Introduction

In this coursework, a spreadsheet has been provided to perform a set of data analysis. The spreadsheet contains the following information: the index of student, gender of student, the programme that a student is enrolled, the grade that the student is in, total marks that a student is awarded and the mark of 5 exam questions (indexed as MCQ, Q1, Q2, Q3, Q4 and Q5).

The index student ranges from 1 to 619. The gender of the student is represented as “1” and “2”. The grade of the student is either “2” or “3”. The programme of the student is represented as “1”, “2”, “3” and “4”. The full mark of the whole exam paper is 100. The full mark for 5 exam questions are 54 marks (MCQ), 8 marks (Q1), 8 marks (Q2), 14 marks (Q3), 10 marks (Q4) and

6 marks (Q5) respectively.

The coursework requires students to extracts features of the data and analyse the distribution of the feature with association of the programme that a student is enrolled.

Tasks

1. Observe the distribution of raw data with a box plot. Discuss what action should be taken in order to reduce the impact of scale for the raw data.

2. Perform Principal Component Analysis (PCA) to the data, observe the distribution of components. Find a set of components that make the classification of programme easier.

3. By your own way, extract the features that are easy to classify the programme that the student belongs to.

4. Visualise and compare raw features, scaled features, PCA features and your resulting features.

CW1 Tasks

1. Observe the distribution of raw data with a box plot. Discuss what action should be taken in order to reduce the impact of scale for the raw data.

2. Perform Principal Component Analysis (PCA) to the data, observe the distribution of components. Find a set of components that make the classification of programme easier.

3. By your own way, extract the features that are easy to classify the programme that the student belongs to.

4. Visualise and compare raw features, scaled features, PCA features and your resulting features.

CW2 Tasks

1. Use decision trees to classify the programme that a student comes from. Comment on the different decision tree resulted. You may want to try random forest and compare the result. You may also try different features to pursue the best result.

2. Classify the programme that a student comes from with SVM (Support Vector Machine). Try the best configuration of SVM with your preferred features.

3. Classify the programme that a student comes from with Naïve Bayes. Try different sets of features. Please explain how the features impact the classification process.

4. Build up an ensemble classifier. Compare the choice of feature that associates with different classifiers.

CW3 Tasks

1. Use GMM (Gaussian Mixture Model) to fit a distribution of raw features / your own features. Find a way that the GMM reflects the distribution of programme information.

2. Use k-means to fit a distribution of raw features / your own features. Find a way that the resulting clusters reflect the distribution of programme information

3. Use hierarchical clustering to fit a distribution of raw features / your own features. Find a way that the resulting clusters reflect the distribution of programme information.

4. Compare and evaluate how the clustering clusters associate with the programme that the student comes from.