CMP 0575 – Clustering (CURE)

Esteban Flores

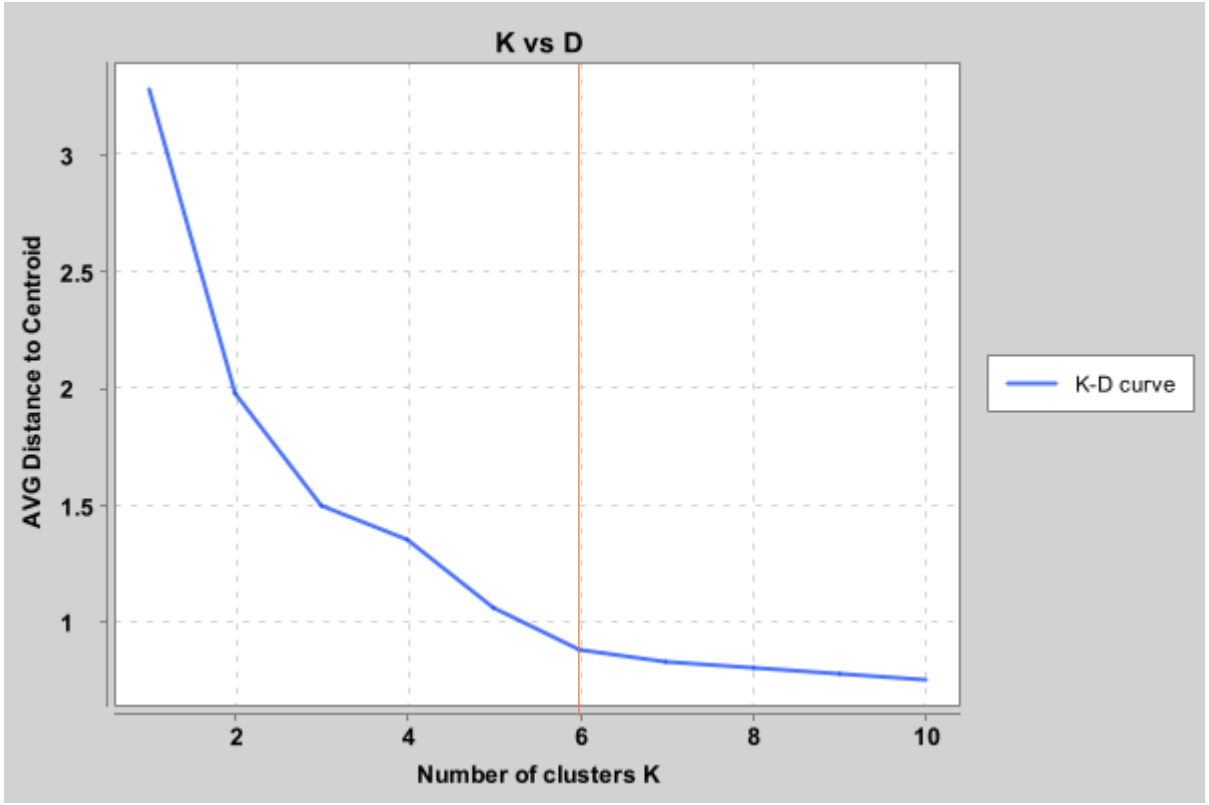
Paso 1 de 2:

- Pick a random sample of points that fit in main memory
- Cluster sample points hierarchically to create the initial clusters
- **Pick representative points:**
  - For each cluster, pick *k* (e.g., 4) representative points, as dispersed as possible
  - Move each representative point a fixed fraction (e.g., 20%) toward the centroid of the cluster

Paso 2 de 2:

- Now, rescan the whole dataset and visit each point ***p*** in the data set
- **Place it in the "closest cluster"**
  - Normal definition of "closest": that cluster with the closest (to *p*) among all the representative points of all the clusters

Centroides Óptimos Solución



Los primeros grupos agregarán mucha información (explicarán mucha varianza), pero en algún punto la ganancia marginal caerá, dando un ángulo en la gráfica. El número de agrupaciones se elige en este punto.

Centroides Óptimos Solución

Cluster	**Centroid**	**# points**
1	(12.95, 13.71, 0.86, 5.35, 3.02, 3.63, 5.09)	130
2	(18.00, 15.99, 0.88, 6.10, 3.64, 3.46, 5.96)	69
3	(13.99, 13.83, 0.92, 5.12, 3.38, 5.23, 4.78)	1
4	(13.80, 14.06, 0.88, 5.47, 3.19, 6.86, 5.22)	2
5	(20.70, 17.09, 0.89, 6.46, 3.95, 5.17, 6.28)	6
6	(12.95, 13.53, 0.89, 5.21, 3.16, 8.39, 5.03)	2