INFS4203 – Report

Part 1: Assumptions & Main functions used

Assumptions made

- The sample will be generated using all rows of the dataset
- For myClassification.r, the training/testing ratio will be 30/70

Main functions used

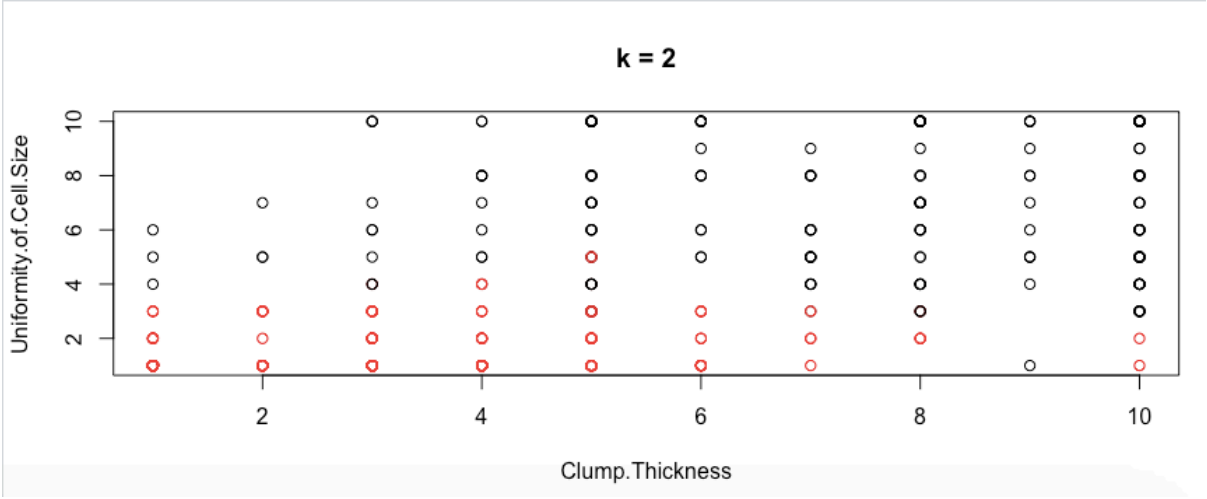
Function	Description	Assumption/Notes
plot()	Plot the classification tree,	Varying width/height for classification tree and dendograms.
set.seed()	Generate a random seed	Use last four digits of student id.
title()	Used to make a title for each plot	N/A
dev.off()	Closes off the current plot	N/A
sample()	Generate a sample	Assume all rows of the dataset are used for the sample for myClustering.
hclust()	Generate hierarchical clusters	No method provided for default hierarchical clustering.
cutree()	Cut a dendogram into different groups.	The k parameter represents the number of cuts that will be made to the dendogram. No h value will be provided.
ctree()	Generate a classification tree	Formula for generating the classification tree will use all variables in the dataset. No other parameters beside the formula and the dataset will be used.
knn()	Find the k-nn for a given training set of data.	No other arguments supplied beside the formula and training set.
predict()	Predict the class labels for the test dataset	Only the classification tree and the test features
as.matrix(table())	Make a matrix with the predicted class label and the actual class label	N/A

Part 2: Evaluation and Plots

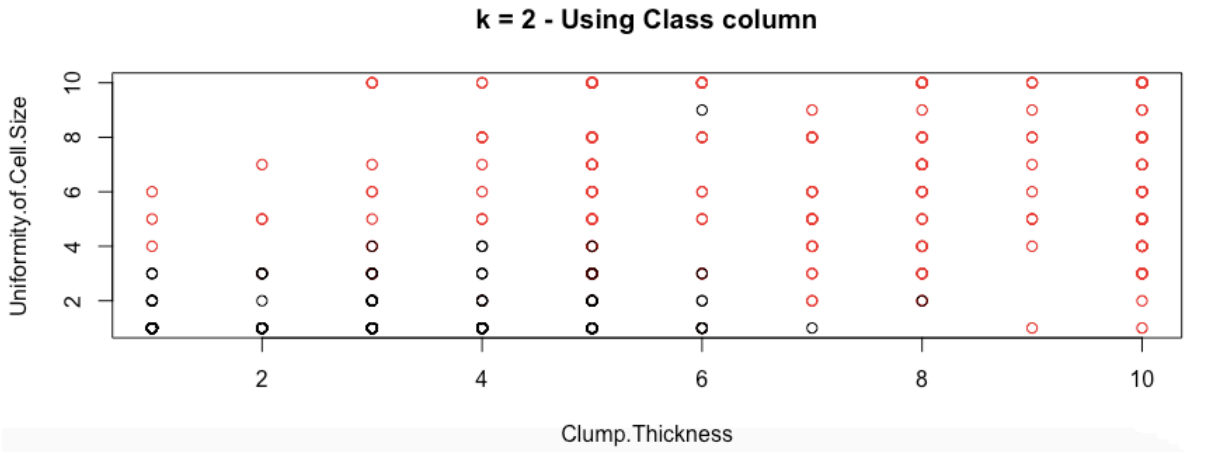
2.4 - Analysis

The following plots represent the k-2 clustering and clustering based on the Class column.

K-2 cluster



K-2 Using the class column



On observation, you can see that the K-2 cluster closely visually represent the benign vs malignant classes.

2.6 - Analysis

From the clustering performed in Ex 2.5 we can see the following result from the kmeans function.

Cluster	***betweenss***	***totalss***	**Ratio (my calculation)**
2-Cluster	23714	42544	0.558
3-Cluster	26784	42544	0.630
4-Cluster	27681	42544	0.650
5-Cluster	28901	42544	0.680

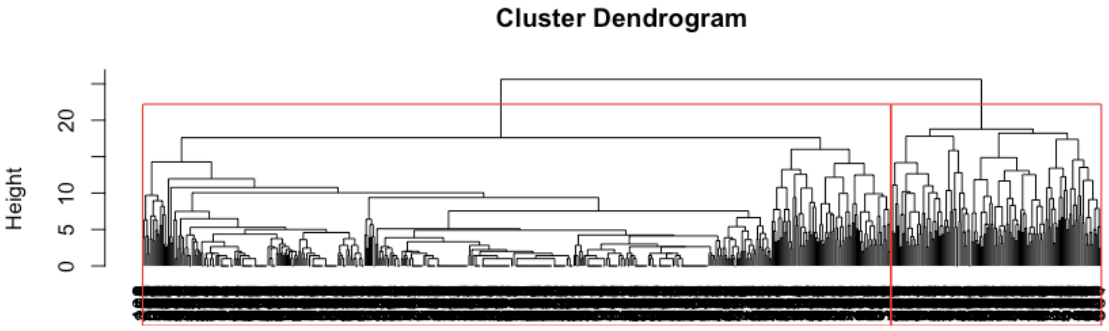
Good clustering is defined by low intra-cluster distance and high inter-cluster distance.
To determine good clustering in this dataset, I looked at the ratio of the SSE between the clusters (betweenss) to the total SSE for the entire dataset (totalss). The higher the ratio, the more defined the clustering is.

Observing the above table, we can see the ratio is highest when we have a k-value of 5. More generally, we can see that the difference between 2 clusters and 3 is where we see the greatest increase in ratio.

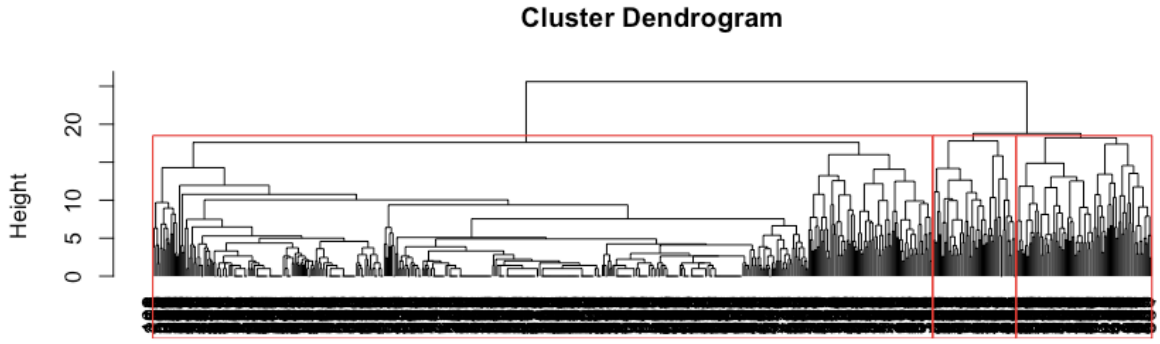
2.8 – Analysis

The following dendograms were created as a result of clustering from k=2 to 5.

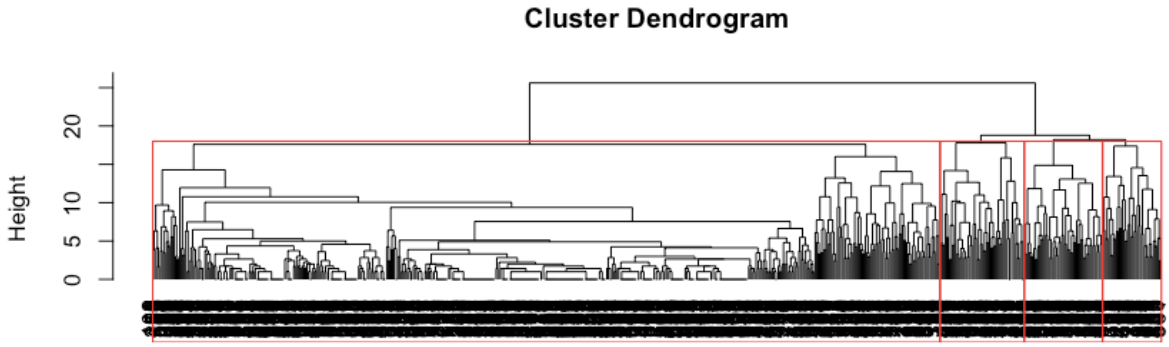
2-Cluster Dendogram



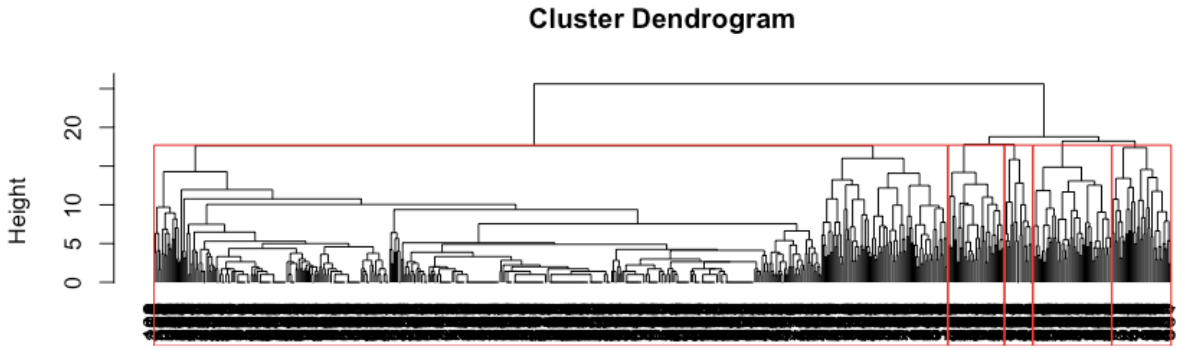
3-Cluster Dendogram



4-Cluster Dendogram



5-Cluster Dendogram

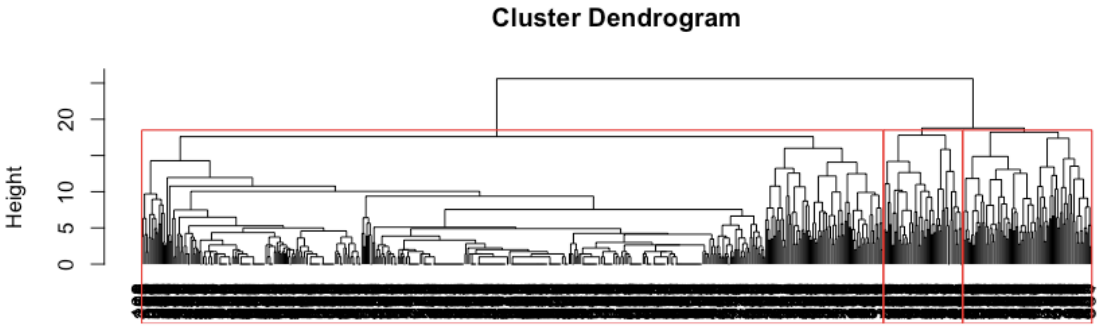


Based off the visualization of the dendogram, there are more than 2 subtypes of diseases. It is arguable that there are 5~6 observable clusters that could be partitioned as rectangles using this default clustering method.

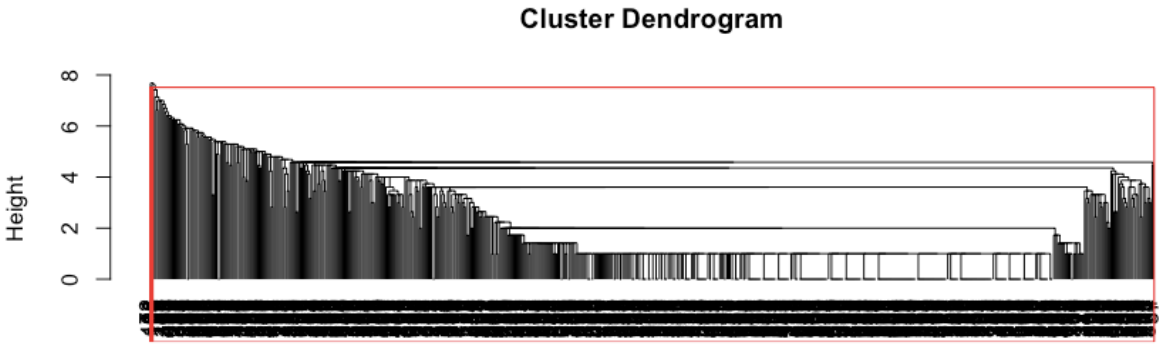
2.9 – Analysis

Observing the difference between single-linkage and complete-linkage we can see that the dataset is sensitive to the agglomeration method that is applied to it.

3-Cluster Complete-linkage

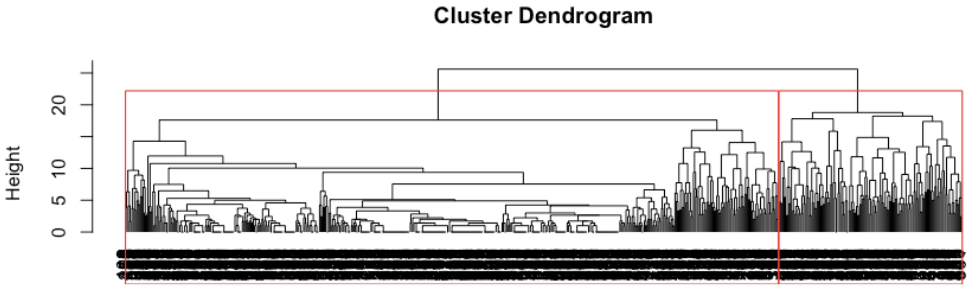


3-Cluster Single-Linkage

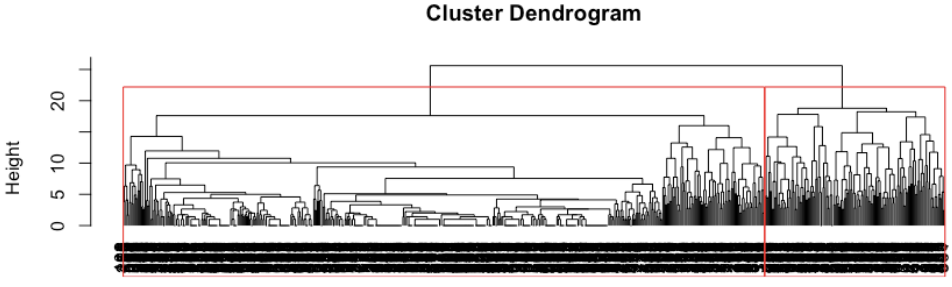


By default, the agglomerative method used is “complete”. This is evidenced by the output of running with default parameters and running with complete as the agglomerative method.

2-Cluster Default



2-Cluster Complete

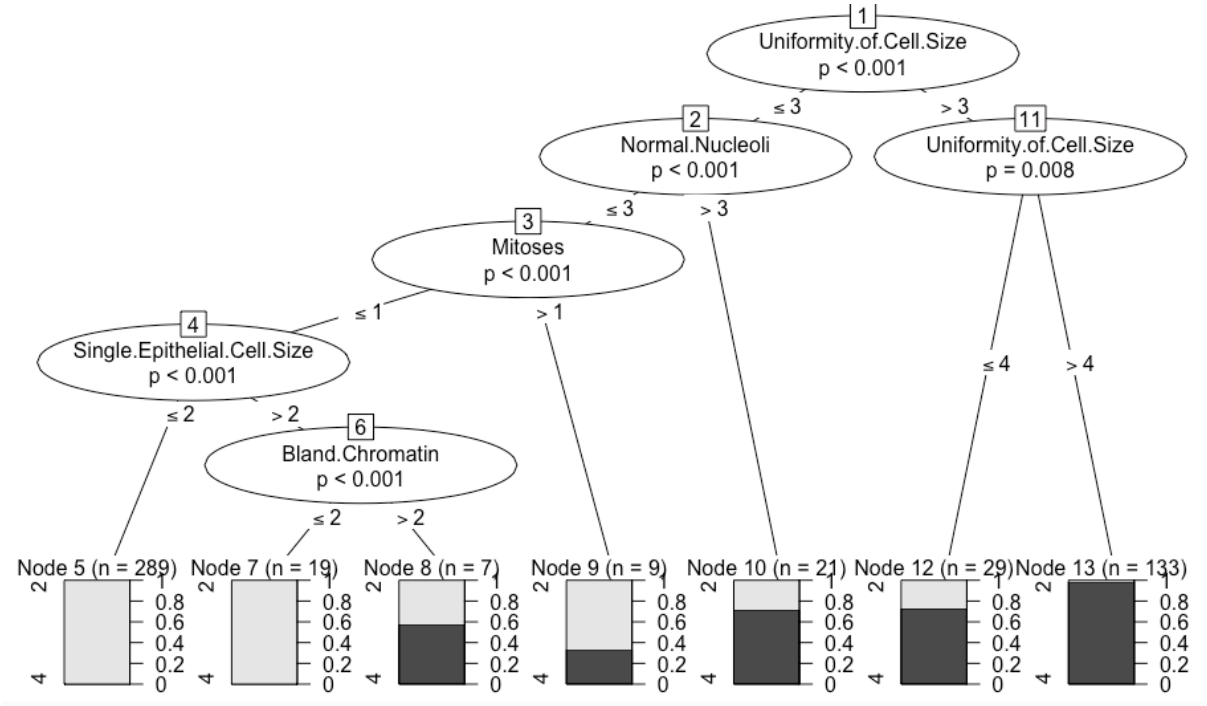


3.2 Analysis

The most significant variables are the leaf nodes that the closest to the root node in the tree. From the plot, it can be seen that the important variables are Uniformity of Cell Size, Normal Nucleoli, Mitoses, Single Epithelial Cell Size and Bland Chromatin. The others, that do not feature in the classification tree, can be labelled as not important variables for classification.

It can be inferred from the classification tree that classifying an observation as benign requires more evaluation in the classification tree. That means, the tree has to be traversed deeper in order to accurately classify a benign observation. Alternatively, for most cases only two evaluations of the classification tree are necessary in order to classify an observation as malignant.

Classification Tree



Summary

Variable (in order)	**Importance**
Clump thickness	Not Important
Uniformity of Cell Size	Important
Uniformity of Cell Shape	Not Important
Marginal Adhesion	Not Important
Single Epithelial Cell Size	Important
Bare Nuclei	Not Important
Bland Chromatin	Important
Normal Nucleoli	Important
Mitoses	Important

Accuracy

Accuracy
0.9602

Precision & Recall

Class Label	**Precision**	**Recall**
2 (Benign)	0.972	0.963
4 (Malignant)	0.939	0.953

3.3 Analysis

The accuracy/precision/recall from above can be improved by applying the following parameter to ctree_control: *ctree_control(mincriterion = 0.5)*. This improves the accuracy, precision and recall.

Accuracy

Accuracy
0.9659

Precision & Recall

Class Label	**Precision**	**Recall**
2 (Benign)	0.973	0.973
4 (Malignant)	0.953	0.953

3.4 Analysis

K-1 Nearest Neighbour

Accuracy

Accuracy
0.960

Precision & Recall

Class Label	**Precision**	**Recall**
2 (Benign)	0.956	0.981
4 (Malignant)	0.967	0.923

K-2 Nearest Neighbour

Accuracy

Accuracy
0.971

Precision & Recall

Class Label	Precision	Recall
2 (Benign)	0.973	0.982
4 (Malignant)	0.968	0.953

K-3 Nearest Neighbour

Accuracy

Accuracy
0.960

Precision & Recall

Class Label	Precision	Recall
2 (Benign)	0.964	0.972
4 (Malignant)	0.953	0.93

K-4 Nearest Neighbour

Accuracy

Accuracy
0.9765

Precision & Recall

Class Label	Precision	Recall
2 (Benign)	0.973	0.973
4 (Malignant)	0.953	0.953

K-5 Nearest Neighbour

Accuracy

Accuracy
0.960

Precision & Recall

Class Label	**Precision**	**Recall**
2 (Benign)	0.964	0.972
4 (Malignant)	0.953	0.938