STATISTICAL LEARNING (AMI22T)

ASSIGNMENT 2

Author: Maha Vajeeshwaran

INTRODUCTION:

This assignment consists of two tasks and it is performed with the help of R studio. I was provided data from the Data Cortex Nuclear data set. There are 38 control mice and 34 mice with Down syndrome in the dataset, which are further divided into 4 categories each (8 categories total, 4 for the control mice and 4 for the Down syndrome mice).

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

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

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.

DATA SET INFORMATION:

The data set consists of the expression levels of 77 proteins/protein modifications that produced detectable signals in the nuclear fraction of cortex. There are 38 control mice and 34 trisomic mice (Down syndrome), for a total of 72 mice. In the experiments, 15 measurements were registered of each protein per sample/mouse. Therefore, for control mice, there are 38x15, or 570 measurements, and for trisomic mice, there are 34x15, or 510 measurements. The dataset contains a total of 1080 measurements per protein. Each measurement can be considered as an independent sample/mouse.

The eight classes of mice are described based on features such as genotype, behavior and treatment. According to genotype, mice can be control or trisomic. According to behavior, some mice have been stimulated to learn (context-shock) and others have not (shock-context) and in order to assess the effect of the drug memantine in recovering the ability to learn in trisomic mice, some mice have been injected with the drug and others have not.

Classes:

c-CS-s: control mice, stimulated to learn, injected with saline (9 mice)

c-CS-m: control mice, stimulated to learn, injected with memantine (10 mice)

c-SC-s: control mice, not stimulated to learn, injected with saline (9 mice)

c-SC-m: control mice, not stimulated to learn, injected with memantine (10 mice)

t-CS-s: trisomy mice, stimulated to learn, injected with saline (7 mice)

t-CS-m: trisomy mice, stimulated to learn, injected with memantine (9 mice)

t-SC-s: trisomy mice, not stimulated to learn, injected with saline (9 mice)

t-SC-m: trisomy mice, not stimulated to learn, injected with memantine (9 mice)

The aim is to identify subsets of proteins that are discriminant between the classes.

METHODS: TASK:1

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

I started by importing the data. The data collection has 82 columns and, 1080 rows, with 77 columns containing numerical data and the remaining four columns containing category data. When looking at the summary, it becomes clear that the dataset has numerous missing values. Removing all missing values is a terrible idea since it may result in the loss of a lot of data. So, to deal with the missing numbers, I used the mean value to fill in the gaps. The dataset also shows that there isn't much variety.

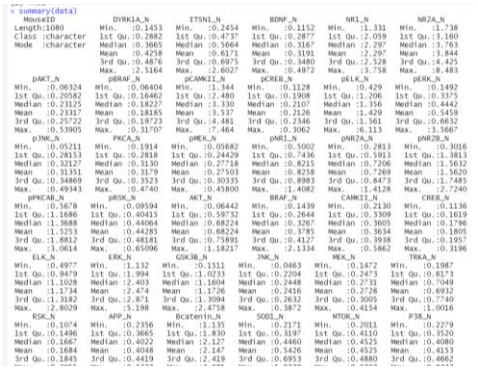


Fig:1 Summary of data after cleaning

I deleted the index column from the data set once the data cleaning process was completed. For multiclass classification, I created 77 protein columns as predictors and used class as the response variable. I used the same 77 proteins as predictors and Genotype as the response variable in the second technique for binary classification.

Y	Model	Accuracy
Binary	Decision Tree	0.858
Binary	Decision Tree with Pruning	0.8611
Multi Class	Decision Tree	0.69
Multi Class	Decision Tree with Pruning	0.69

Table:1 Results for Decision Tree

We can compare the results achieved using the validation set strategy by picking 70% of the data for training and 30% for testing from Table:1. I first tested the model with a Decision Tree without pruning, which yielded 85.8% for binary data and 0.69 percent for multi class data. There is a minor improvement in accuracy for the Binary class data (Genotype) after trimming, but there is no improvement for the multiple class data (Class).

Model	Accuracy	
	Binary (Genotype)	Multiple Class (Class)
SVM- Linear Kernel	0.65	0.565
SVM- Linear Kernel (parameter tuned)	0.9475(cost=1)	0.9414 (cost:0.1)
SVM- Radial Kernel	0.52	0.11
SVM- Radial kernel (parameter tuned)	0.963 (cost=5, gamma= 0.001)	0.953 (cost = 5, gamma = 0.001)
SVM- Polynomial Kernel	0.9815	0.95
SVM- Polynomial Kernel (parameter tuned)	0.99 (cost=100,degree=3, gamma = 1)	0.987 (cost=10, degree=3)

Table:2 Results for SVM model

Then, in accordance with the criteria, I ran the model via SVM using the same approach that I did for the Decision Tree. I tested the performance of the SVM model first with linear, radial, and polynomial kernels without tweaking any parameters, then with tuned parameters. When comparing the SVM model with parameter tuning to the decision tree model, Table 2 shows that the SVM model with parameter tuning outperforms the decision tree model. Overall, the SVM-Polynomial Kernel approach with the parameters (cost = 100, degree = 3, gamma = 1) offers a binary answer accuracy of 99 percent. The accuracy of the SVM-polynomial kernel with the parameters (cost = 10, degree = 3) in multiple class response is 98.7%.

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 (4 points).

PCA

Y	Model	Accuracy
Binary	Decision Tree	0.907
Binary	Decision Tree with CV.tree()	0.8611
Multi Class	Decision Tree	0.679
Multi Class	Decision Tree with CV.tree()	0.679

Table:3 Results for Decision Tree model using PCA

For the predictors, I used PCA (Principal Component Analysis) in this work for dimension reduction. With the use of a cumulative proportion of variation explained plot, it was discovered that 20 PCA has 90 percent of variance explained. Figure 2 illustrates this point. Data is trained and tested with a decision tree for binary and multiclass responses, just as it was in Task 1.a. For binary class responses, a decision tree without cross validation outperforms a model with cross validation and pruning. Both methods get the same results for multiclass responses. By examining Tables 1 and 2, I believe that the model with K-fold cross validation outperforms the validation set technique, since there will be no fluctuation in the findings.

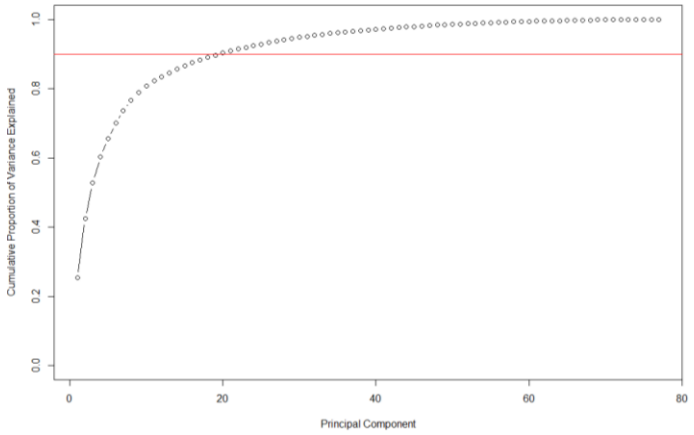


Fig:2 PCA vs Cumulative proportion variance explained

Model	Accuracy	
	Binary (Genotype)	Multiple Class (Class)
SVM- Linear Kernel	0.929	0.885
SVM- Linear Kernel (parameter tuned)	0.929(cost=0.01)	0.92 (cost:1)
SVM- Radial Kernel	0.8951	0.833
SVM- Radial Kernel (parameter tuned)	0.941(cost=1, gamma= 0.001)	0.9877 (cost = 1, gamma = 0.1)
SVM- Polynomial Kernel	0.93	0.1852
SVM- Polynomial Kernel (parameter tuned)	0.9877(cost=10 , degree=3, gamma = 1)	0.9877(cost=5, degree=3)

Table:4 SVM model using PCA

From the Table 4 we can see that SVM model with the polynomial kernel by tuning the parameter performs better when compared to all other methods for both binary and multiclass response. For Binary Class response method SVM- Polynomial Kernel with parameters (cost=10 ,degree=3, gamma = 1) shows the accuracy of 98.7% for Multiple class response method SVM- Polynomial Kernel with parameters (cost=5, degree=3) shows the accuracy of 98.7%. Both Binary and Multiple class method shows the same accuracy. For the multiple class response method it is noted that SVM- Radial Kernel with parameter (cost = 1, gamma = 0.1) also gives the same result.

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

Model	Accuracy	
	Binary (Genotype)	Multiple Class (Class)
Random forest	0.99	0.9753
Bagging	0.963	0.9907
Boosting	0.9845 (n.trees =5000, interaction.depth =4)	0.978(n.trees =5000, verbose = F, shrinkage = 0.01, interaction.depth =4)

Table:5 Random Forest, Bagging, Boosting model results

From the above table we can see that for the Binary class data Random Forest performs better with the accuracy of 99% when compared to Bagging and Boosting. But for the Multiple class response method Bagging performs better with the accuracy of 99% when compared Random Forest and Boosting. There is not much difference in the results.

Discussion:

While looking in to the results from the Task 1(a, b, c) I understood that accuracy and prediction depends on the type of data, ML models, parameters used in the models. Both with and without PCA method performs same even though I propose to use PCA for the dimension reduction this may remove features that are correlated and improves performance. Here there is not much difference in results for SVM, Random Forest, Bagging, boosting but for the binary class response Random Forest gives better results 0.99 and for the multiple class Bagging gives better result 0.9907. For the large data set computation time for the SVM will be more so I propose to use Random Forest and Bagging based on confusion matrix and accuracy scores.

TASK:2 METHOD:1

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

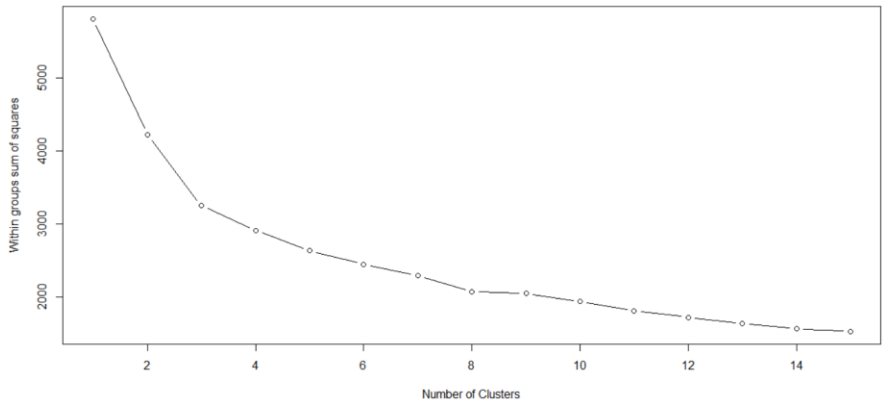


Fig:3. Elbow plot to choose optimal no of clusters

To begin, all the necessary libraries are installed, and the data is imported. The dataset is determined to have numerous missing values. The mean values for all the columns are used to fill in the missing data. As this is a clustering and unsupervised approach, we must perform the model with only predictors and no response. The total distance between the locations and their associated centroids was determined using the WSS technique. The data was then clustered using KMeans clustering with K = 2 and K = 8, and it was discovered that K = 2 is the best option based on the findings and also for the variables genotype, treatment, and behavior clustering, whereas in k=8 it is not well clustered. it is found that Initial partition into clusters can be random, or based on domain knowledge.

	Memantine	Saline
1	308	215
2	262	295

Fig:4. Clustering for Treatment variable

As per the Fig:4 for the treatment variable 308 Memantine lies in cluster 1 and 262 lies in cluster 2. For Saline 215 points lies in cluster 1 and 295 lies in cluster 2.

	C/S	S/C
1	218	305
2	307	250

Fig:5. Clustering for Behavior variable

As per the Fig:5 for the Behavior variable 218 C/S falls in cluster 1 and 307 falls in cluster 2. For S/C 305 points falls in cluster 1 and 250 points falls in cluster 2.

	Control	Ts65Dn
1	292	231
2	278	279

Fig:6. Clustering for Genotype variable

As per the Fig:6 for the Genotype variable 292 Control falls in cluster 1 and 278 falls in cluster 2. For Ts65Dn 231 points falls in cluster 1 and 279 points falls in cluster 2.

	c-CS-m	c-CS-s	c-SC-m	c-SC-s	t-CS-m	t-CS-s	t-SC-m	t-SC-s
1	78	69	85	60	48	23	97	63
2	72	66	65	75	87	82	38	72

Fig:7. Clustering for Class variable

From the above figure we see that for the cluster 1, c-CS-m recorded 78 points and in cluster 2 it recorded 72 points. For c-CS-s 69 points recorded in cluster 1 and 66 points recorded in cluster 2. For c-SC-m 85 points recorded for the cluster 1 and 65 points recorded for the cluster 2, for the c-SC-s 60 points recorded for the cluster 1 and 75 points recorded for the cluster 2. For t-CS-s 23 points recorded for the cluster 1 and 82 points recorded for the cluster 2. For t-SC-m 97 points recorded for the cluster 1 and 38 points recorded for the cluster 2.

Hierarchical clustering

In the Hierarchical clustering data distance is calculated based on the Euclidian distance. In the three linkages (Average, Single, Complete) complete linkage method is well clustered as per the figure below. Cluster is cut at the optimal distance of 8 as it seems like well clustered.

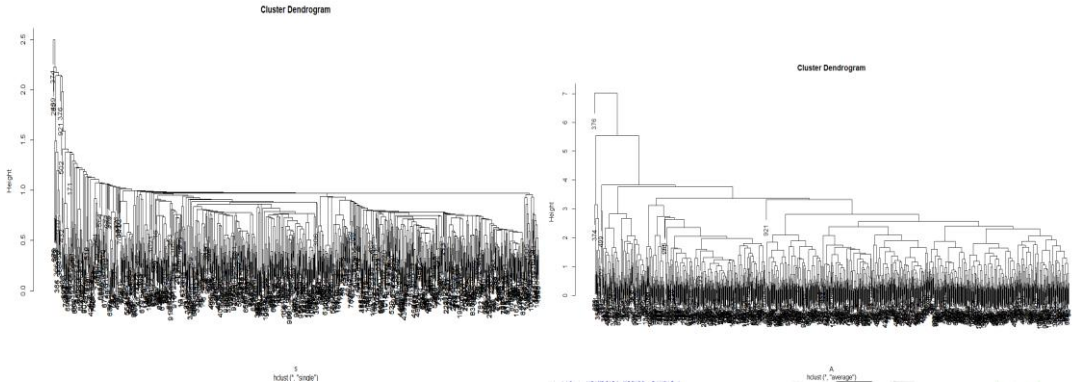


Fig:8. Single Linkage & Average Linkage

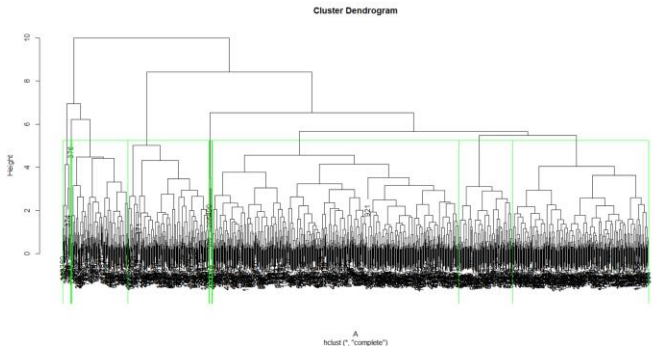


Fig:9. Complete Linkage

METHOD:2 USING PCA

Now PCA is performed for the dimension reduction and while looking in to the below figure cluster 4 makes much sense so I made k = 4.

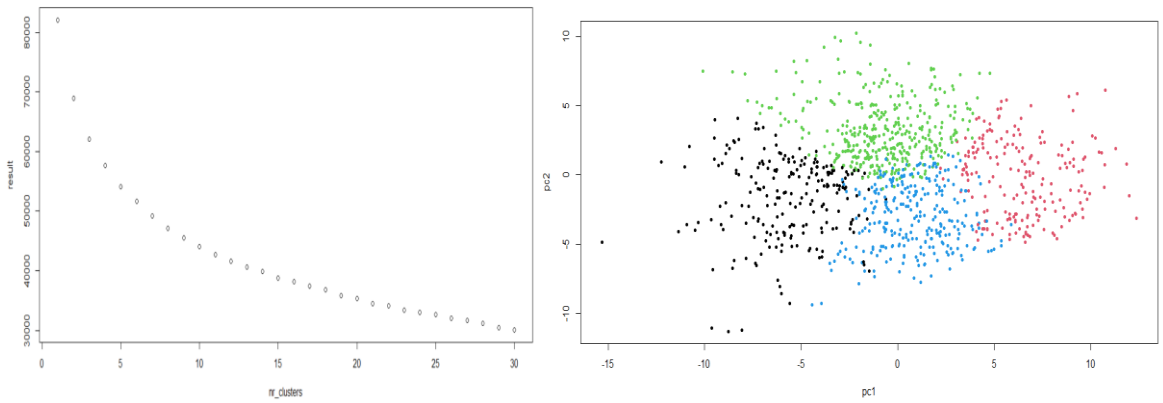
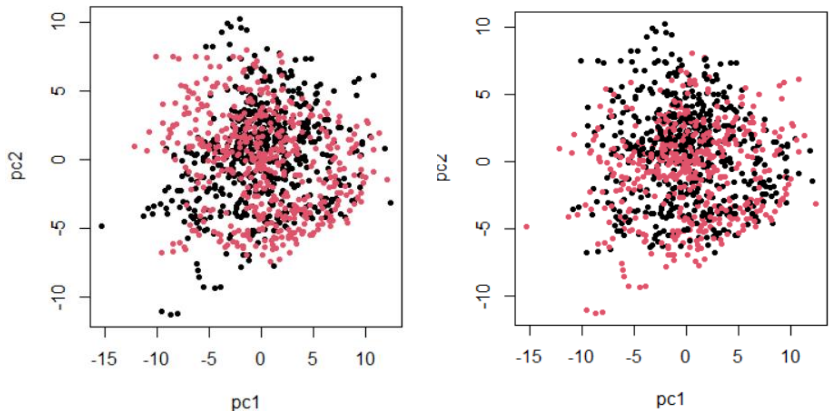


Fig:10. Elbow plot and K-means clustering with 4 clusters

While looking into the Fig:10 we can notice that it is well clustered for the pca1 and pca2. Let us try to explore further for the all four class.



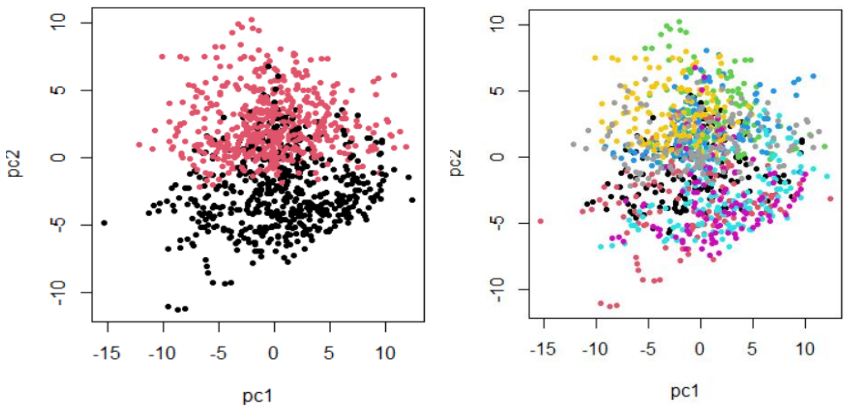
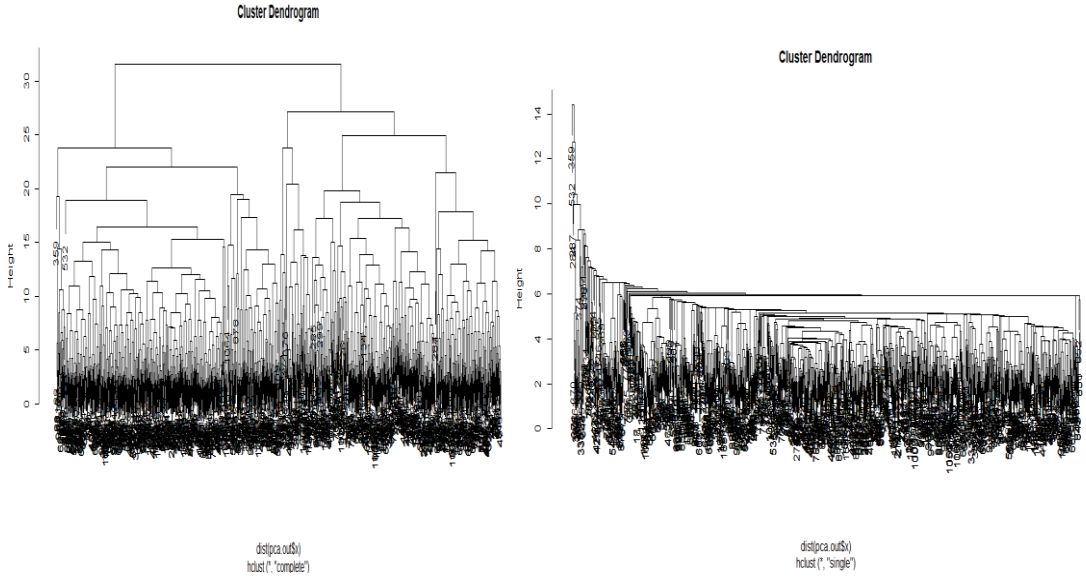


Fig:11 Clustering for four class variables

From the above four class it is understood that cluster is not well separated for the class variables Genotype, Treatment, Behavior, and Class.



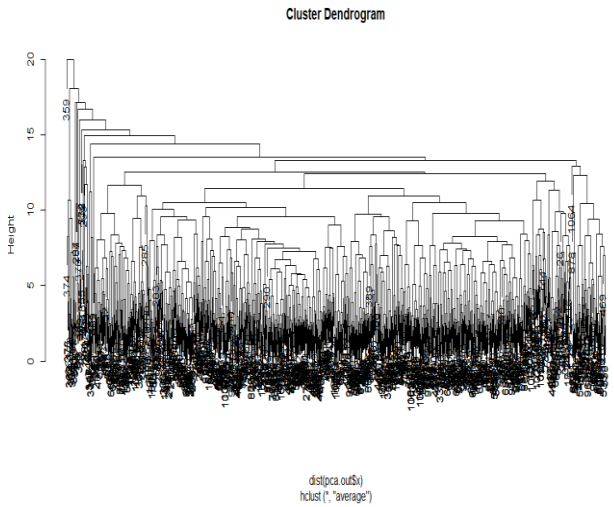


Fig:12 Hierarchical clustering using PCA

Then I used Hierarchical clustering through this by analyzing the result it is found that complete linkage method is well clustered when compared to single and average linkage using Euclidean distance.

CONCLUSION:

Task 1.

Wile looking in to results the decision tree method accuracy is improved in PCA method when compared to the normal method (without PCA) . Overall SVM with Polynomial kernel with tuned parameter performed better when compared to the Decision tree and SVM with Linear and Radial method. As support vectors are nonlinear Polynomial kernel performs better. If we look into the differences in binary and multiclass response binary class response performs better for some algorithms and multiple class performs better for some algorithms as per the Table1,2,3,4. Overall Random Forest model performs better for the binary class response (99%) and bagging model performs better for the multiple class response (99%). Finally, the model's accuracy is found to be dependent not only on the dataset with the number of fields employed, but also on the multiple training-testing split, parameters. It is critical to choose the right algorithm. Since a result, it is advised that the dataset be applied to a series of models and the best one be chosen using different training-testing partitions and by checking its confusion matrix.

Task 2:

In K-means clustering for both normal method and PCA, I tried to explore using 8 and 2 clusters I found 2 clusters performs better and well clustered. For Hierarchical clustering complete linkage is well clustered and I found optimal distance is 8 while looking into the dendrogram.