K - MEANS CLUSTERING

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

Recall that the goal of clustering is to assign
(A) the data to disjoint subsets called _____
so that points in the same cluster are more similar to each other than to points in other clusters.

A clustering can be represented by a _____ _____, which is a function

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

where K is the number of clusters.

K-means criterion

Choose C to minimize

(B) $$W(C) =$$

where

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

Note that K is assumed fixed and known.

$W(C)$ is sometimes called the <u>within cluster scatter</u> because of the following property.

<u>Exercise</u> Show that

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{i: C(i)=k} \left[\frac{1}{n_k} \sum_{j: C(j)=k} \|x_i - x_j\|^2 \right]$$

Avg. dissimilarity to points in same cluster

Solution

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

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

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

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

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

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

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

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

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

Algorithm

(c) Minimizing the K-means criterion is a ____ optimization problem. The number of possible cluster maps C is

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

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

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

Recall 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,

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

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

(E)

1) $m_k^* =$

2) $C^*(i) =$

K-means Clustering Algorithm

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

Repeat

(F)

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

Until clusters don't change

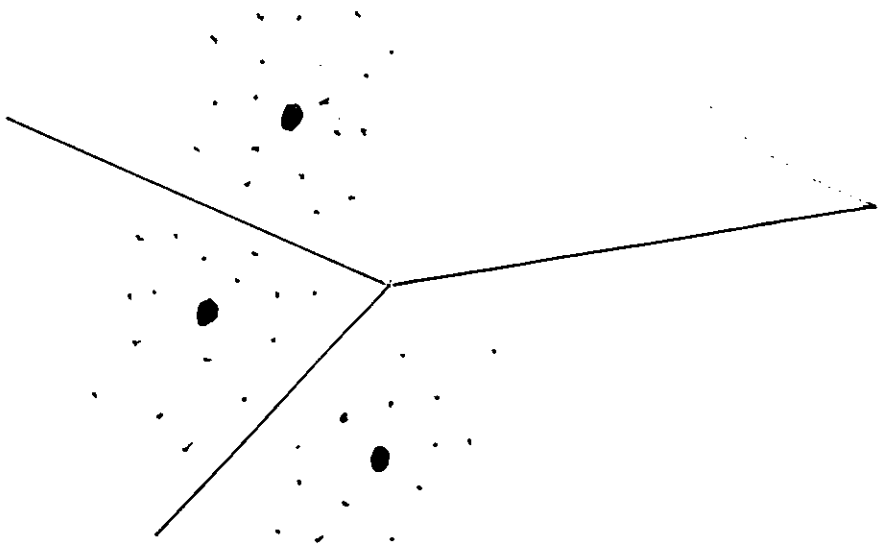
Remarks

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

Cluster Geometry

Clusters are "nearest neighbor" regions or _____ cells defined with respect to the cluster means.

Therefore the cluster boundaries are _____



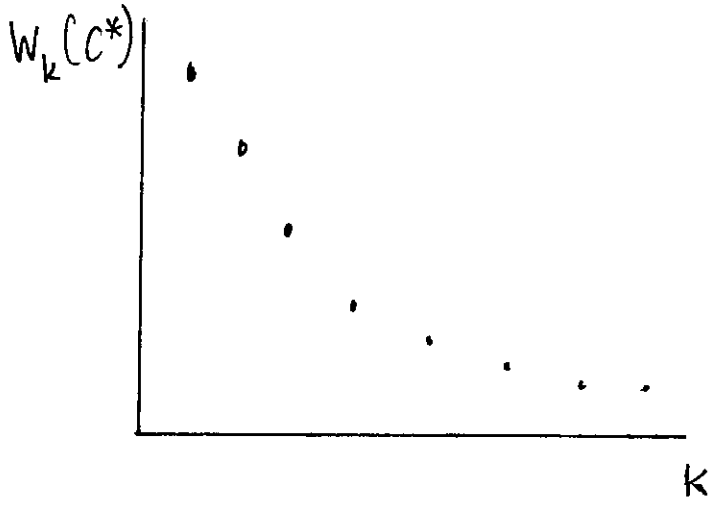
K = 3

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

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

Key

A. clusters, cluster map

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

C. combinatorial

D. $\bar{x}_k$

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

$$C^*(i) = \arg\min_k \|x_i - m_k\|^2$$

F. $C(i) = \arg\min_k \|x_i - \bar{x}_k\|^2$

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

G. Voronoi, hyperplanes, nonconvex