Introduction to Clustering

Prof. Sharon McNicholas

STATS 780/CSE 780

Introduction

- We move from classification to clustering today.
- In clustering applications, all of the observations are unlabelled (or treated as such).
- Hierarchical clustering is a famous approach, and agglomerative hierarchical clustering is quite popular.

Agglomerative Hierarchical Clustering

- A bottom-up approach.
- First, each observation is assigned to its own cluster.
- Then, the two closest clusters are joined into a single cluster.
- The process is repeated until there is only one cluster.

Two Decisions to Make

- How do we decide how close two observations are?
- How do we decide how close two clusters are?
- The first question is answered by choice of **dissimilarity**.
- The second question is answered by choice of **linkage**.

Some Dissimilarity Options

- Euclidean

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{m=1}^{M} (x_{im} - x_{jm})^2}.$$

- Manhattan

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{M} |x_{im} - x_{jm}|.$$

Some Linkage Options

- Complete

$$d(A, B) = \max_{\mathbf{x} \in A, \mathbf{y} \in B} d(\mathbf{x}, \mathbf{y}).$$

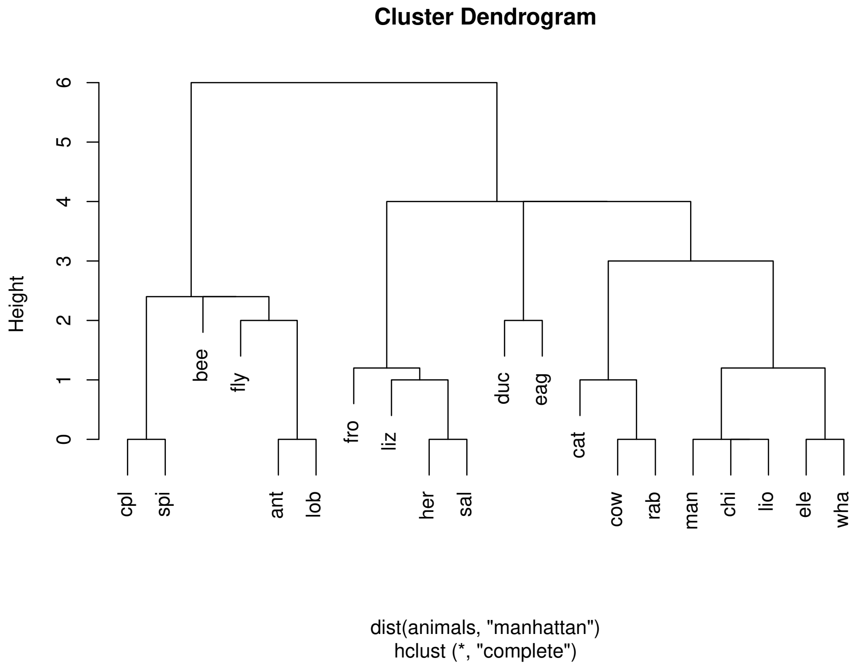
- Single

$$d(A, B) = \min_{\mathbf{x} \in A, \mathbf{y} \in B} d(\mathbf{x}, \mathbf{y}).$$

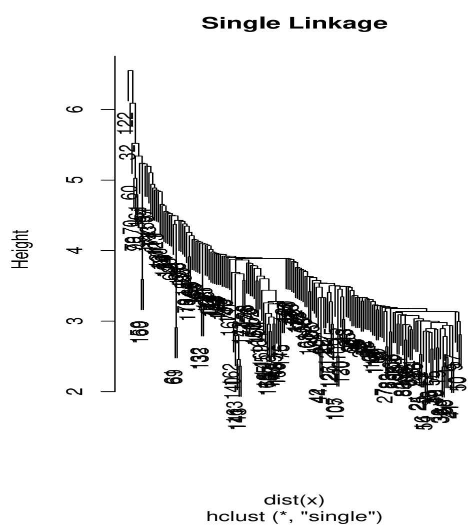
- Average

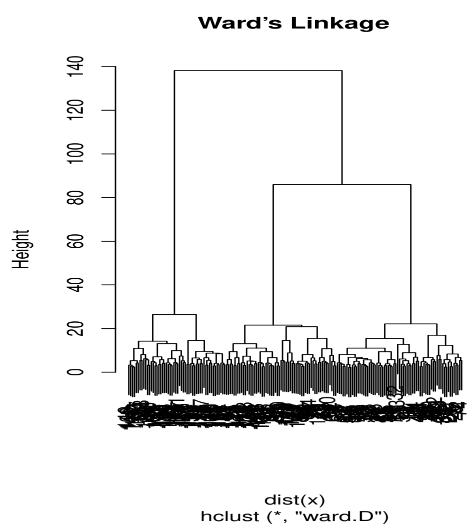
$$d(A, B) = \frac{1}{|A||B|} \sum_{\mathbf{x} \in A} \sum_{\mathbf{y} \in B} d(\mathbf{x}, \mathbf{y}).$$

Dendrogram



Chaining Example





Chaining

- Chaining is a problem that often occurs when single linkage is used.
- From the figure on the previous slide, you can see where the name comes from.
- In general, the chaining phenomenon leads to solutions that are of no practical use.
- Note the these dendrograms come from the `coffee` data set that are available in `pgmm` package.

Comments

- There is also the choice of how many clusters.
- While there are some “automatic” approaches, this is often done by eye.
- Hierarchical clustering solutions are naturally nested.
- For now, let’s look at some examples of agglomerative hierarchical clustering using `hclust()` in R.

Divisive Hierarchical Clustering

- We have seen that agglomerative hierarchical clustering is "bottom up".
- Divisive hierarchical clustering is "top down".
- Divisive hierarchical clustering starts with all observations in one cluster and then splits clusters to get new clusters.
- Divisive hierarchical clustering stops when each point is its own cluster.
- Let's look at some examples.

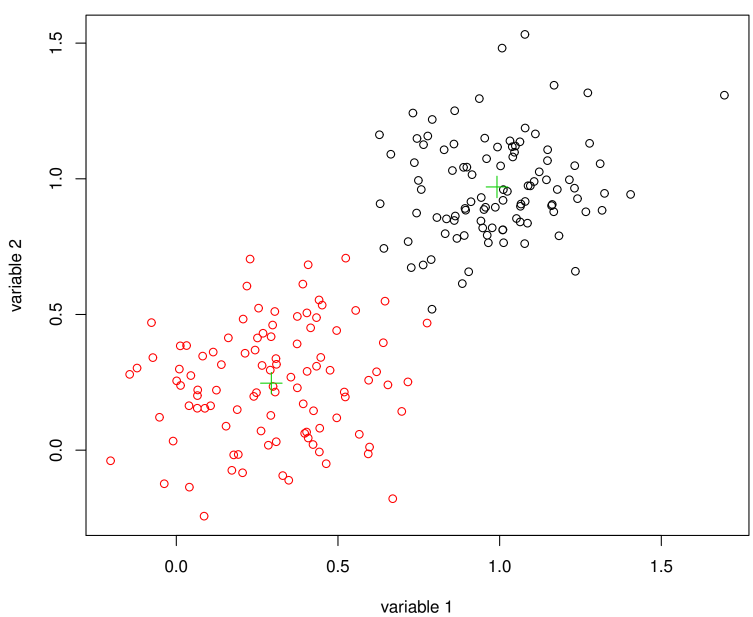
Partitioning Methods

- First, note that the nomenclature around clustering approaches is not universal.
- That said, a common view is that the alternatives to hierarchical clustering are partitioning methods.
- In a nutshell, partitioning methods cluster points around k cluster centres.
- Note that k, i.e., the number of clusters, is pre-specified.
- However, the locations of the centres are learned.

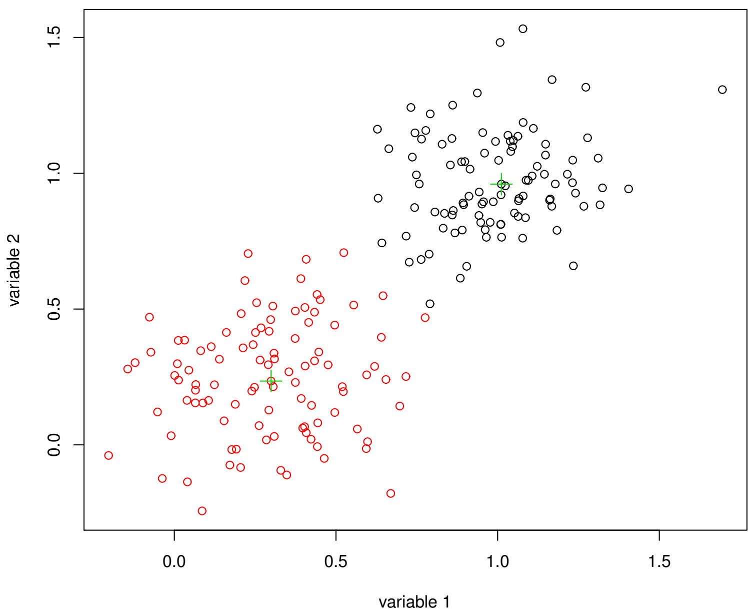
k-Means and k-Medoids Clusterting

- Choose clusters that minimize the distances between points within clusters and the cluster centres.
- There are k cluster centres.
- For k-means clustering, the cluster centres are means.
- For k-medoids clustering, the cluster centres are medoids.
- Note: the k here is the number of clusters whereas the k in kNN was the size of the neighbourhood.

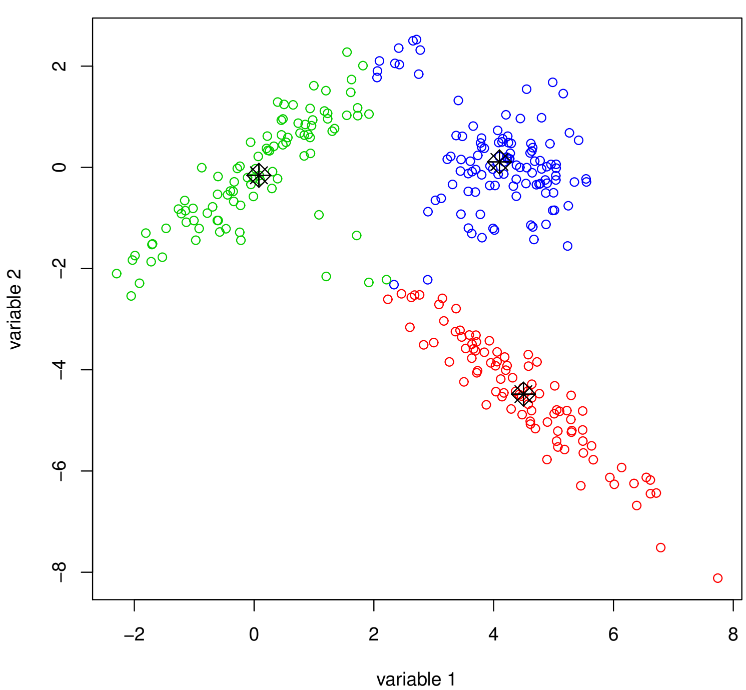
Bivariate Normal Example: k-Means



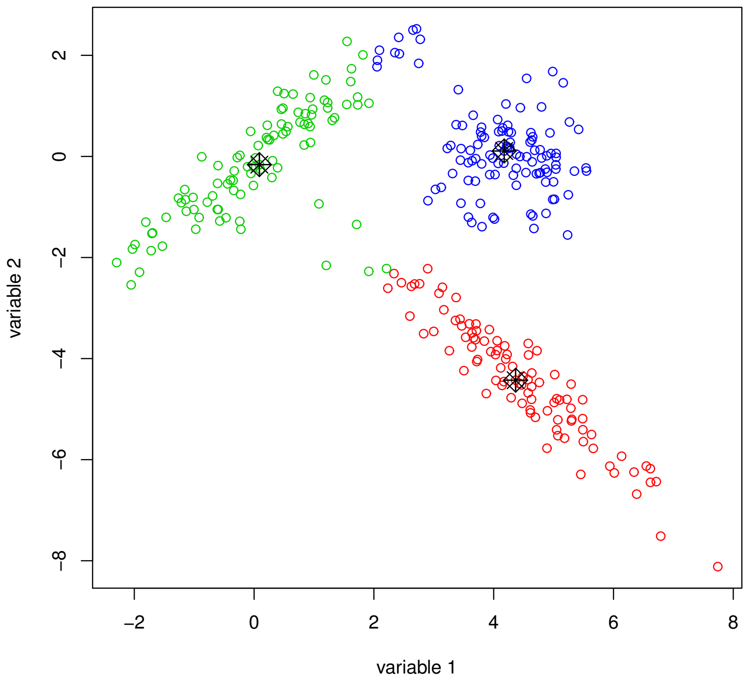
Bivariate Normal Example: k-Medoids



x2 Data: k-Means



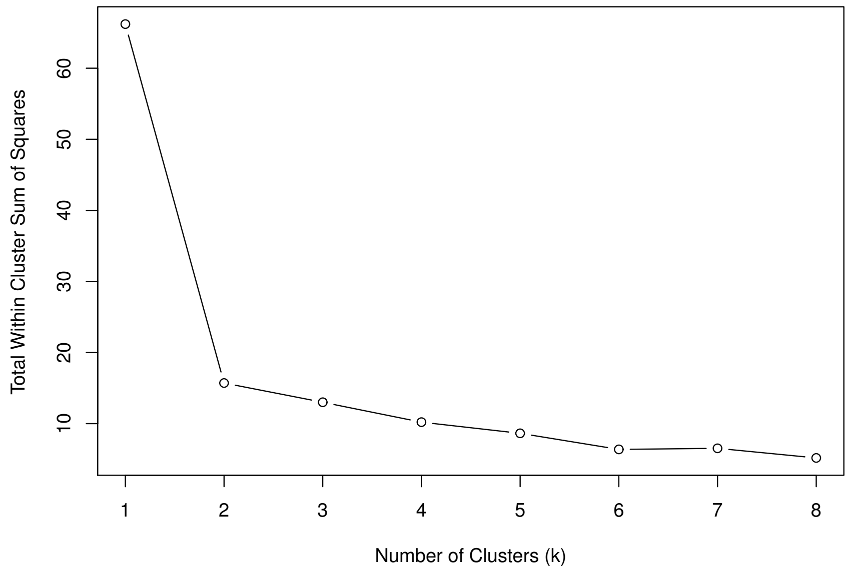
x2 Data: k-Medoids



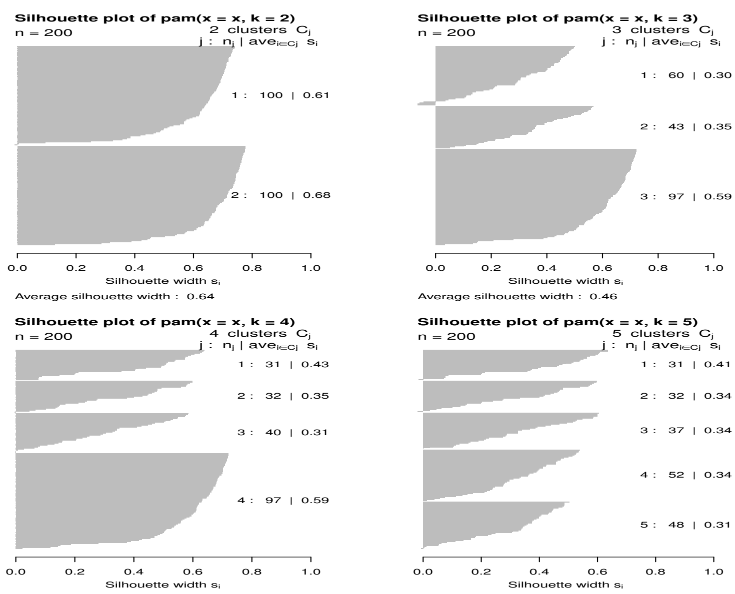
Choosing k

- The choice of the number of clusters, k, is very important.
- With hierarchical clustering, we choose the number of clusters after the algorithm is done, and the cluster solutions will necessarily be nested, e.g., the three-cluster hierarchical clustering solution can be thought of as merging two clusters in the four-cluster solution or splitting a cluster in the two cluster solution.
- For a given run of k-means or k-medoids, we must specify k up-front.
- We can run the k-means algorithm for different values of k and then choose the "best" k; however, the clustering solutions will not (in general) be nested.
- Same applies to k-medoids.

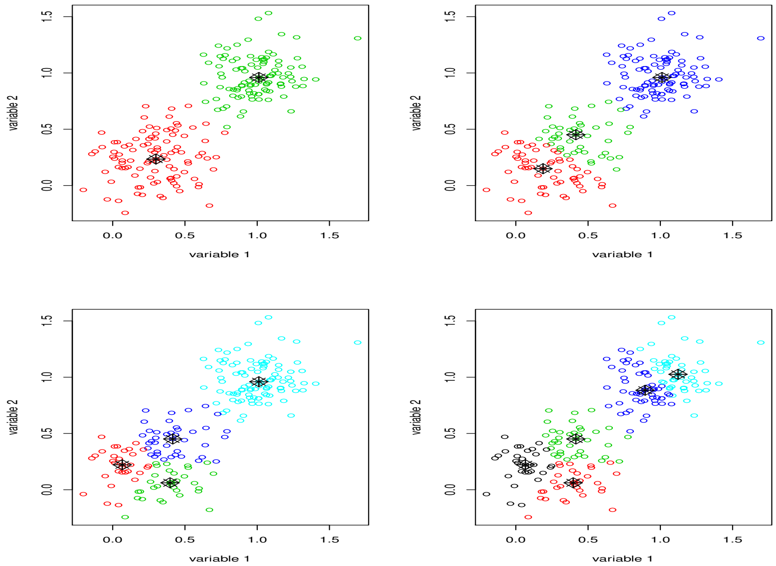
Choosing k for k-Means ("elbow")



Choosing k for k-Medoids (silhouette)



Choosing k for k-Medoids contd.



Comments

- The silhouette approach can also be used for other clustering methods.
- We have seen examples where k-means and k-medoids work well.
- However, problems can arise (e.g., non-spherical clusters and selecting the wrong k).
- Now, let's look at some more examples.
- Next, we will start to look at mixture model-based clustering.