Clustering and Randomised Algorithms

*Statistic for Computer Science*

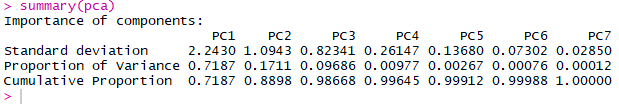
# Part 1

## Section 1

### task 1

1. PCA results using the first 7 attributes:





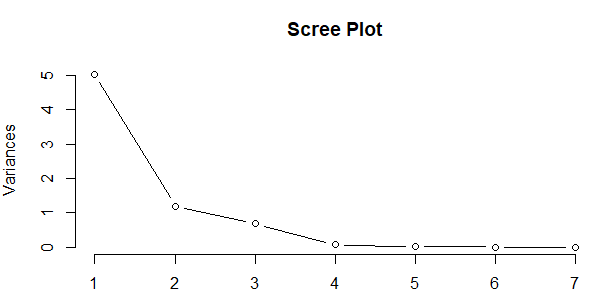
1. Eigenvalues for each of the attributes in descending order:



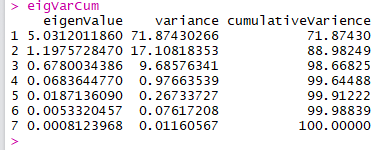
|  |  |
| --- | --- |
| **PC1** | 5.0312011860 |
| **PC2** | 1.1975728470 |
| **PC3** | 0.6780034386 |
| **PC4** | 0.0683644770 |
| **PC5** | 0.0187136090 |
| **PC6** | 0.0053320457 |
| **PC7** | 0.0008123968 |

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

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** is perhaps the most common algorithm used because it is a simple and easy way to classify a given data set through a certain number of clusters. The main idea is to define k centres, one for each cluster. These centres should be placed in a calculated way since different location cause different results. The better choice is to place them as far away as possible from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centre. When no point is pending, the first step is complete, and an early group age is done. At this point we re-calculate k new centroids as barycentre of the clusters resulting from the previous step. After we have these k new centroids, a new binding must be done between the same data set points and the nearest new centre. Here a loop is generated and as a result we notice that the k centres change their location step by step until no more changes are done. This is a randomised clustering algorithm as the centres are randomly selected.

K-means is Fast, robust and easier to understand. It is Relatively efficient and Gives best result when data set are distinct or well separated from each other. However, it requires specification of the number of clusters. it is not invariant to non-linear transformations therefore with different representation of data we get different results. The Euclidean distance measures can unequally weight underlying factors. This algorithm is Unable to handle noisy data and outliers and it fails for non-linear data set

**Hierarchical** clustering is an algorithm that builds a hierarchy of clusters. This algorithm starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. The algorithm terminates when there is only a single cluster left. This is a non-randomised clustering algorithm. An advantage to hierarchical clustering is that it does not require information about the number of clusters that are needed. It is also easy to implement and gives best result in some cases. On the downside you can never undo what was previously done and a time complexity of at least O(n2 log n) is required (n being the number of data points). Moreover, no objective function is directly minimized and sometimes it is difficult to identify the correct number of clusters by the dendrogram.

Differences between the above algorithms:

* Hierarchical clustering can’t handle big data well unlike K Means clustering since the time complexity of K Means is linear (O(n)) while that of hierarchical clustering is quadratic (O(n2))
* Since K Means clustering starts with a random choice of clusters, the results produced by running the algorithm multiple times might differ. While results in Hierarchical clustering are reproducible
* K Means works well when the shape of the clusters is hyper spherical (2D circle, 3D sphere)
* K Means requires prior knowledge of K i.e. no. of clusters you want to divide your data into. But, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram

The **fuzzy c-means** clustering algorithm works by assigning membership to each data point corresponding to each cluster centre based on distance between the cluster centre and the data point. The closer the data is to the cluster centre, the closer is its membership towards that particular cluster centre. Fuzzy c-means allows one piece of data to belong to two or more clusters. This method is frequently used in pattern recognition. Fuzzy c-means gives best result for overlapped data set and is comparatively better then k-means algorithm. Unlike k-means where data point must exclusively belong to one cluster centre here a data point is assigned membership to each cluster centre as a result of which a data point may belong to more than one cluster. Disadvantages to this algorithm include the fact that the number of clusters need to be specified. Also, with lower value of β (the termination criterion between [0, 1]) we get better results but at the expense of more number of iteration. The Euclidean distance measures can unequally weight underlying factors.

### Task 2

Discuss and describe how you will initialise the algorithm. Give justification of your initialisation method. (1 marks)

Discuss the metric that you will be using for the analysis. Give justification of the metric used. (1 marks)

K-means clustering is a type of unsupervised learning, used when you have unlabelled data (data without defined categories or groups). The goal of the algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are the centroids of the K clusters, which can be used to label new data and labels for the training data (each data point is assigned to a single cluster). Clustering allows you to find and analyse the groups that have formed organically. Each centroid of a cluster is a collection of feature values which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.

In the k-mean algorithm the first thing to do is to randomly initialize k points, also called means. Then categorize each item to its closest mean and update the mean’s coordinates, which are the averages of the items categorized in that mean so far. This process is repeated for a given number of iterations and at the end, we get the clusters.

K points are called means because they hold the mean values of the items categorized in it. To initialize these there are many options. One method is to initialize the means at random items in the data set. Another method is to initialize the means at random values between the boundaries of the data set. The following is the pseudocode for the process explained above:

initialize k with randomly;

for (a number of iterations){

foreach(iterate through items){

find the mean closest to the item;

assign item to mean;

update mean;

}

}

K-means algorithm works well when the data fits the cluster model where clusters are spherical; the data points in a cluster are centred around that cluster and where the spread/variance of the clusters is similar therefore each data point belongs to the closest cluster. If any of these principles does not hold then the local minimal of the k-means algorithm will be counter-intuitive, which is an issue that can occur when using this algorithm.

To find the value of K, or the number of clusters in the data, the K-means clustering algorithm must be run for a range of K values and then the results are compare. A common used metrics to compare results of K is the mean distance between data points and their cluster centroid. Since increasing the number of clusters will always reduce the distance to data points, increasing K will hence decrease this metric, until zero is reached when K is the same as the number of data points. Thus, this metric cannot be used as the sole target. Instead, mean distance to the centroid as a function of K is plotted and the "elbow point," where the rate of decrease sharply shifts, can be used to roughly determine K. Other techniques for validating K include cross-validation, information criteria, the information theoretic jump method, the silhouette method, and the G-means algorithm.

Using R libraries the K-means algorithms can be initialised and used on any data set

### Task 3

Implement the chosen k-means variation.

If you are using an existing library or tool that supplies the actual code, then for the purposes of the assignment, the implementation is the code/script/formula that calls this method.

Place screenshots of the k-means implementation. There is no need to take screenshots of the supporting code, only the k-means and it initialisation.

Apply your implementation to the data set, and describe your resulting clusters, giving screenshots of the cluster visualisation.

### Task 4

Visually inspect your cluster visualisation. Comment whether the resulting clusters appear to be good, explaining why. (1 mark)

What is the accuracy of the cluster for identifying the correct variety for the given sample? (1 mark).

The accuracy identified above might not reflect the accuracy for new kernels that were not used in the original sample to build the clusters.

Split the original data set into a training set and a test set and perform the same test again.

Find the sensitivity, specificity and accuracy of the clusters. (3 marks)

Apart from the statistics mentioned above, research another method to evaluate the resulting clusters.

Write down information about the selected method. (1 mark)

Use the selected method on your clusters. (1 mark)

Perform the following: (3 marks)

* Try different parameters, metrics and initialisations, and compare the resulting clusters.
* For the optimal setup found, repeat the experiment using different training sets and data sets. Discuss how the clusters vary, both in location and shape as well as the other methods used in evaluating clusters (sensitivity, specificity, accuracy and the method chosen).

# part 2

## Section 1

### Task 1

### Task 2

### Task 3

## Section 2

### Task 1

### Task 2

# references