MPhil in Data Intensive Science

Sunday 26th November 2023

M1: Applied Data Science - Coursework Assignment

This is the coursework assignment for M1: Applied Data Science.

*You should attempt **all** questions. The appropriate number of marks allocated to each question, or each part of a question, is indicated in the square brackets, [...], to the right of each question or part of question.*

You should write a report of no more than 3000 words to accompany the code that you write to solve the problems. Your report should mirror the structure of the questions (i.e. Section A, Q1, (a) ... etc.)

*Your code and a pdf of your report should be submitted to a private GitLab project which can be found at https://gitlab.developers.cam.ac.uk/phy/data-intensive-science-mphil/m1_assessment/<your_crsid>. Only you will have access to this project and your access will expire after the deadline for submission which is **23:59 on 20th December 2023**. You should ensure that your code is well documented with regular comments.*

The GitLab project should contain a README.md file which describes exactly how to run the code. You should also provide a Conda environment.yml file[1] that lists any Python packages you have used and their versions. You should be able to run your code within a new Conda environment created from this file.

[1] https://conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html#sharing-an-environment

This coursework is based around analysing four datasets:

- ADS_baselineDataset.csv
- A_NoiseAdded.csv
- B_Relabelled.csv
- C_MissingFeatures.csv

The baseline dataset comprises 500 observations (stored as rows) of 1000 features (stored as columns). In the final column, you will find *a priori* defined labels for the observations. The three other datasets (A, B and C) have been derived from the baseline dataset and contain fewer observations and/or fewer features.

SECTION A

Attempt **all** *questions from this Section. [50 marks]*

This section uses datasets A, B and C and requires you to demonstrate that you can use DS/ML techniques to understand characteristics of your data.

1 Dataset A: Exploration, Dimensionality Reduction and Clustering [15]

(a) Generate density plots for the first 20 features. Include the figure in your report and state what you observe.

(b) Apply PCA to visualise the features in 2D and include this visualisation in your report. Comment on what you observe and refer this back to the density plots.

(c) Partition the data into two training sets of equal sizes. Apply k-means clustering to each training set, using the default scikit-learn parameters. In each case, the unused data can be mapped onto the learned clusters. In your report, briefly explain how and why this is possible. Compare both clusterings for the combined training set, using a contingency table.

(d) Comment on the size of the clusters and the cluster stability in your report, and repeat k-means clustering with a different number of clusters.

(e) Identify the k-means clusters within the PCA figure. Comment on the differences between performing (i) k-means followed by PCA visualisation or (ii) PCA followed by k-means. Which do you think is better in general?

2 Dataset B: Missing Labels and Duplicated Observations [15]

(a) Summarise the frequency of the labels in the dataset, and display this information as a table in your report.

(b) Identify duplicated observations (which may or may not have been mislabelled). In your report, state these duplicate observations and explain a strategy that you could take to address this. Implement this process in your code.

(c) In your report, briefly describe two approaches for handling observations with missing labels, and outline the pros and cons for each. Briefly mention the difference between missing at random (MAR) and missing not at random (MNAR).

(d) For observations without missing data, predict the classification labels using multinomial logistic regression or k-nearest neighbours. Use this model to predict the missing labels. Recalculate the overall summary of labels and compare this to the original summary.

3 Dataset C: Missing Data and Outliers [20]

(a) Summarise the missing data in this dataset. In your report, identify which features are affected.

(b) In your report, describe two methods for imputing missing values across multiple features: one static and one model-based. Outline the pros and cons for both. Briefly describe the advantages of using a multiple imputation.

(c) Perform the imputation using an appropriate model-based approach. In your report, justify the model you have chosen to use and compare the original and imputed distributions.

(d) Implement a standardisation approach to detect outliers. In your report, explain how the standardisation approach works and identify the outliers you have found.

(e) Implement a model-based approach to correct for outlier values. In your report, justify the approach and compare the original and imputed distributions.

SECTION B

Attempt **all** *questions from this Section. [50 marks]*

This section uses the baseline dataset (ADS_baselineDataset.csv) and requires you to demonstrate that you can implement supervised and unsupervised learning methods.

4 Baseline Dataset: Supervised Learning and Random Forests [30]

(a) Briefly describe the differences between a decision tree, bagging and a random forest (for classification). Describe one criterion you can use to measure the quality of a split within a classification tree. Identify two hyperparameters of a single tree within a random forest, and relate one of these to a heuristic involving the number of features.

(b) Pre-process and filter the data in preparation for the classifier. In your report, explain how you have pre-processed the data and justify any steps you have taken.

(c) Apply the classifier on training data using the default hyperparameters and summarise the output of the model. Evaluate the test set classification error.

(d) Optimise the algorithm with respect to the number of trees in the random forest. You should be able to do this without explicitly performing cross-validation.

(e) Calculate the feature importance. In your report, describe and interpret the feature importances. Retrain the model using a subset of the most important features. In your report, indicate which features you have chosen and compare the retrained classifier with the original classifier.

(f) Repeat steps (b), (c) and (e) for one other supervised learning classifier. Compare and contrast the two classification approaches and your results for each.

5 Baseline Dataset: Unsupervised Learning - Clustering [20]

(a) Apply two different clustering techniques to the dataset (excluding the class labels). In your report, briefly explain the differences between the two methods and in more detail explain the difference in outputs between the two methods when applied to this dataset. Include a contingency table to summarise the number of observations assigned to each cluster using each technique. You may need to choose an appropriate number of clusters and justify this choice.

(b) For each of the clusterings you have obtained in (a), train a classifier to predict cluster membership and use this to identify the most discriminative features. In each case, apply the same clustering technique using only this subset of features. In your report, compare the partitions using the subset of features against the ones obtained using the entire dataset.

(c) Visualise the clusters in a lower-dimensional space and colour each point according (i) the cluster membership, (ii) the value of the most discriminative feature, (iii) the value of the next most discriminative feature.

END OF PAPER