Clustering

Machine Learning

Data we will work with

- Customer Spend Data
  - AVG_Mthly_Spend: The average monthly amount spent by customer
  - No_of_Visits: The number of times a customer visited in a month
  - Item Counts: Count of Apparel, Fruits and Vegetable, Staple Items purchased

	Cust_ID	Name	Avg_Mthly_Spend	No_Of_Visits	Apparel_Items	FnV_Items	Staples_Items
1	1	A	10000	2	1	1	0
2	2	B	7000	3	0	10	9
3	3	C	7000	7	1	3	4
4	4	D	6500	5	1	1	4
5	5	E	6000	6	0	12	3
6	6	F	4000	3	0	1	8
7	7	G	2500	5	0	11	2
8	8	H	2500	3	0	1	1
9	9	I	2000	2	0	2	2
10	10	J	1000	4	0	1	7

- Can we cluster similar customers together?

Connectivity Based: Hierarchical Clustering

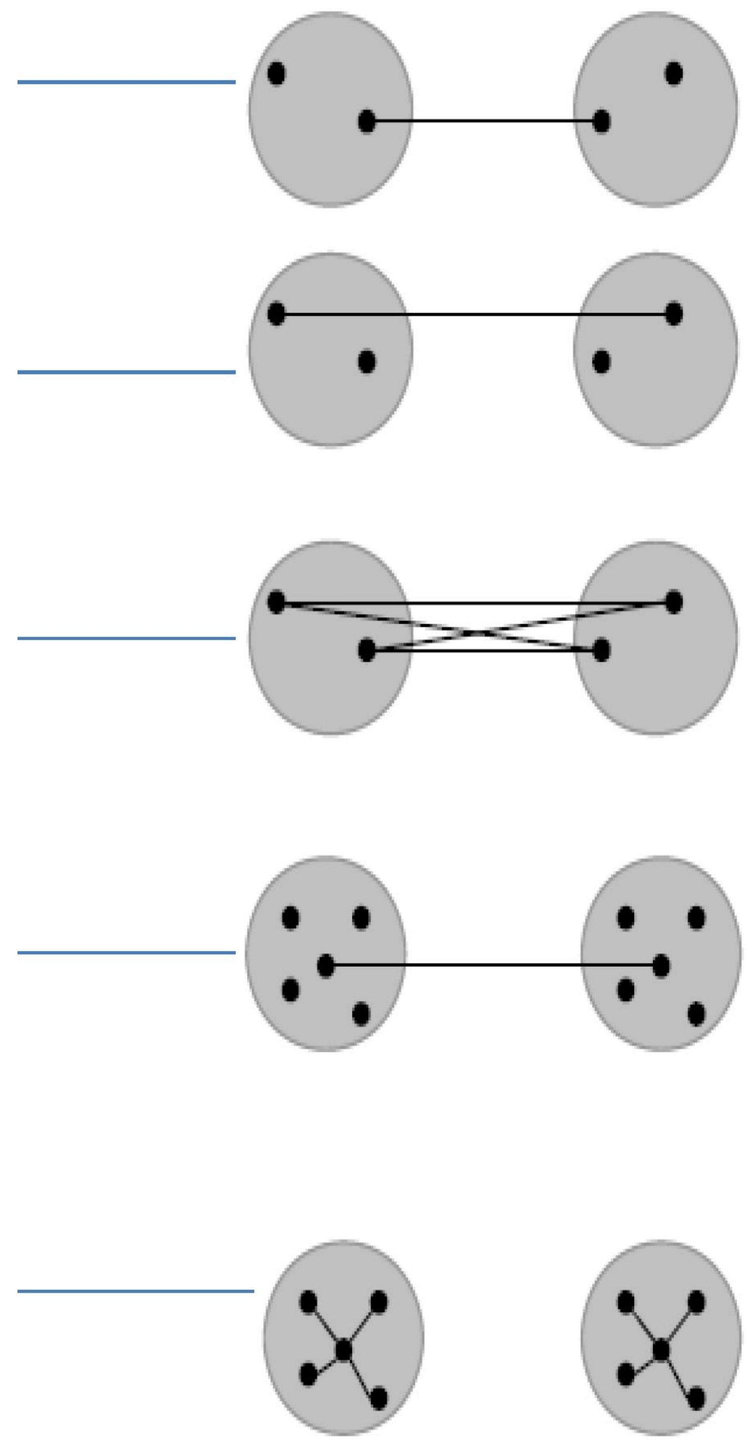
- Hierarchical Clustering techniques create clusters in a hierarchical tree like structure
- Any type of distance measure can be used as a measure of similarity
- Cluster tree like output is called Dendogram
- Techniques either start with individual objects and sequentially combine them (Agglomerative), or start from one cluster of all objects and sequentially divide them (Divisive)

Agglomerative

- Starts with each object as a cluster of one record each
- Sequentially merges 2 closest records by distance as a measure of similarity to form a cluster.
- How would we measure distance between two clusters?

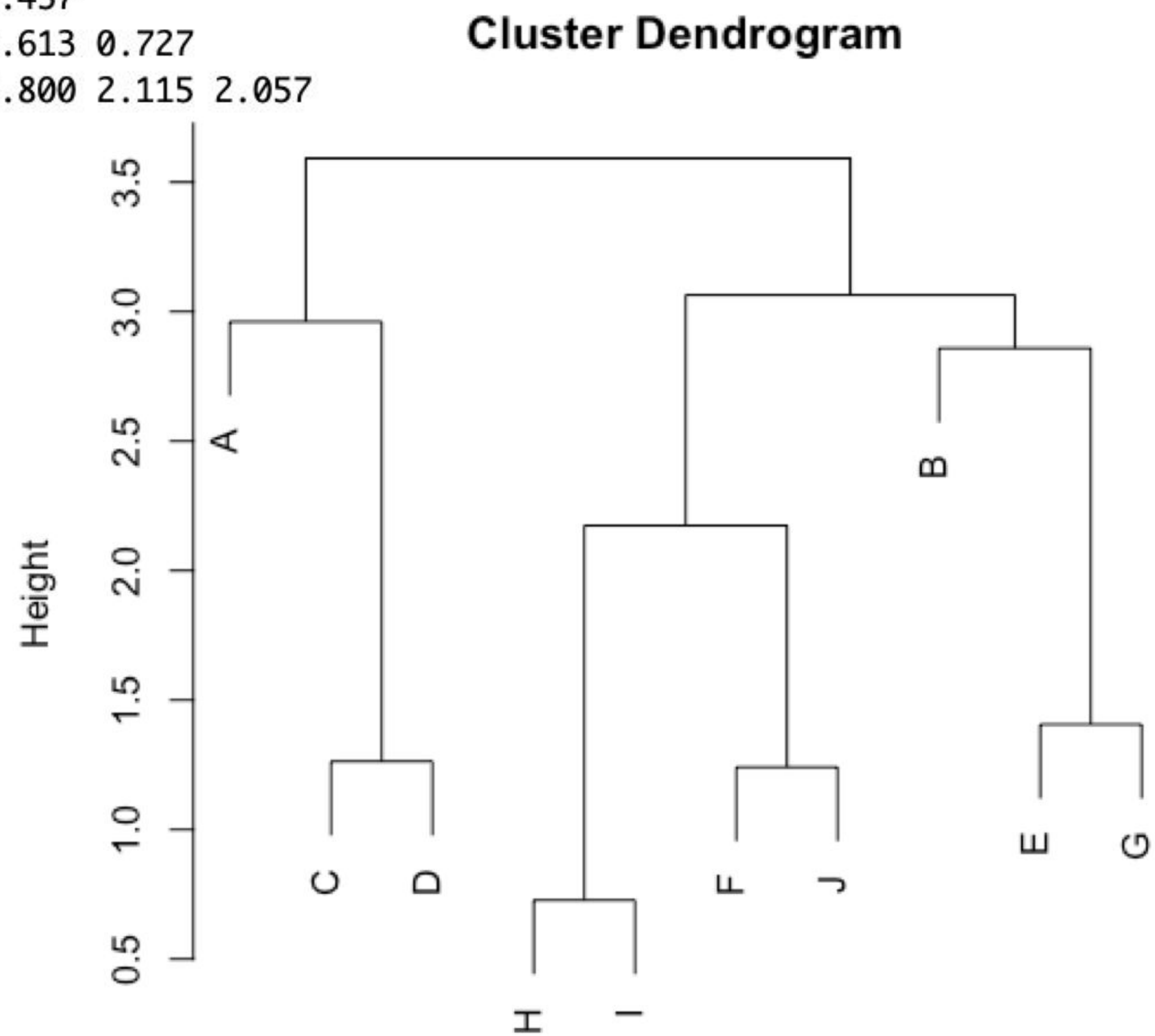
Distance between clusters

- Single linkage – Minimum distance or Nearest neighbor
- Complete linkage – Maximum distance or Farthest distance
- Average linkage – Average of the distances between all pairs
- Centroid method – combine cluster with minimum distance between the centroids of the two clusters
- Ward’s method – Combine clusters with which the increase in within cluster variance is to the smallest degree



Distance between objects

	1	2	3	4	5	6	7	8	9
2	4.252								
3	3.411	3.838							
4	2.512	3.473	1.264						
5	4.268	2.697	2.922	3.204					
6	3.980	2.208	3.579	2.853	3.431				
7	4.378	3.021	3.384	3.345	1.406	3.171			
8	3.396	3.603	3.663	2.927	3.244	2.350	2.457		
9	3.534	3.395	4.054	3.213	3.482	2.175	2.613	0.727	
10	4.550	2.967	3.591	3.041	3.408	1.241	2.800	2.115	2.057



Centroid based: K-Means Clustering

- K-Means is probably the most used clustering technique
- Aims to partition the n observations into k clusters so as to minimize the within-cluster sum of squares (i.e. variance).
- Computationally less expensive compared to hierarchical techniques.
- Have to pre-define K, the no of clusters

Lloyd's algorithm

1. Assume K Centroids

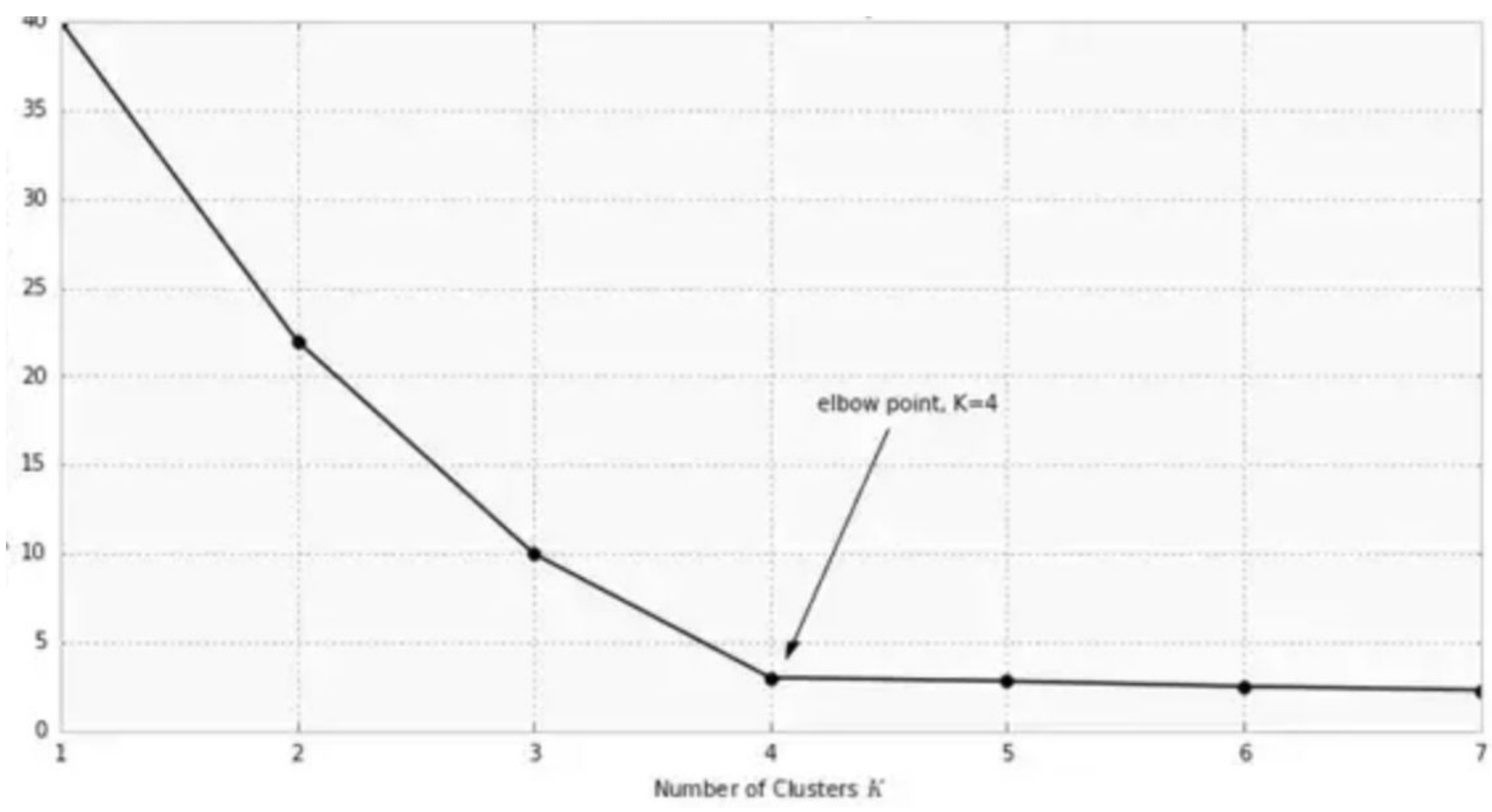
2. Compute Squared Eucledian distance of each objects with these K centroids. Assign each to the closest centroid forming clusters.

3. Compute the new centroid (mean) of each cluster based on the objects assigned to each clusters.

4. Repeat 2 and 3 till convergence: usually defined as the point at which there is no movement of objects between clusters

Choosing the optimal K

- Usually subjective, based on striking a good balance between compression and accuracy
- The "elbow" method is commonly used



Lloyd’s algorithm

1. Assume K Centroids

2. Compute Squared Eucledian distance of each objects with these K centroids. Assign each to the closest centroid forming clusters.

3. Compute the new centroid (mean) of each cluster based on the objects assigned to each clusters.

4. Repeat 2 and 3 till convergence: usually defined as the point at which there is no movement of objects between clusters