusters. Present the scatterplot of the first two principal components using different colours for the instances on each cluster (three colours for three clusters). Compare the group characteristics to the group characteristics obtained in Task 2.

Task 4: Repeat the analysis of Task 3 using the K-means clustering (with K=3). That is, use the first two principal components score vectors as features and set the initial centroids of the K-means as the group means obtained from the hierarchical clustering on Task 3. Compare the results from the K-means clustering to the results from the hierarchical clustering of Task 3. Which one do you think provides a better result? For the K-means clustering, you should use the Euclidean distance and set `random_state` to "5508".