Clustering and Randomised Algorithms

*Statistic for Computer Science*

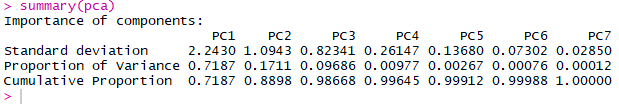
# Part 1

## Section 1

### task 1

1. PCA results using the first 7 attributes:





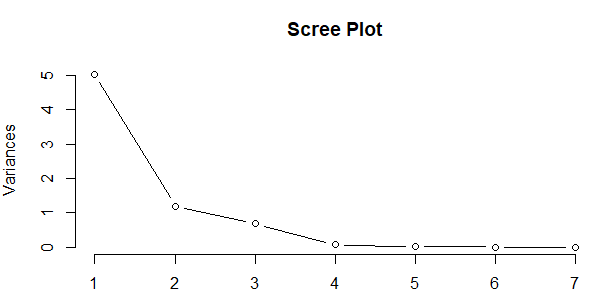
1. Eigenvalues for each of the attributes in descending order:



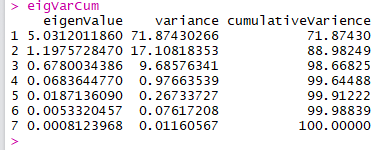
|  |  |
| --- | --- |
| **PC1** | 5.0312011860 |
| **PC2** | 1.1975728470 |
| **PC3** | 0.6780034386 |
| **PC4** | 0.0683644770 |
| **PC5** | 0.0187136090 |
| **PC6** | 0.0053320457 |
| **PC7** | 0.0008123968 |

1. Scree plot using the eigenvalues found above:





1. After performing the PCA the 7 original scaled attributes have been combined to form 4 leaner combinations or Principal Components. However, the first Principal Component is generally the most important as it explains most of the data, followed by Principal Component 2 then 3 and so on. Using the summary function for the PCA we get 3 values; the standard deviation, proportion of variance and the cumulative proportion. The standard deviation is the standard deviation of the data along a single principal component; a measure of the variability across that principal component. The proportion of variance represent the proportion of all the variability in the original data explained by the principal component for instance PC1 explains 71.87% of the data followed by PC2 which explains 17.11%, followed by PC3 which explains 9.69%. The cumulative proportion of PC1 explains 71.87% of the data, PC1 and PC2 combined explain 88.98%, PC1, PC2 and PC3 combined explain 98.67% of the data. This goes on until a 100% of the data is explained after combining all 7 principal components. All this is represented in the scree plot above. The table below shows the eigenvalue, percentage of variance and percentage of cumulative variance for each principal component.



From this table it is apparent that the first principal component is the best one as at explains the data the most from all 7 principal components.

The following table show the first 10 rows of the data set with the scores from the first 2 principal components in the last 2 columns.

## Section 2

### Task 1

Research and compare different clustering algorithms that can be applied to your data set.

You must compare *at least* 1 randomised and 1 non-randomised clustering algorithm.

You must compare *at least* 3 clustering algorithms in total.

For each of the chosen algorithms, give information like expected results, advantages and disadvantages of the algorithm with respect to this data set. (5 marks)

Clustering is the process of partitioning a data set into homogeneous groups based on given features such that similar objects are kept in a group while dissimilar objects are placed in a different group. It is the most important unsupervised learning problem as it deals with finding structure in a collection of unlabelled data.

There are many clustering algorithms that can be used in any given scenario. The most popular ones include k-means, hierarchical clustering, fuzzy c-means and mixture of gaussians.

K-means

### Task 2

### Task 3

### Task 4

# part 2

## Section 1

### Task 1

### Task 2

### Task 3

## Section 2

### Task 1

### Task 2

# references