Clustering

Goal: divides set of n-vectors (denoted by $x_1, x_2, \ldots x_n$) into K groups or clusters (where vectors within each cluster are "close" to each other)

* In most case K is smaller than N.

Automatic Topic Discovery: Each vector represents a word histogram of a document. Clustering documents into groups can reveal documents with similar topics / genres / authors.

Clustering Objective:

- We have N n-vectors x_1 & $x_2, \ldots, x_N$ that need to be grouped into K-clusters.
- Group assignments: $N=5$; $K=3$ & $c = (3,1,1,1,2)$.
  x_1 is in cluster 3 ; x_2, x_3, x_4 are in cluster 1 ; x_5 is in cluster 2.
- Another representation : $G_1 = \{2,3,4\}$; $G_2 = \{5\}$; $G_3 = \{1\}$

Group representations: c_i = group that x_i is in.

$c_1 = 4$ is read as : 1st element is in group 4.

* z_{c_i} = avg of the vectors in its group.

* Each cluster j has a group representative z_j (n-vector) summary / center of group

Goal ? ⟶ make distance between x_i & z_{c_i} as small as possible.

clustering objective J_{clust}

$$J_{clust} = \frac{1}{N} \sum_{i=1}^{N} \| x_i - z_{c_i} \|^2$$

If this is small, then we consider it to be a good cluster.

Q simplified optimization problems:

① partitioning vectors (when representatives are fixed):

- we assign x_i to cluster with representative z_j as: $j = \arg\min_{j=1,\ldots,K} \| x_i - z_j \|$
- This means that each vector is assigned to its nearest representative.

② optimizing group representatives (when assignments are fixed)

- we minimize J_{clust} by setting representative of each group to the centroid of the vectors in the group.

$$z_j = \frac{1}{|G_j|} \sum_{i \in G_j} x_i$$

$|G_j|$ ↑ number of elements in G_j

avg of the values of x_k within the subset defined by G_j

K-means Algorithm:

Goal: partition a set of N vectors $x_1, x_2, \ldots x_N$ into K clusters, such that each vector is assigned to a group with representative (centroid) as close as possible.

Objective: minimize the sum of squared distances between each vector and its assigned group representative. (J_{clust} decreases in each step till z_j stops changing)

Additional comments and Clarifications:

- **Ties in step 1:** If vector is equally close to more than 1 representative, you can break the tie by assigning it to the group with the smallest j index. (Ex: if we had 2 & 5; 2 wins)
- **Empty Groups:** We drop the empty group, meaning final no. of groups is $< K$
- **Stopping Condition:** If group assignments remain same in 2 successive iterations, rep's will also remain unchanged & algo will stop.
- **Initial Group Representatives:** randomly choose them at the start / start by assigning mean.

* K-means algorithm is heuristic (it cannot guarantee that partition it finds min's J clust)
∴ we generally run the algo with different initial representatives & choose 1 among them that is most optimal.

Interpretation of z:

→ If 4th component of vectors represents the age of the voters, then the 4th component of centroid z_3 for group 3 $((z_3)_4)$ = 37.8 means that the average age of that group is 37.8

Complexity: 1. partitioning = distance x comparison = $3Kn$ flops.
2. updating centroids = Nn flops.
3. Flops per iteration: $(3K+1)Nn = NKn$ flops → for multiple iterations: $1000NKn$

Ex **: $N = 100{,}000$, $n = 100$, $K = 10$

Total flops = $1000NKn = 10^3 \times 10^5 \times 10^2 \times 10^1 = 10^{11}$ flops.

on a computer that can process 1 Gflops/sec [1 billion flops] this would take about $10^{11} / 10^9 = 100$ seconds.