K-MEANS CLUSTERING

Let $x_1, \dots, x_n \in \mathbb{R}^d$.

Recall that the goal of clustering is to assign th data to disjoint subsets called

(A) ____________ so that points in the same cluster are more similar to each other than to points in other clusters

Therefore, at the heart of every clustering algorithm is a notion of ____________.
Often it is more convenient to work with a _______________.

Dissimilarity

A dissimilarity matrix is an $n \times n$ matrix

$$D = [d_{ij}]_{i,j=1}^{n}$$

which has the following properties

- $d_{ii} = 0$
- $d_{ij} = d_{ji}$
- $d_{ij} \geq 0$

Conceptually, if x_i is more similar to x_j than to x_k, then

(B)

A dissimilarity matrix need not satisfy the triangle inequality:

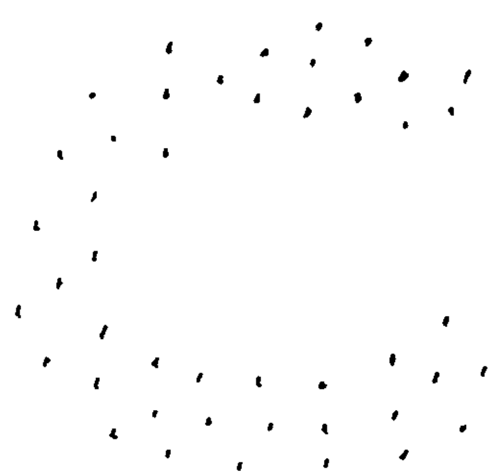
Examples

- Euclidean distance

c

- Squared Euclidean distance

- kNN - based distance



K-means criterion

A cluster map is a function

$$C: \{1,2,\dots,n\} \longrightarrow \{1,2,\dots,K\}$$

that partitions the data into K clusters.

In K-means clustering we

- assume K is known (more on this later)
- adopt the squared Euclidean distance as a dissimilarity

(D) $$d_{ij} =$$

- seek to minimize the ________ ________ ________

$$W(C) =$$

where

$$n_k =$$

Algorithm

(E) The K-means criterion is a ________________ optimization problem. The number of possible cluster maps C is

$$\frac{1}{K!} \sum_{k=1}^{K} (-1)^{K-k} \binom{K}{k} k^n \qquad \text{(Jain \& Dubes, 1988)}$$

$$\begin{cases} = 34{,}105 & \text{if } n=10,\ K=4 \\ \approx 10^{10} & \text{if } n=19,\ K=4 \end{cases}$$

There is no known efficient search strategy for this space. Therefore we resort to an iterative, suboptimal algorithm.

Exercise | Show that

$$W(C) = \sum_{k=1}^{K} \sum_{i:C(i)=k} \| x_i - \bar{x}_k \|^2$$

where

$$\bar{x}_k := \frac{1}{n_k} \sum_{i:C(i)=k} x_i$$

Solution

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i:C(i)=k}\;\sum_{j:C(j)=k} \underbrace{\|x_i - \bar{x}_k - (x_j - \bar{x}_k)\|^2}$$

$$\langle x_i - \bar{x}_k - (x_j - \bar{x}_k),\ x_i - \bar{x}_k - (x_j - \bar{x}_k)\rangle$$

$$= \|x_i - \bar{x}_k\|^2 - 2(x_i - \bar{x}_k)^T(x_j - \bar{x}_k) + \|x_j - \bar{x}_k\|^2$$

$$= \frac{1}{2}\sum_{k=1}^{K}\frac{1}{n_k}\Bigg[\sum_{i:C(i)=k}\;\sum_{j:C(j)=k}\|x_i - \bar{x}_k\|^2$$

$$-2\sum_{i:C(i)=k}\;\sum_{j:C(j)=k}(x_i - \bar{x}_k)^T(x_j - \bar{x}_k)$$

$$+\sum_{i:C(i)=k}\;\sum_{j:C(j)=k}\|x_j - \bar{x}_k\|^2\Bigg]$$

$$= \frac{1}{2}\sum_{k=1}^{K}\frac{1}{n_k}\Bigg[n_k\cdot\sum_{i:C(i)=k}\|x_i - \bar{x}_k\|^2$$

$$+ n_k\sum_{j:C(j)=k}\|x_j - \bar{x}_k\|^2\Bigg]$$

$$= \sum_{k=1}^{K}\;\sum_{i:C(i)=k}\|x_i - \bar{x}_k\|^2$$

Therefore we seek to solve:

$$C^* = \arg\min_C \sum_{k=1}^{K} \sum_{i:C(i)=k} \|x_i - \bar{x}_k\|^2$$

Note that for fixed C and k,

(F) $$= \arg\min_m \sum_{i:C(i)=k} \|x_i - m\|^2$$

Therefore

$$C^* = \arg\min_{C,\{m_k\}_{k=1}^K} \underbrace{\sum_{k=1}^{K} \sum_{i:C(i)=k} \|x_i - m_k\|^2}_{W(C,\{m_k\}_{k=1}^K)}$$

This suggest an iterative algorithm

1) Given C, choose $\{m_k\}_{k=1}^K$ to minimize $W(C, \{m_k\}_{k=1}^K)$

2) Given $\{m_k\}_{k=1}^K$, choose C to minimize $W(C, \{m_k\}_{k=1}^K)$

G 1) $m_k^* =$

2) $C^*(i) =$

K-means Clustering Algorithm

Initialize $\bar{x}_k$, $k=1,\ldots,K$

Repeat

- $C(i) =$
- $\bar{x}_k =$

Until clusters don't change

H

Remarks

- The algorithm is typically initialized by setting each $\bar{x}_k$ to be a random data point
- Since the algorithm often finds a local min, several random initializations are recommended.

Cluster Geometry

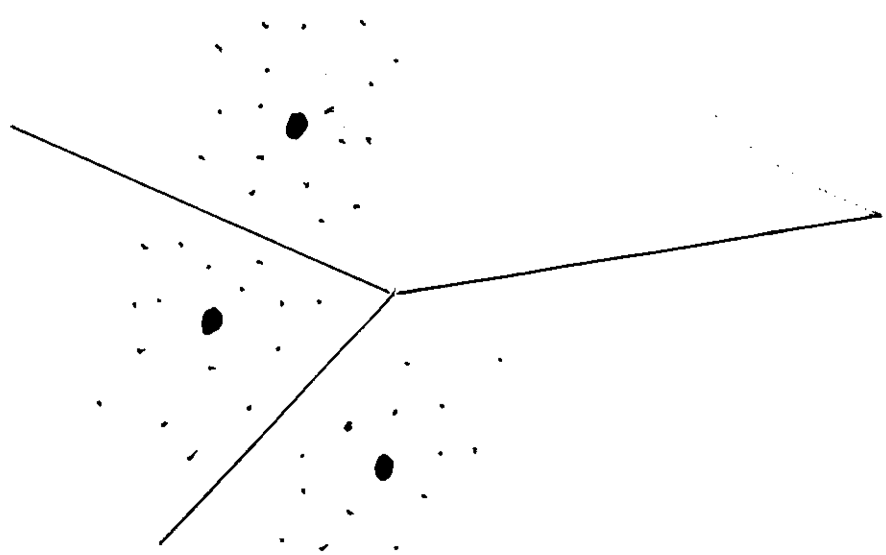
Clusters are "nearest neighbors"

(I) regions or ____________ cells

defined with respect to the cluster means.

Therefore the cluster boundaries are

____



$K = 3$

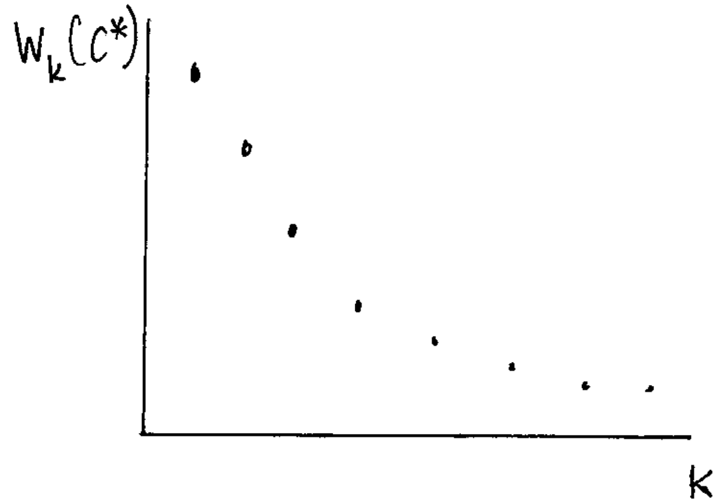
K-means will fail if clusters are

______________ .

Model selection

How to choose K?

If $W_K(C^*)$ is the within-cluster scatter based on K clusters, we have a plot like this



If the "right" number of clusters is K^*, we expect

- for $K < K^*$, $W_k(C^*) - W_{k-1}(C^*)$ will be <u>large</u>
- for $K > K^*$, $W_k(C^*) - W_{k-1}(C^*)$ will be <u>small</u>

This suggests choosing K near the "knee" of the curve.

$\boxed{\text{Key}}$ A. clusters, similarity, dissimilarity

B. $d_{ij} < d_{ik}$

$d_{ij} + d_{jk} \ngeq d_{ik}$

C. • $\|x - y\| = \left(\sum_{j=1}^{d} (x^{(j)} - y^{(j)})^2\right)^{\frac{1}{2}}$

• $\|x - y\|^2$

• length of shortest path on k-nearest neighbor graph, for some k

D. $d_{ij} = \|x_i - x_j\|^2$, within cluster scatter

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{i:C(i)=k} \underbrace{\left[\frac{1}{n_k} \sum_{j:C(j)=k} \|x_i - x_j\|^2\right]}_{\text{avg. dissim. to points in same cluster}}$$

$$n_k = \sum_{i=1}^{n} \mathbb{1}_{\{C(i)=k\}}$$

E. Combinatorial F. $\bar{x}_k$

G. $$m_k^* = \frac{1}{n_k} \sum_{i:C(i)=k} x_i$$

$$C^*(i) = \arg\min_k \|x_i - \quad \|$$

H. $$C(i) = \arg\min_k \|x_i - \bar{x}_k\|$$

$$\bar{x}_k = \frac{1}{n_k} \sum_{i:C(i)=k} x_i$$

I. Voronoi, hyperplanes, nonconvex