(K-MEANS) CLUSTERING

→ Clustering discovers the underlying structure of the data

↳ Unsupervised task — do not require labelled data to build a model

→ Clustering methods show us how many sub-populations there are in the data

→ Clustering algorithms can be characterised by various properties:

HARD vs. SOFT

- an instance can only belong to 1 cluster
- an instance can belong to >1 cluster

POLYTHETIC vs. MONOTHETIC

- Cluster instances are 'similar' according to some measure across many/all attribute values
- Cluster instances share some common attribute values

FLAT vs. HIERARCHICAL

- each cluster is a grp. of instances w/ no relationship between clusters
- some clusters group other clusters rather than instances directly (Taxonomy)

→ Clustering methods we looked / will look at include:

1) K-D trees ——— Monothetic, hard boundaries, hierarchical
2) K-means clustering ——— Polythetic, hard boundaries, flat
3) Gaussian mixtures ——— Polythetic, soft boundaries, flat
4) Agglomerative clustering —— Polythetic, hard boundaries, hierarchical

K-MEANS

→ Data is partitioned into K sub-populations

→ A cluster is characterised by a 'centroid', which has, for each attribute, the avg. of that attribute's values over all the instances in the cluster.

APPLICATIONS

1) Discover classes in an unsupervised manner

e.g. Cluster images of handwritten digits (w/ K=10)

2) Used to perform dimensionality reduction

→ Data is not regular enough

→ Can use K-means clustering on a dataset and then build a model for each subgroup [Cluster-then-predict]

K-MEANS CLUSTERING ALGO.

1) We start w/ a set of points $x_1 \dots x_n$ and K

2) Place centroids $c_1 \dots c_K$ at random locations

3) Repeat until no centroid changes:

Distance (e.g. Euclidean) between instance x_i and center c_j

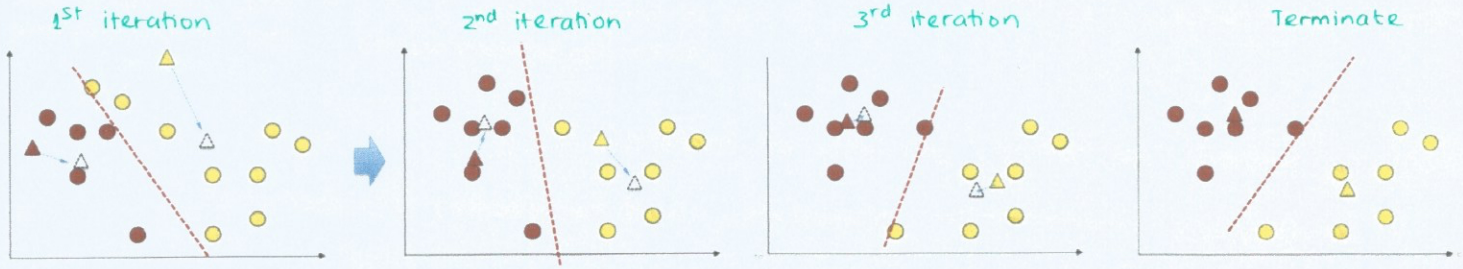
a) For each point x_i, find nearest centroid c_j $\left[\underset{j}{\text{argmin}}\, D(x_i, c_j) \right]$

and assign x_i to cluster j

b) For each cluster $j = 1 \dots K$, change the centroid c_j to be the mean of all points x_i assigned to cluster j in prev. step

$$\left[c_j(a) = \frac{1}{n_j} \sum_{x_i \to c_j} x_i(a) \quad \text{for} \quad a = 1 \dots d \right]$$

↳ The complexity is $O(\#\text{iterations} \times \#\text{clusters} \times \#\text{instances} \times \#\text{dimensions})$



K-MEANS PROPERTIES

1) Minimizes the aggregate intra-cluster distance

Same as variance if Euclidean distance is used

total squared distance from point to center of its cluster $\sum_j \sum_{x_i \to c_j} D(c_j, x_i)^2$

2) Converges to a local minimum

→ So different starting points give very different results

→ Run several times w/ random starting points and pick clustering that yields smallest aggregate distance

3) Nearby points may not end up in the same cluster

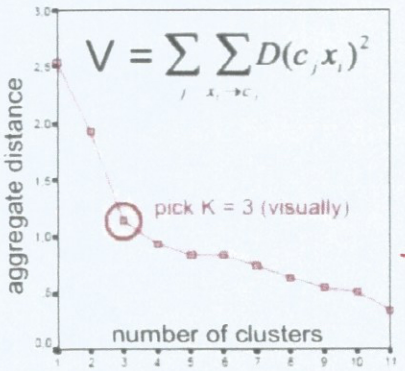
i.e.

HOW TO PICK VALUE OF K ?

1) Class labels (of the whole population)

   i.e. digits (0..9)

2) Visually from a scree plot



a.) Looking at the knee of the scree plot

b.) The maximum of the 2nd derivative of V

↳ 'ELBOW' method
→ Point where rate of decline changes the most

→ This is why we have to determine value of K because aggregate distance ↓ as the num. of centroids ↑. It'll reach a point where each point has its own centroid.

EVALUATING CLUSTERING ALGO.

EXTRINSIC evaluation

→ is measuring the improvement you get on another task if you add the cluster number to the input data for that task

INTRINSIC evaluation

→ Does clustering help you to understand the structure of your data?

→ Does clusters correspond to classes?

→ If no natural notion of classes, you can get human judgements