EE769 Intro to ML
Unsupervised Learning - Clustering

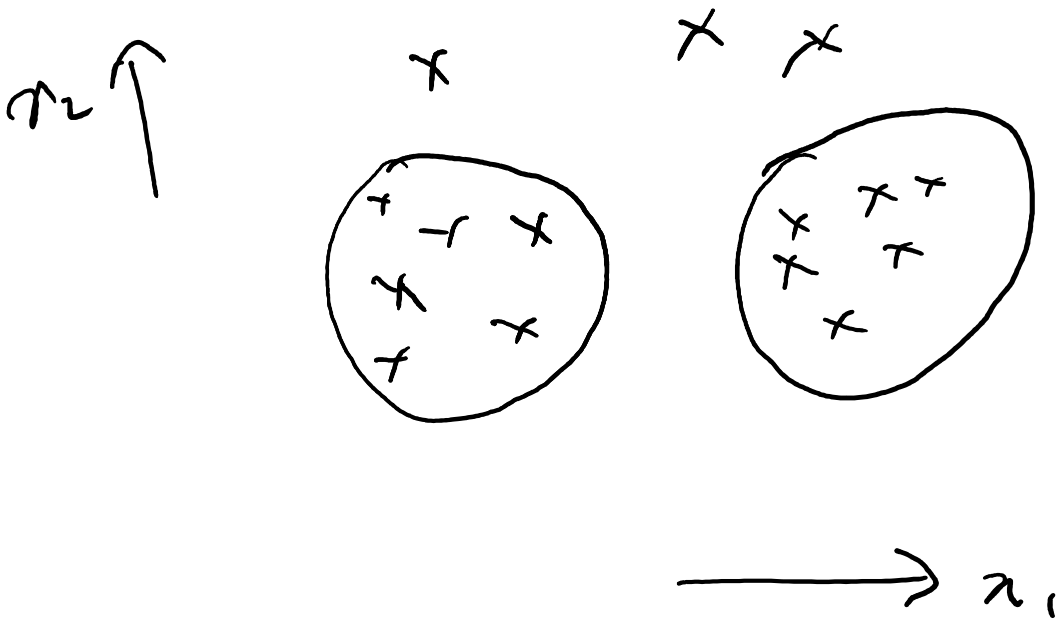
Amit Sethi
Faculty member, IIT Bombay

Learning objectives

- List applications of clustering
- Write objectives and algorithms for various clustering algorithms
- List methods to assess goodness of clustering

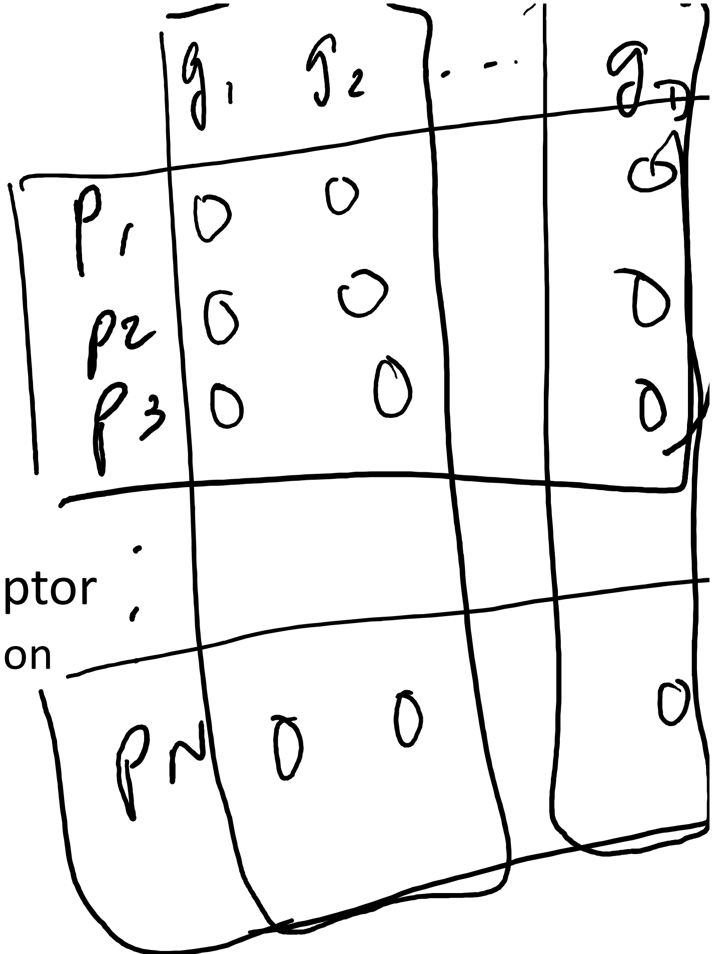
Objective of clustering

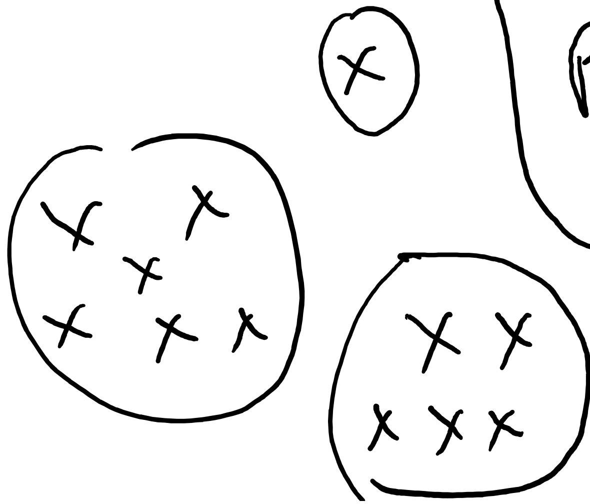
- Find natural subsets of data samples



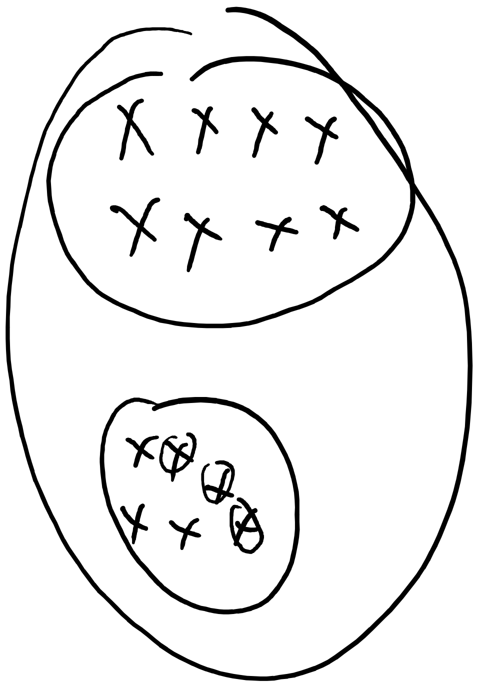
Applications of clustering

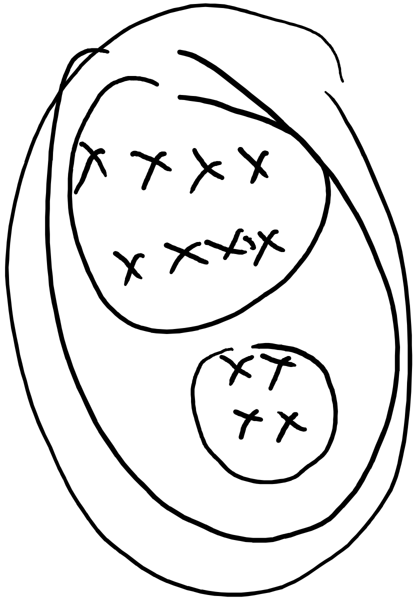
- Find natural groups
  - E.g., for customer segmentation
- Reduce data and deal with only cluster descriptor
  - E.g., for cancer sub-types based on gene expression
- Find outliers
  - E.g., three door cars





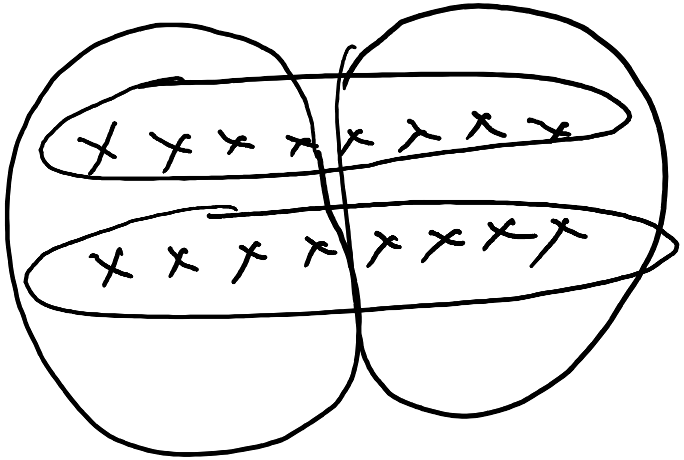
Clustering is hard because it is unsupervised





N → N cluster

1 cluster



Hard partitioning using K-means clustering

$K \rightarrow$ Hyperparameter

- Minimize sum of squared distances from cluster center

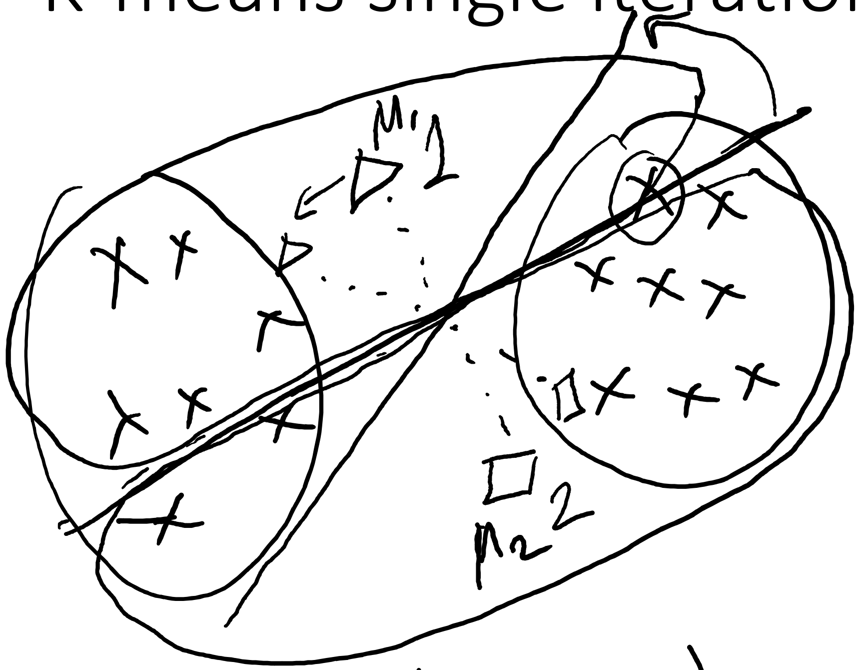
If ; $x_i \in C_k \Rightarrow x_i \notin C_j$, $j \neq k$

S_k is the set of points belonging to C_k

objective $\min_{C_k} \sum_i \sum_k \mathbb{1}_{x_i \in C_k} \|x_i - \mu_k\|^2$

Centroid $\mu_k = \frac{1}{|C_k|} \sum_{x_i} \mathbb{1}_{x_i \in C_k} x_i$

K-means single iteration



$K = 2$

Randomly initialize centroids

Iteration loop

① Compute membership of each point

② Recompute μ_j $\forall j$

Until change in centroid locations $< \varepsilon$

① $\arg\min_j d(x_i, \mu_j)$

② $\mu_j = \frac{1}{|C_j|} \sum x_i \text{ if } x_i \in C_j$

K-means initialization

1. Random initialization
2. Pick furthest points
   Pick first point randomly
   Pick next pt. that is furthest
   from the previosly picked pt.s

Complete K-means

Stopping criteria

k^{n}

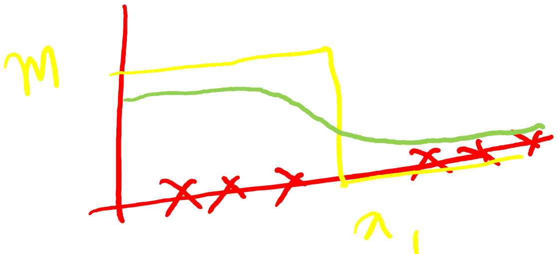
ε

How to choose k

Soft-assignment

Not guaranteed to be optimal

Fuzzy c-means objective

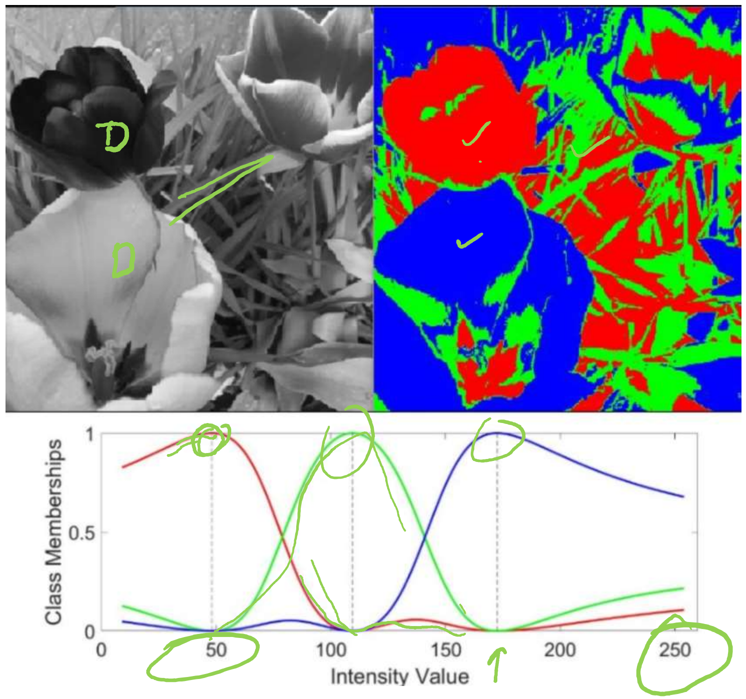


- Minimize weighted sum of squared distances from cluster center
- $w_{ij} = [\, \sum_k \{ d(x_i,c_j) / d(x_i,c_k) \}^{2/m-1} \,]^{-1}$
- Weight is w_{ij}^m

m > 1

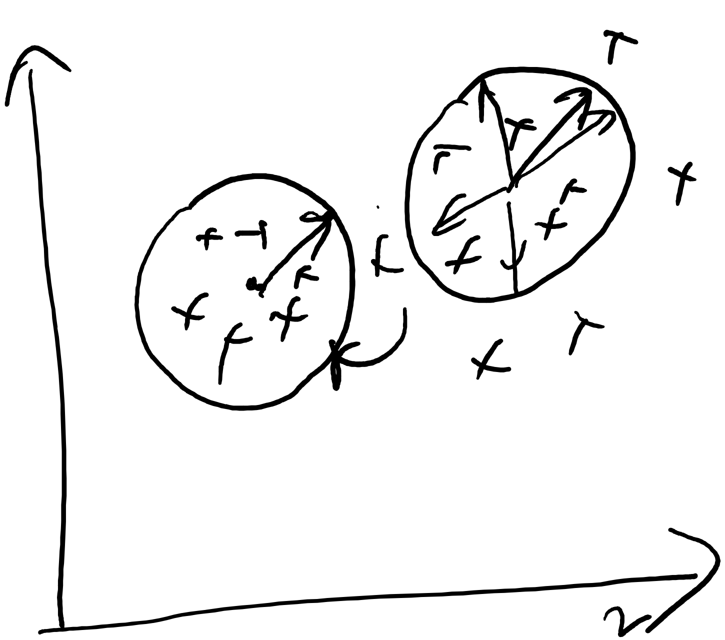
m = 1

⇒ k-means

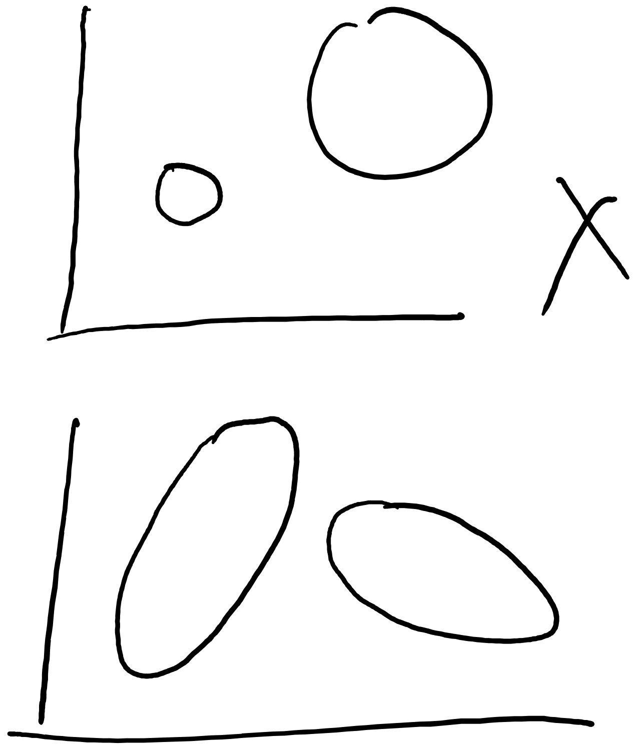


Prior: Equal-sized hyperspheres

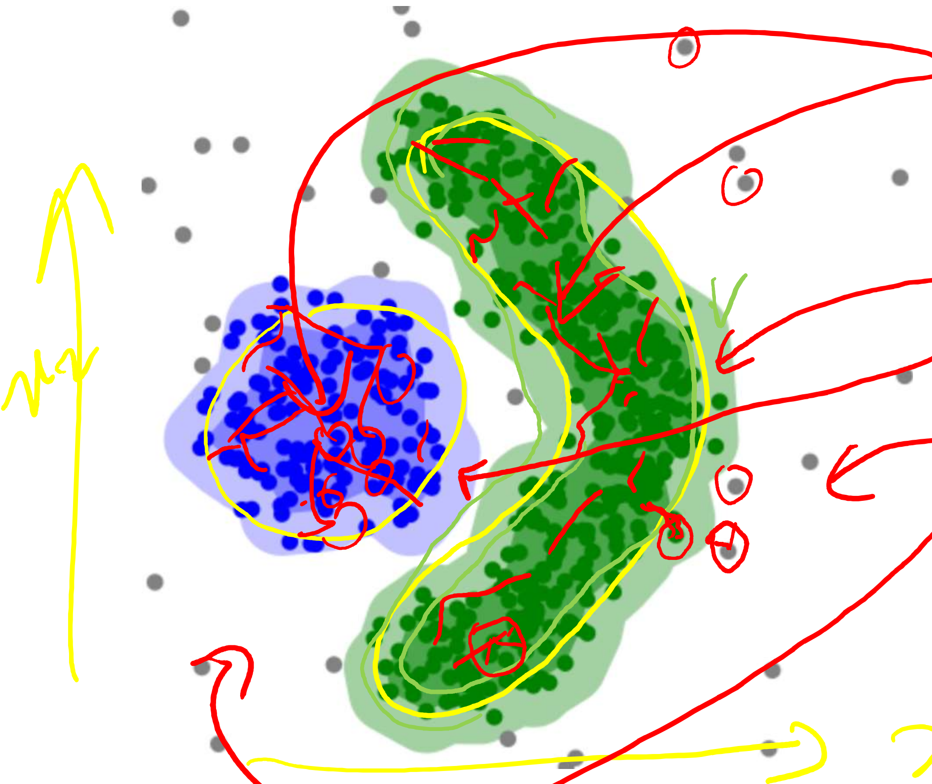
$$\frac{w_{ij}}{\sum_k w_{ik}}$$



$$\min \|x_i - c_j\|^2$$



DBSCAN – density-based clustering

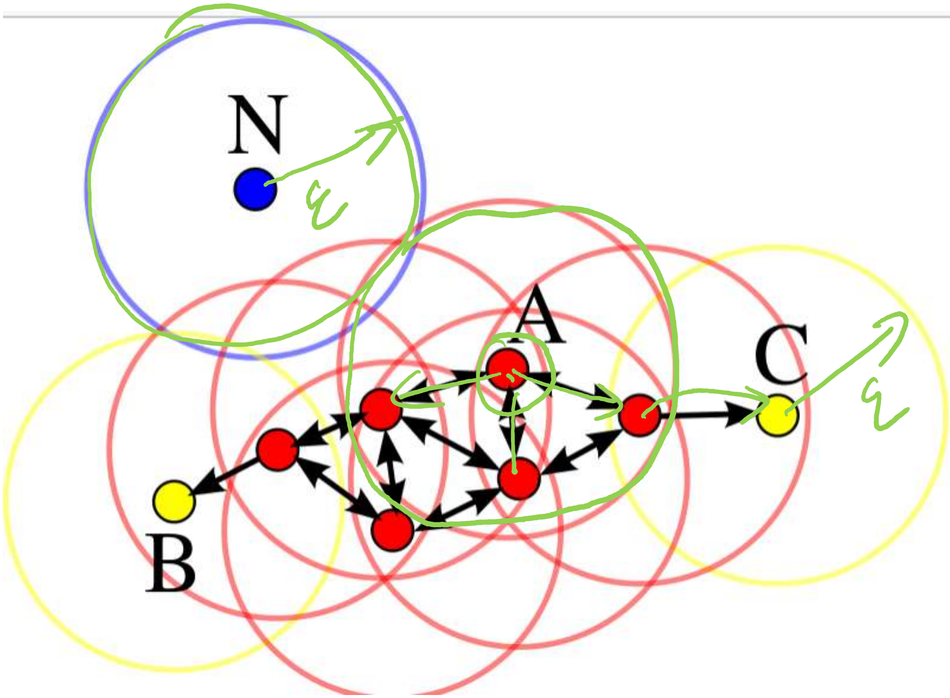


- High-density cores of arbitrary shapes
- Medium-density peripheries
- Low-density outlier regions

x_1

$\rightarrow p(x_1, x_2)$

DBSCAN critiera



- Hyperparameters:
  - Min points m
  - Tolerance ε

  2

- Three types of points:
  - Core: with m points in ε radius High density
  - Reachable: Not core, but there exists a path from a core point with hops < ε each Medium density
  - Outliers: Others low density
- Clusters are connected core points and their reachable points

A crude DBSCAN algorithm

Adjacency discovery.

- For each sample x_i
  - For each other sample x_j
    - Mark j as neighbor of i, if $d_{ij} < \varepsilon$
    - Increment number of neighbors of n_i of x_i

- For each sample x_i
  - If neighbors >= *m* then mark as *core*
  - else If neighbors > *0* then mark as *reachable*
  - else If neighbors = 0, then mark as *outlier*

marking

mark all points as unclustered

- For each sample x_i
  - If core and unclustered, then mark all connected samples with this cluster

Cluster discovery

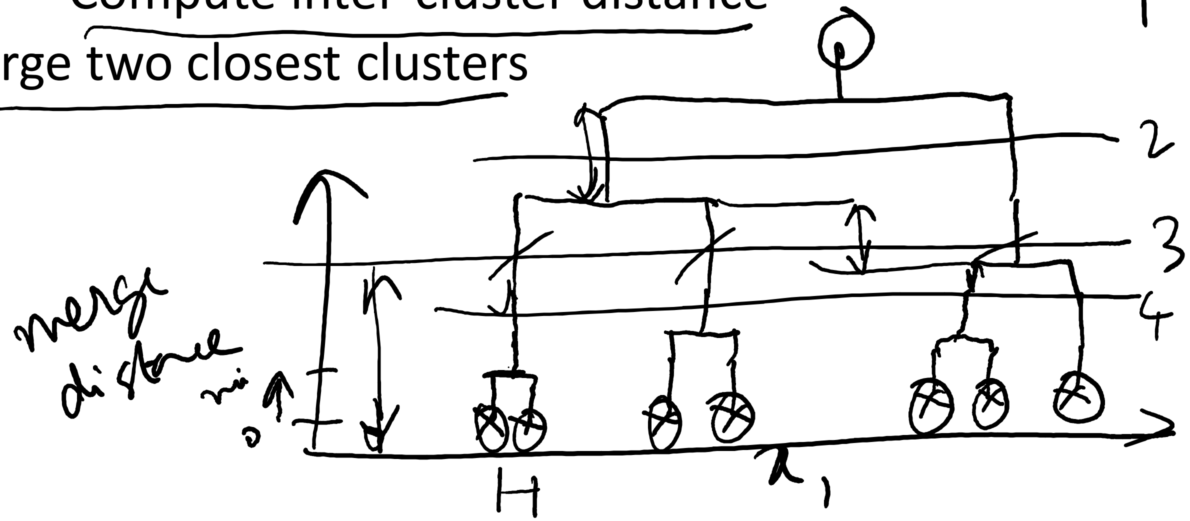
Hierarchical clustering gives us all possible clusters as a dendrogram

Variance of the data

Variance of the cluster

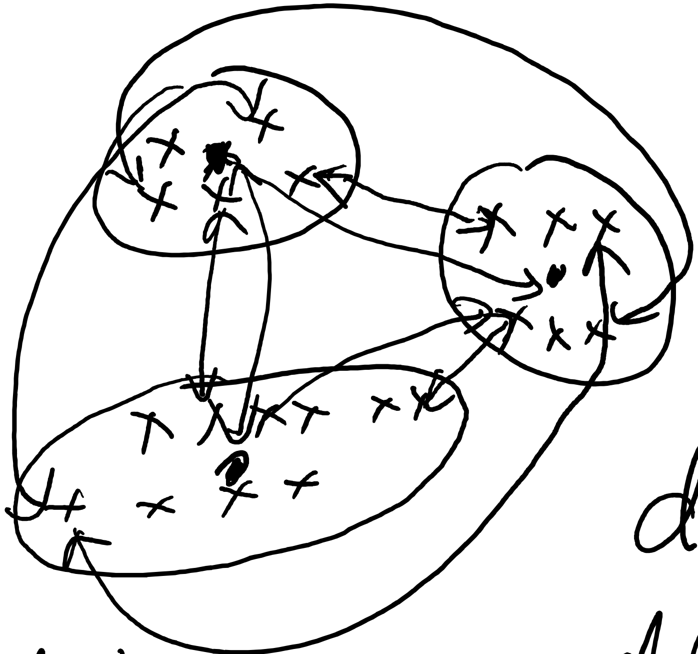
- Start with each sample as its singleton cluster
- For each merge iteration (N)
  - For each cluster
    - For each other cluster
      - Compute inter-cluster distance
  - Merge two closest clusters

Dendrogram



Some types of cluster distances

- Min: Single-linkage
- Max: Complete-linkage
- Average: Centroid distance



$d(x_i, x_j)$

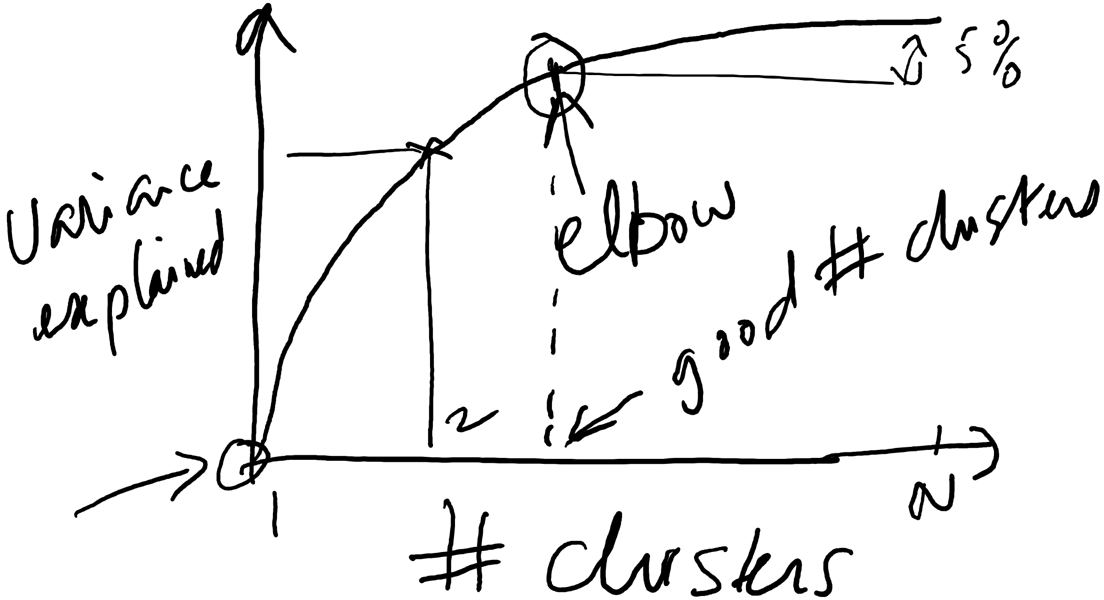
$d(x_i, \mu_j)$

Euclidean dist

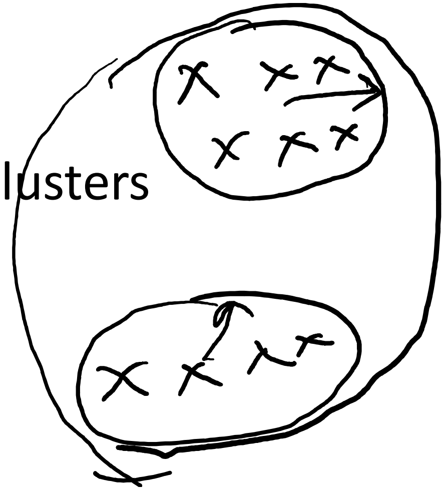
Mahalanobis dist.

Some dissimilarity

Clustering criteria



- Several methods:
  - Dunn's Validity Index
  - Silhouette method
  - C–index
  - Goodman–Kruskal index
  - Elbow method
  - Davies Bouldin index



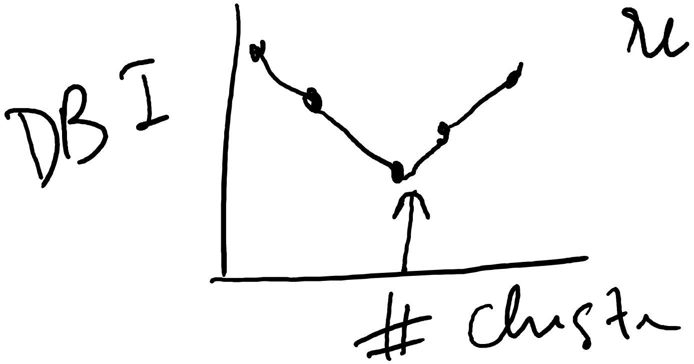
- Elbow method: Variation explained versus number of clusters
  - Between group variance versus total variance
  - E.g., Total variance – average or max within cluster variance

Clustering criteria – Davies Bouldin index

- Minimize ratio of intra-cluster versus inter-cluster variation
- Average$_i$ of max$_j$ [$(S_i + S_j) / M_{ij}$]

pt.s in a cluster should be similar to each other

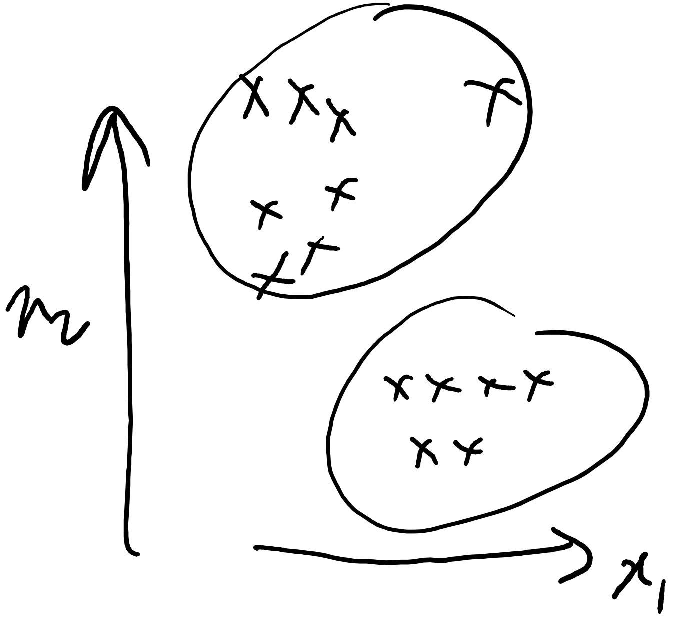
relative to pt. outside the cluster

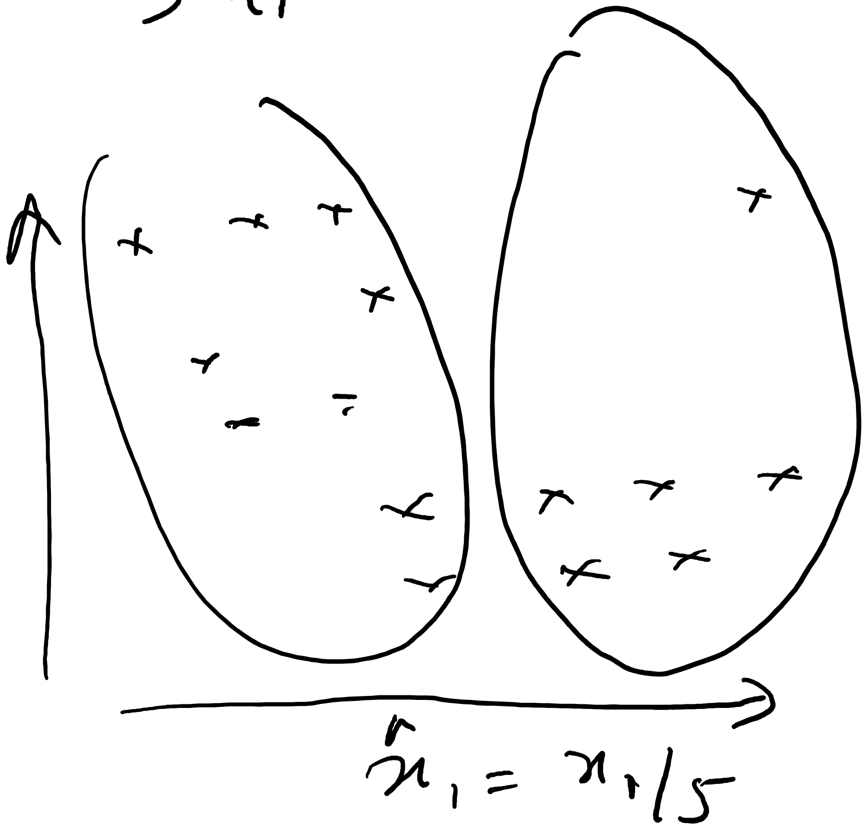


Impact of variable transforms on clustering

- Normalizing variables S
- Taking log or power tranforms

$\rightarrow x_1$ is revenue / customer





Advanced topics

- Gaussian mixture models using EM algorithm
- Spectral clustering