**DALARNA UNIVERSITY 2020**



**STATISTICAL LEARNING**

**HOME EXERCISE - 2**

**KIRAN SAI TANNEERU**

## **h19kirta@du.se**

**Introduction:**

We have been given the Data Cortex Nuclear data set which is collected from a nuclear fraction of cortex on Mice.

The dataset consists 1080 instances and 82 attributes. For the experiment, 72 mice are considered of which 38 are control mice and 34 are trisomic mice (down syndrome). There were 15 measurements done on each mice so 38\*15 = 570 instances on control mice and 34\*15 = 510 instances for down mice. In the 82 attributes the first attribute is Mouse ID, the last four attributes are categorical Genotype, Treatment, Behavior and class. They are numerical and categorical values. The remaining attributes consists of data of different levels of proteins.

There are numerical and categorical variables on the dataset. After handling the missing values appropriately, answer the following questions.

**Methods:**

At the second part I have performed PCA and using the PCA components I have reperformed Decision Tree and SVM classification. At the third part I have performed Random Forest, Bagging and Boosting. For the Clustering task, I have performed K-means and Hierarchical clustering.

**1. a) Use the 77 proteins as predictors for decision trees and support vector machines models to make binary and multiple class classification.**

For these tasks, we should perform Binary and Multiple class classification. For the Binary class I have considered Genotype as response variable and for the Multiple class I have considered class feature as the response variable. On both we have to perform decision trees and support vector machine models.

Firstly, we do the decision tree process for both the binary class classification and the multi class classification and find their accuracy.

Then for the support vector machines we use tune() function lowers the error rate and best possible output. This process is done for all the remaining steps. In support vector machines we did the same procedure for the three kernel types linear, radial and polynomial in both binary and multiple class classification.

A picture containing table

Description automatically generatedA screenshot of a computer

Description automatically generated

|  |  |  |
| --- | --- | --- |
|  | Binary class | Multi class |
| Decision tree | 85.7% | 74.5% |
| SVM-Linear | 79.6% | 91.0% |
| SVM-Radial | 100% | 99.6% |
| SVM-Polynomial | 99.3% | 100% |

From the above we can that in case of decision tree, binary class performed well compared to multi class classification. In case of support vector machine, multi class have performed better than binary class classification.

**b) Perform principal component analysis on the 77 numerical features. Use an appropriate number of principal components as predictors and perform the same classification task.**

Principal component analysis is an unsupervised learning algorithm, in this all the components are not considered, only a few important components are taken into consideration. For principal component analysis, we divided the data of 77 components into two parts training data and testing data and considered only 10 components.

|  |  |  |
| --- | --- | --- |
|  | Binary class | Multi class |
| Decision tree | 78% | 59% |
| SVM-Linear | 73% | 73% |
| SVM-Radial | 98% | 95.5% |
| SVM-Polynomial | 93% | 94 % |

After performing principal component analysis, in decision tree still binary class is performing better. In case of support vector machine, it is complex as linear both of them gave the same output, in radial binary class is better and finally in polynomial, multi class has marginal higher accuracy.

**c) Using bagging, random forest, and boosting perform the same classification task. Compare the results of the three methods.**

**Random Forest:**

Random forest is similar to that of decision tree but it only considers the best part of the many decision trees.

**Bagging:**

Bagging is also called bootstrap aggregating, it is a ensemble meta-algorithm designed to improve the stability and accuracy of algorithms used in classification regression. It also reduces variance and helps to avoid overfitting.

**Boosting:**

Boosting refers to converting weak learners to strong learners by joining the multiple weak rules to a single strong prediction rule after running many iterations.

|  |  |  |
| --- | --- | --- |
|  | Binary class | Multi class |
| Bagging | 94% | 97% |
| Random Forest | 98% | 98% |
| Boosting | 98.4% | 99% |

From the above table, we can observe that both Boosting has higher prediction accuracy and Bagging has less accuracy compared to others.

**2. Use the dataset to perform clustering. You should try both k-means clustering and hierarchical clustering. In every case, find a number of clusters that make sense and try to explain what each cluster describes.**

For this task I have replaced the null values with the mean values

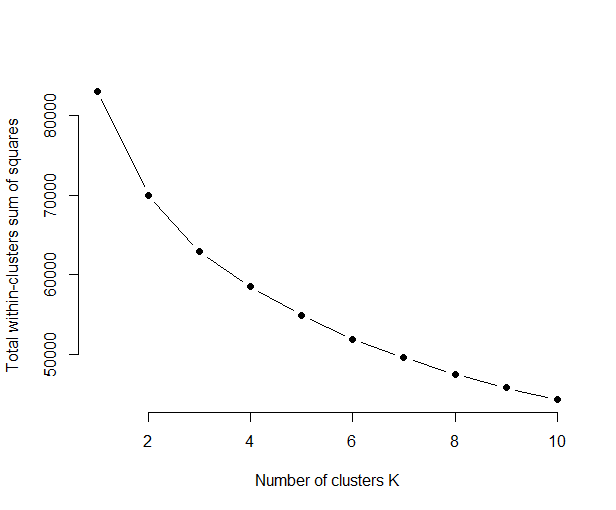
**K-Means Clustering:**

Firstly, we find the cluster plot for the whole as it is, as we can observe the below graph it is very hard to analyze.

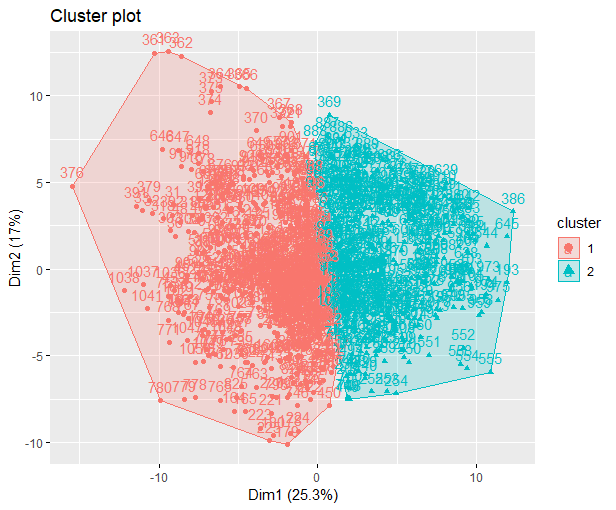
A close up of a map

Description automatically generated

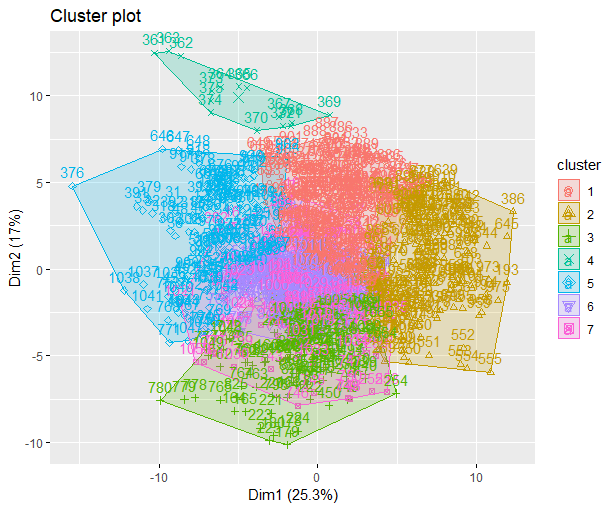
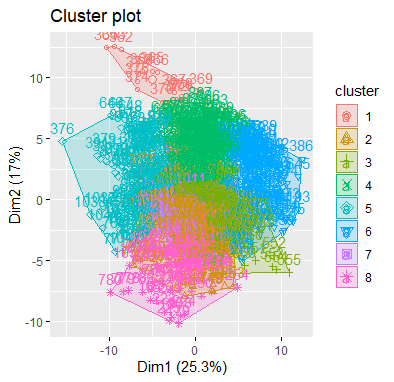
The number of clusters which are to be considered is based on the elbow plot, we can see that there is a drastic change at 2, 4 and 6 clusters and at 7, 8 the changes seen are minimal. Also based on the attribute class given in the data, we have an idea to divide the data into 8 parts. So, it is more meaningful to consider number of clusters to be 8.



So now I start doing the cluster plots with 2 clusters just to check the output.



Now I repeat the process with 7 clusters and 8 clusters.

**Hierarchical clustering:**

Hierarchical clustering considers every data point to be its own subset, then each subset is formed as a cluster. Now all the similar data types are grouped into a single cluster and non-similar one to be different cluster.

A screenshot of a cell phone

Description automatically generated

From the above dendrogram, we visualize the ideal number of clusters. Euclidean distance is used as the distance metric and from my observation 5 clusters can be taken into consideration from the above diagram. But overall we can say that the clustering part is not very clear to give a detailed explanation on each cluster.