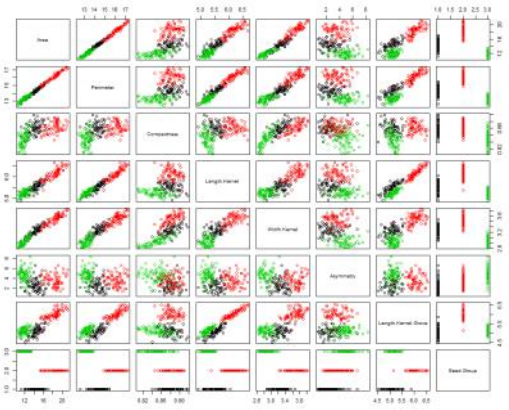
Clustering different varieties of wheat

Saturday, March 7, 2020 7:53 PM

Data Source : Seeds.txt

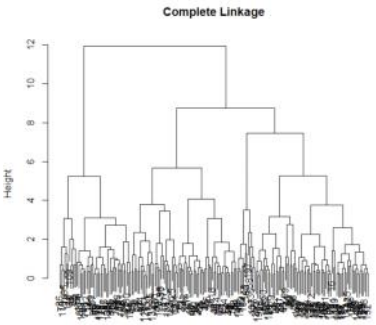
Observation:

- The data contains 199 records and 7 predictors. The predictors are the geometric properties of kernel belonging to three different classes.
- The highly correlated attributes are Area, perimeter, length of kernel, width of kernel and length of kernel groove. The expectation is that if one of these attribute value is high, the other will also be high for that wheat grain



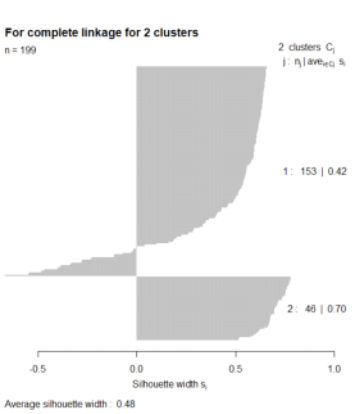
A) Cluster the data based on single-linkage, average linkage, and complete-linkage agglomerative hierarchical clustering. Decide on the groupings, and justify it, for all three methods. The justification should be based on a measure (you select which) that we learned in class. Which method "performed" the best and which method performed the worst? Was the result in line with your expectations?

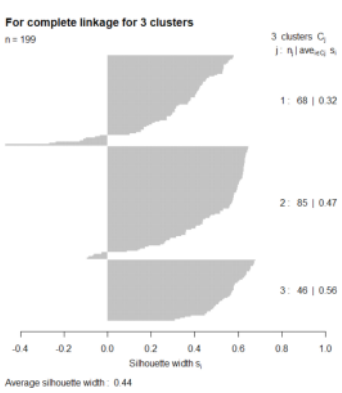
Complete Linkage

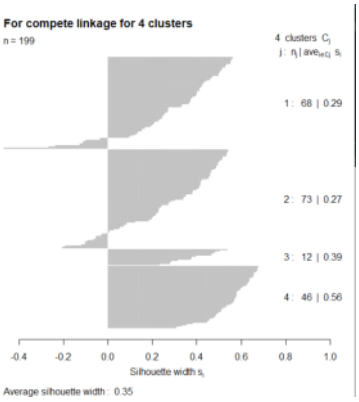


Analyzing the silhouette plot for different groupings to decide on the grouping
The analysis is based on the Silhouette plot for 2, 3 and 4 clusters for seed group shown below:

- For 2 cluster case, one of the cluster size is huge and we see significant negative silhouette width for the huge cluster. Thus making 2 clusters does not look like a good idea.
- For 4 cluster case, there is negative silhouette width as well as one of the cluster is extremely small.
- For 3 cluster case, the 3 clusters are of comparable size and negative silhouette width is the least out of the different choices. Taking 3 clusters looks like a good idea.







Cutting the dendrogram when there are 3 groupings and compared the clusters formed with the actual cluster results
The comparison

ct	A	B	C
1	46	22	0
2	20	0	65
3	0	46	0

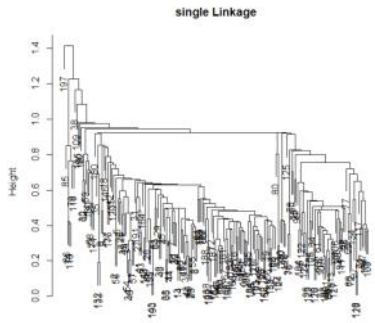
Results

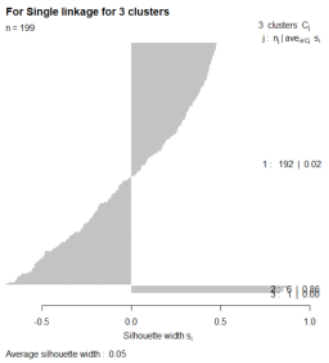
Misclassification Rate : 21.1%

2/3rd of the Wheat variety A is in one single cluster but 1/3rd is misclassified. Wheat Variety B and C have been completely misclassified.
Using Complete linkage for agglomerative hierarchical clustering does not give good clusters

Single Linkage

For 3 clusters using single linkage, most of the seeds have become a part of one cluster. The other two clusters have barely got any representation.
This is the disadvantage of single linkage as it tends to yield extended clusters to which single leaves are fused one by one.





Results

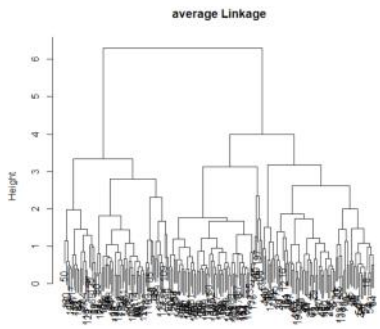
As expected the clustering results are very bad with single linkage. All the seeds have been clustered as one type

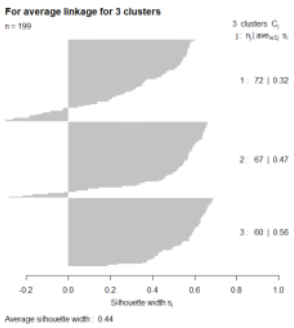
ct	A	B	C
1	66	62	64
2	0	6	0
3	0	0	1

Misclassification Rate : 63.3%

Average Linkage

- Average linkage gives a balanced dendrogram with a good representation in all three clusters.
- The Silhouette plot shows the error as being very limited on the negative width side





Result

As expected. Average Linkage has shown the best results with respect to classification.
There are 6 misclassifications for variety A, 4 misclassification for B and 8 misclassification for type C
Misclassification Rate : 10.55%

ct	A	B	C
1	60	4	8
2	3	64	0
3	3	0	57

Inference:

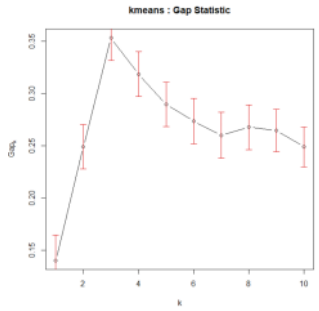
Average Linkage Hierarchial Clustering is performing better than Complete and Single linkage.

b) Cluster the data based on K-means or K-medoids. Use an analytical technique to justify your choice in "k". How did the performance compare to the hierarchical clustering of part a? Which did you feel was a better method for this data?

Choice of Analytical Technique

- Plotting Gap Statistic to choose the number of clusters.
- The gap statistic compares the total intra-cluster variation for different values of k with their expected values under null reference distribution of the data (i.e. a distribution with no obvious clustering)
- The more the gap statistic value, we can say that when the take this particular k as the number of cluster, we get the maximum gain from clustering.

Kmeans Clustering



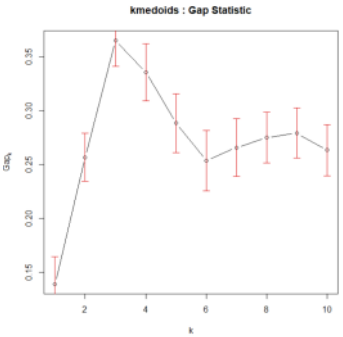
Results :

Misclassification rate : 39.19%

```
   A  B  C
1 53  0 65
2 13 68  0
```

Kmedoids Clustering

- Initially two random data points are choisen as centroids and points closer to each are assigned to them.
- This step is similar to the intial step in Kmeans. Kmedoids takes the data point nearest to the centroid of the cluster as the new central point for the cluster in the next step and this process keeps repeating.



Results

- Based on the above plot, the highest gap statistic value is maximum for 3 clusters. Let's go ahead with 3 clusters.
- 3 clusters seems to perform well.
- **Misclassification Rate = 10%**
- The result is as per our expectation as we know that there are 3 varieties of wheat types in our data.

```
   A  B  C
1  8  0 64
2  1 59  0
3 57  9  1
```

Inference : Performance Comparison

Clustering Technique	Misclassification Rate
Kmedoids	10%
Average Linkage - Hierarchical	10.55%
Complete Linkage - Hierarchical	21.10%
Kmeans	39.90%
Single Linkage - Hierarchical	63.30%

Kmedoids and Hierarchical clustering with average linkage have the best performance in clustering.