Ch11-B Clustring Analysis

Contents

Textbook: James et al. ISLR 2ed.

11B Subsections

[ToC]

B.1 Clustering Analysis

- Find homegeneous subgroups among the observations.
- K-means clustering
- Hierarchical clustering

B.2 K-means Clustering

- Must choose k first.
- Good clustering is one for which the within-cluster varaiation is small
- Must choose a measure for within-cluster variation $W(C_k)$.
- Typically squared Euclidian distance

$$W(C_k) = \frac{1}{N_k} \sum_{i,j \in C_k} \sum_{\ell=1}^{p} (x_{i\ell} - x_{j\ell})^2$$

  where N_k is a number of obs in cluster k.

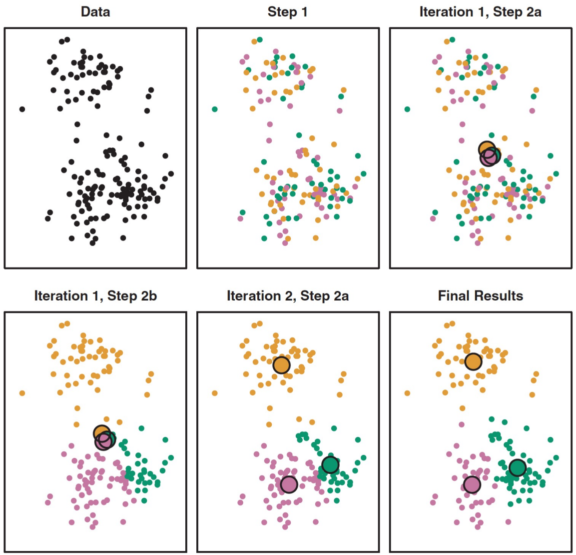
- Want to minimize $W(C_1) + \cdots + W(C_k)$.
- Cluster Centroid: mean observations in the cluster.

B.3 K-means Cluster Algorithm

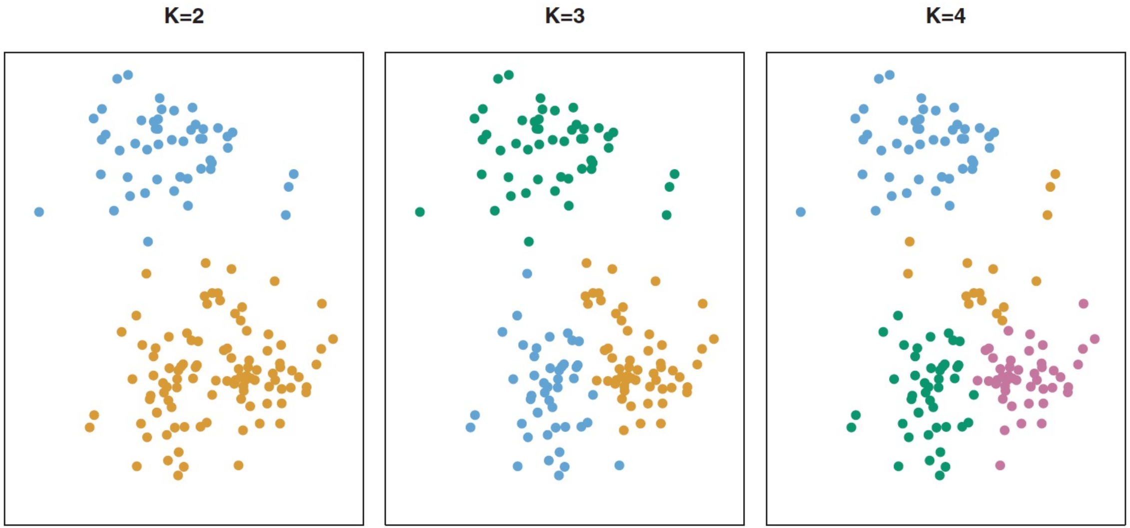
1. Randomly assign each of the obs. to a cluster (1-K).
2. For each of the K clusters, compute the cluster centroid.
3. Assign each observation to the cluster whose centroid is closest.

    - Initial assignment is random
    - Have to choose k.
    - Each iteration, $W(C_1) + \cdots + W(C_k)$ will be reduced (local minimum. local given the initial position.)

B.4 Ex:


Data
Step 1
Iteration 1, Step 2a
Iteration 1, Step 2b
Iteration 2, Step 2a
Final Results

B.5 Ex:



B.6 Pros and Cons

Pros

- Relatively simple to implement.
- Scales to large data sets.
- Guaranteed convergence.
- Can warm-start the positions of centroids.

Cons

- Choosing k
- Being dependent on initial values (should repeat couple of times)
- hard to find clusters of uneven sizes and density. (does have some generalization)
- Centroid is influenced by outliers.
- Scaling with number of dimensions. (curse of dimensionality)

B.7 Hierarchical Clustering

- No need to choose k apriori.
- Produces chart called dendrogram.
- You can decide on the number of clusters looking at the dendrogram afterwards
- Bottom-up (agglomerative) clustering

B.8 Hierarchical Clustering

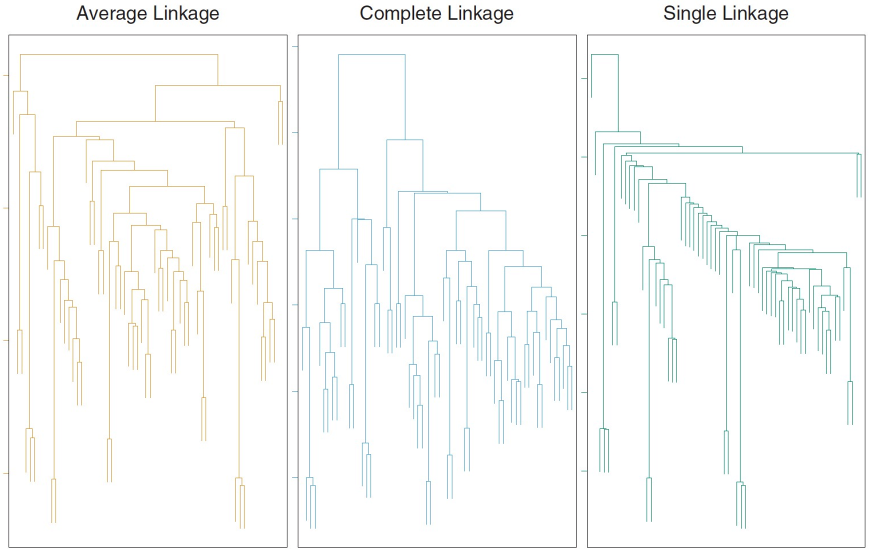
1. Start as each observation being a cluster. There are n clusters.

2. Look at intercluster dissimilarity measure (ICD) of all possible pairs.

3. Merge the two that has least ICD. Repeat.

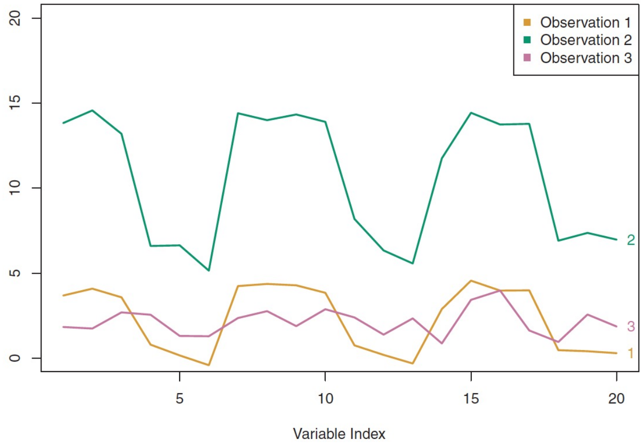
B.9 Intercluster Dissimilarity Measure

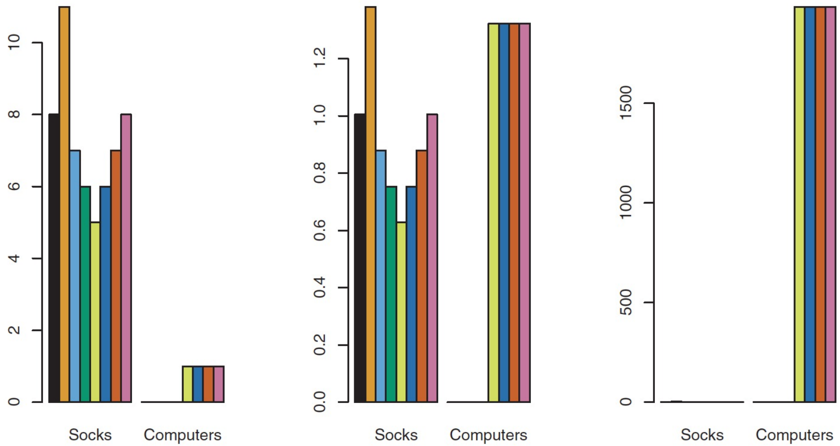
- Usually euclidian distance.
- Intercluster Dissimilarity Measure can take many forms (Linkage Function)
  - Complete Linkage: Look at all pairwise dissimilarities in A and B. Take max.
  - Single Linkage: Take min.
  - Average Linkage: Take average.
  - Centroid Linkage: Dissimilarities betwen centroid of A and centroid of B.


Average Linkage
Complete Linkage
Single Linkage

B.10 Online Retailer Example

- Scaling Issue (Scale or not?)
- Correlation-based dissimilarity measure can be used


Observation 1
Observation 2
Observation 3
0
5
10
15
20
5
10
15
20
2
3
1
Variable Index



0
2
4
6
8
10
Socks
Computers
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Socks
Computers
0
500
1000
1500
Socks
Computers

Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani