

3 Clustering

Clustering refers to a broad set of techniques for finding *subgroups* in a data set.

We seek to partition observations into distinct groups so that

- observations within a group are similar ← need to define
- observations in different groups are dissimilar ← depend on domain!

For instance, suppose we have a set of n observations, each with p features. The n observations could correspond to tissue samples for patients with breast cancer and the p features could correspond to measurements collected for each tissue sample:

- clinical measurements, e.g. tumor stage or grade
- gene expression measurements.

↗ diverse in character

We may have reason to believe there is heterogeneity among the n observations.

e.g. different unknown subtype of cancer.

This is *unsupervised* because

We are trying to discover structure (distinct clusters) in the absence of a response.

VS.

Supervised problems we have the goal of prediction of a response.

Both clustering and PCA seek to simplify the data via a small number of summaries.

- PCA - finds a low dimensional representation of the observations that explain a good fraction of the variance.
- Clustering finds homogenous subgroups among observations

Since clustering is popular in many fields, there are many ways to cluster.

We will focus on 2 best-known clustering approaches.

- K -means clustering

Seeks to partition the observations into a pre-specified # of clusters.

- Hierarchical clustering

We do not know in advance how many clusters we want.

We obtain clusterings for $1, \dots, n$ # of clusters

\hookrightarrow can view in a tree-like visualization called a "dendrogram"

In general, we can cluster observations on the basis of features or we can cluster features on the basis of observations.

\downarrow
identify subgroups
among observations

\downarrow
identify subgroups
among the features.

We will focus on ①

But we can perform ② by transposing the data matrix.

$$\begin{matrix} X & \rightarrow & X^T \\ n \times p & & p \times n \end{matrix} \rightarrow \text{clustering.}$$

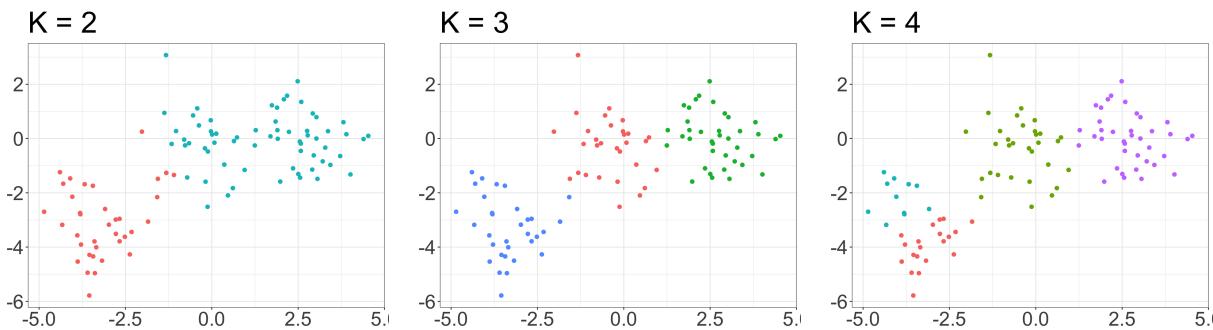
3.1 K-Means Clustering

Simple and elegant approach to partition a data set into K distinct, non-overlapping clusters.

We must first specify how many clusters K .

Then k-means assigns each observation to one of the K clusters.

e.g. clustering $n=100$ observations into K clusters using $p=2$ features.



The K -means clustering procedure results from a simple and intuitive mathematical problem. Let C_1, \dots, C_K denote sets containing the indices of observations in each cluster. These satisfy two properties:

e.g. if observation i is in cluster k , $i \in C_k$

$$1. \quad C_1 \cup \dots \cup C_K = \{1, 2, \dots, n\}$$

each observation belongs to one of the K clusters.

$$2. \quad C_k \cap C_{k'} = \emptyset \quad \forall k \neq k'$$

The clusters are non overlapping.

Idea: "good clustering" is one for which the within-cluster variation is as small as possible.

The *within-cluster variation* for cluster C_k is a measure of the amount by which the observations within a cluster differ from each other.

Call this $W(C_k)$.

Then we want to solve the problem:

$$\underset{C_1, \dots, C_K}{\text{minimize}} \left\{ \sum_{k=1}^K W(C_k) \right\}$$

we want to partition observations into K clusters s.t. total within-cluster variation is minimized.

To solve this, we need to define within-cluster variation.

Many ways we can do that.

Most common way: squared Euclidean distance:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i:j \in C_k} \sum_{j=1}^p (x_{ij} - \bar{x}_{ij})^2$$

\uparrow # obs in k^{th} cluster.

This results in the following optimization problem that defines K -means clustering:

$$\underset{C_1, \dots, C_K}{\text{minimize}} \underbrace{\left\{ \sum_{k=1}^K \frac{1}{|C_k|} \sum_{i:j \in C_k} \sum_{j=1}^p (x_{ij} - \bar{x}_{ij})^2 \right\}}_{\text{objective function}}$$

This is very difficult to solve exactly! $\approx K^n$ ways to partition n obs. into K clusters.

\downarrow "pretty good solution"

A very simple algorithm has been shown to find a local optimum to this problem:

1. randomly assign a number from 1 to K to each observation
these are initial cluster assignments for the observations

2. iterate until cluster assignments stop changing. \checkmark vector of the p feature

(a) for each of the K clusters compute cluster centroid \checkmark means for observation i^{th} each cluster (K).

(b) assign each observation to the closest centroid cluster.

\uparrow Euclidean distance.

Algorithm is guaranteed to decrease value of objective function at each step.

When cluster assignments stop changing this is a local minimum.

\hookrightarrow not necessarily global min \Rightarrow clustering depends on initial (random) cluster values (Step 1).

What to do? Run the algorithm multiple times from different initial configurations and choose clustering w/ smallest objective function.

Problem: We still must choose K ! More later...

3.2 Hierarchical Clustering

One potential disadvantage of K -means clustering is that it requires us to specify the number of clusters K . *Hierarchical clustering* is an alternative that does not require we commit to a particular K .

ahead of time

Hierarchical clustering also results in a tree-based representation of the observations.

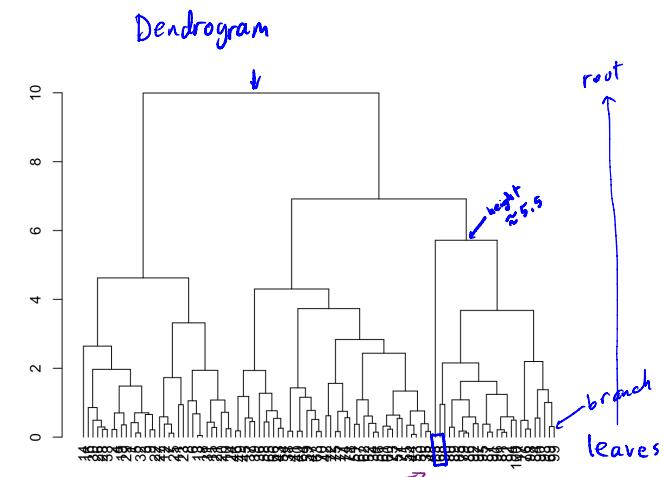
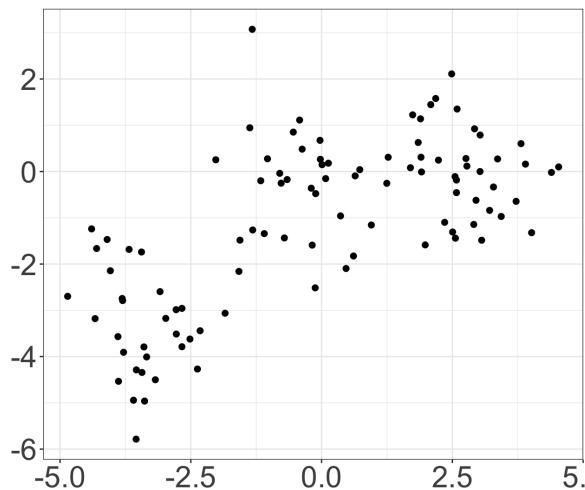
We will discuss *bottom-up* or "agglomerative" clustering.

Start with every observation in its own cluster and merge (fuse) clusters until all observations are in a single cluster (n clusters of size 1 \rightarrow 1 cluster of size n).

"bottom-up" refers to representation of clusters in tree diagram w/ leaves on the bottom.

3.2.1 Dendograms

Same simulated data as before: $n=100$ observations w/ $p=2$ features.



*branches = clusters w/ more than 1 observation
leaves = cluster w/ 1 observation
even though these observations are right next to each other on x-axis of dendrogram, they are quite different b/c height of first fusion.*

Each *leaf* of the dendrogram represents one of the 100 simulated data points.

As we move up the tree, leaves begin to fuse into branches, which correspond to observations that are similar to each other.

- as we move higher up the tree, branches fuse with other branches or w/ leaves.
- the lower a fusion occurs, the more similar the observations are.
- observations that fuse high up in the tree can be quite different

More precisely:

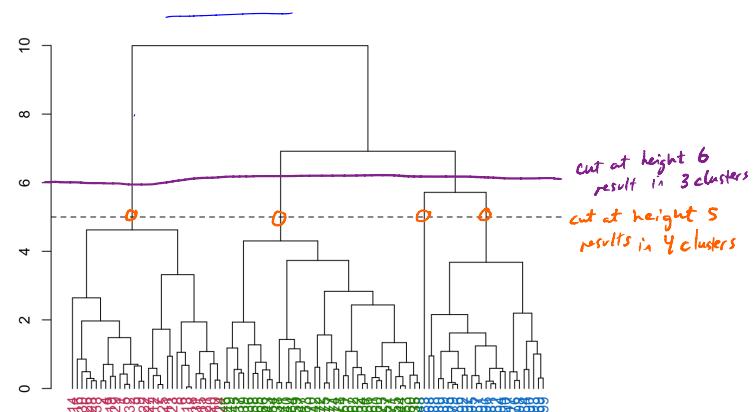
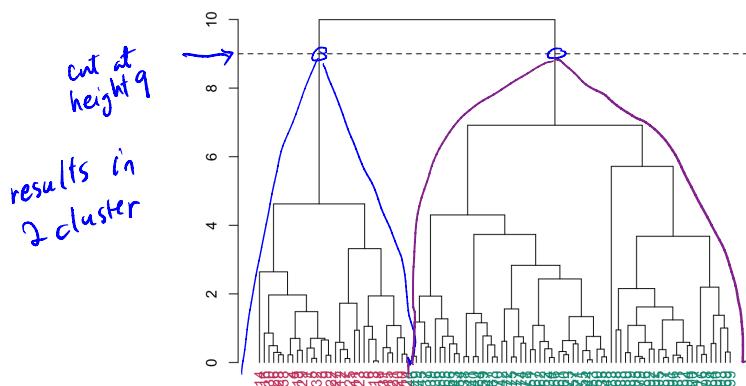
For any two observations, we can look for the point in the tree where branches containing those two observations are first fused.

The height of this fusion indicates how different they are!

We draw conclusions about similarity of two observations based on the location on the vertical axis where branches containing those observations are first fused.

How do we get clusters from the dendrogram?

We make horizontal "cuts" across the dendrogram.



We can cut at a height that corresponds to $1, \dots, n$ clusters. (i.e. height of cut is similar to K in K-means).
 \Rightarrow A single dendrogram can be used to obtain any number of clusters!

In practice: people inspect dendrogram and choose where to cut based on heights of fusion and # cluster results (subjective).

The term *hierarchical* refers to the fact that clusters obtained by cutting the dendrogram at a given height are necessarily nested within the clusters obtained by cutting the dendrogram at a greater height.

This hierarchical assumption may or may not be realistic.

e.g. suppose have group of observations, 50-50 split M/F and evenly split American, Japanese, French.

Maybe 2 clusters results in clustering by sex ↩
 3 clusters results in a clustering by nationality ↩
 ↩ not nested.

3.2.2 Algorithm

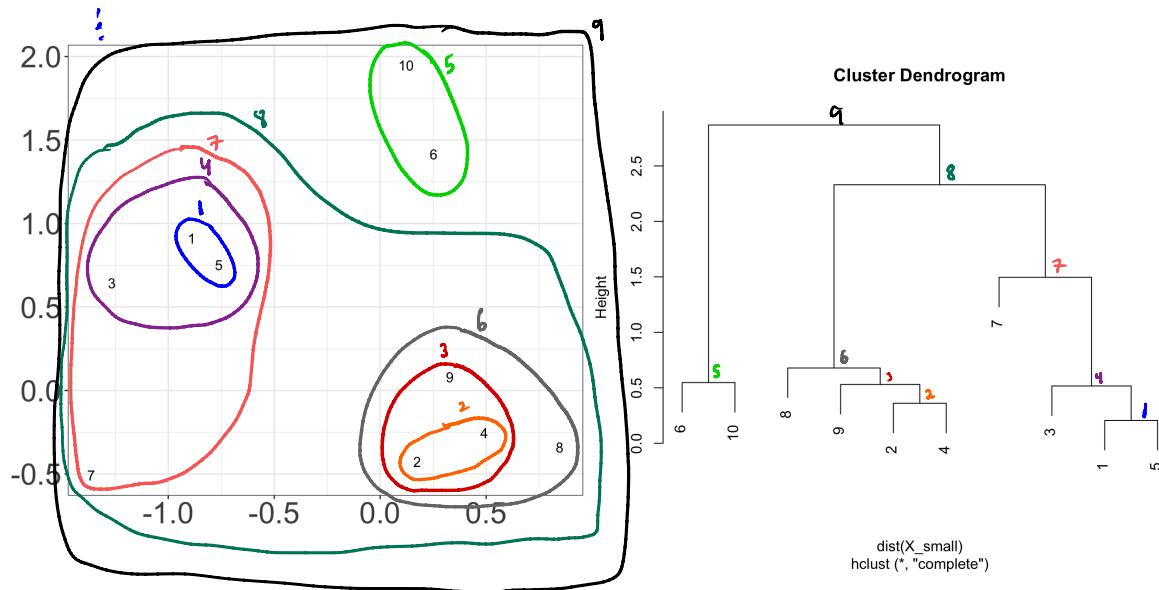
First, we need to define some sort of *dissimilarity* metric between pairs of observations.

Most often Euclidean distance is used.

Then the algorithm proceeds iteratively.

start at the bottom of dendrogram each obs. is in own cluster

find 2 most similar clusters and fuse.



More formally,

One issue has not yet been addressed.

How do we determine the dissimilarity between two clusters if one or both of them contains multiple observations?

1.

2.

3.

4.

3.2.3 Choice of Dissimilarity Metric

3.3 Practical Considerations in Clustering

In order to perform clustering, some decisions should be made.

-
-
-

Each of these decisions can have a strong impact on the results obtained. What to do?