Clustering - IV

Prof. Dan A. Simovici

UMB

- Partitional clustering algorithms aim to discover partitions of a set of objects that optimize certain criteria and, generally, do this through iterative processes.
- These algorithms begin with a set of initial centroids as seeds for the clusters, assign objects to these tentative centers, and recompute these centroids and their corresponding clusterings as they try to optimize the clustering criteria.

The notion of *inertia* a finite subset X of $\mathbb{R}^m$ relative to a vector $\mathbf{z}$ originates in mechanics of solids.

Definition

Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of vectors in $\mathbb{R}^m$. The *inertia of* X *relative to a vector* $\mathsf{z} \in \mathbb{R}^m$ is the number

$$I_{\mathbf{z}}(X) = \sum_{j=1}^{n} \| \mathbf{x}_j - \mathbf{z} \|_2^2 .$$

The special case of the inertia of X relative to the vector

$$\mathbf{c}_X = \frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j$$

is referred to as the *sum of square errors* of X. We denote $I_{\mathbf{c}_X}(X)$ by $\mathrm{sse}(X)$.

The *mean square error* of the set X is the number $r(X)$ defined by

$$r(X) = \frac{\mathrm{sse}(X)}{|X|}.$$

Theorem

(Huygens' Inertia Theorem)
Let $X = \{x_1, \ldots, x_n\}$ be a finite set of vectors in $\mathbb{R}^m$. We have:

$$I_z(X) - I_{c_X}(X) = n \parallel c_X - z \parallel_2^2,$$

for every $z \in \mathbb{R}^m$.

Proof

The inertia of X relative to $\mathbf{c}_X$ is

$$\begin{aligned} I_{\mathbf{c}_X}(X) &= \sum_{j=1}^{n} \| \mathbf{x}_j - \mathbf{c}_X \|_2^2 = \sum_{j=1}^{n} (\mathbf{x}_j - \mathbf{c}_X)'(\mathbf{x}_j - \mathbf{c}_X) \\ &= \sum_{j=1}^{n} (\mathbf{x}_j'\mathbf{x}_j - \mathbf{c}_X'\mathbf{x}_j - \mathbf{x}_j'\mathbf{c}_X + \mathbf{c}_X'\mathbf{c}_X). \end{aligned}$$

Similarly, we have

$$I_{\mathbf{z}}(X) = \sum_{j=1}^{n} (\mathbf{x}_j'\mathbf{x}_j - \mathbf{z}'\mathbf{x}_j - \mathbf{x}_j'\mathbf{z} + \mathbf{z}'\mathbf{z}).$$

Proof cont'd

This allows us to write

$$
\begin{aligned}
I_{\mathbf{z}}(X) - I_{\mathbf{c}_X}(X) &= \sum_{j=1}^{n} (\mathbf{c}_X - \mathbf{z})' \mathbf{x}_j + \sum_{j=1}^{n} \mathbf{x}_j' (\mathbf{c}_X - \mathbf{z}) + \mathbf{z}'\mathbf{z} - \mathbf{c}_X'\mathbf{c}_X \\
&= (\mathbf{c}_X - \mathbf{z})' \sum_{i=1}^{n} \mathbf{x}_j + \left(\sum_{j=1}^{n} \mathbf{x}_j \right)' (\mathbf{c}_X - \mathbf{z}) + n(\mathbf{z}'\mathbf{z} - \mathbf{c}_X'\mathbf{c}_X) \\
&= n(\mathbf{c}_X - \mathbf{z})'\mathbf{c}_X + n\mathbf{c}_X'(\mathbf{c}_X - \mathbf{z}) + n(\mathbf{z}'\mathbf{z} - \mathbf{c}_X'\mathbf{c}_X) \\
&= n \parallel \mathbf{c}_X - \mathbf{z} \parallel_2^2,
\end{aligned}
$$

which is the equality of the theorem.

Corollary

Let $X = \{x_1, \ldots, x_n\}$ be a set of vectors in $\mathbb{R}^m$. The minimal value of the inertia $I_z(X)$ is achieved for $z = c_X$.

This is an immediate consequence of Huygens Theorem.

Corollary

The sum of all squared distances between the members of a set divided by its cardinality equals the sum of the square errors of that set.

Proof: By Huygens' Theorem, the inertia of X relative to one of its members $\mathbf{x}_k$ is

$$\sum_{i=1}^{n} \| \mathbf{x}_i - \mathbf{x}_k \|^2 = I_{\mathbf{x}_k}(X) = I_{\mathbf{c}_X} + n \| \mathbf{c}_X - \mathbf{x}_k \|_2^2 .$$

Therefore,

$$\begin{aligned} \sum_{k=1}^{n}\sum_{i=1}^{n} \| \mathbf{x}_i - \mathbf{x}_k \|^2 &= 2\sum\{\| \mathbf{x}_i - \mathbf{x}_k \|^2 \mid 1 \leqslant k < i \leqslant n\} \\ &= n I_{\mathbf{c}_X} + n\sum_{k=1}^{n} \| \mathbf{c}_X - \mathbf{x}_k \|_2^2 = 2n I_{\mathbf{c}_X}, \end{aligned}$$

which implies the statement of the corollary.

Definition

For a set X and a partition $\pi = \{U_1, \ldots, U_k\}$ of X, the *sum of the squared errors* of π is the number $\text{sse}(\pi)$ given by:

$$\text{sse}(\pi) = \sum_{i=1}^{k} \text{sse}(U_i) = \sum_{i=1}^{k} \sum \{ \| \mathbf{x} - \mathbf{c}_{U_i} \|^2 \mid \mathbf{x} \in U_i \}.$$

Corollary

The sum of square errors of a partition $\pi = \{U_1, \ldots, U_k\}$ of a finite subset X of $\mathbb{R}^m$ equals the sum over all blocks of mean square errors, $\sum_{j=1}^{k} r(U_j)$.

This statement follows immediately.

Lemma

Let W be a subset of $\mathbb{R}^m$ and let $\sigma = \{U, V\}$ be a bipartition of W. We have:

$$sse(W) = sse(U) + sse(V) + \frac{|U|\,|V|}{|W|} \parallel c_U - c_V \parallel^2 .$$

Proof

By applying the definition of the sum of square errors we have:

$$\begin{aligned}
&\text{sse}(W) - \text{sse}(U) - \text{sse}(V) \\
&= \sum\{\| \mathbf{x} - \mathbf{c}_W \|^2 \,|\mathbf{x} \in U \cap V\} \\
&\quad - \sum\{\| \mathbf{x} - \mathbf{c}_U \|^2 \,|\mathbf{x} \in U\} - \sum\{\| \mathbf{x} - \mathbf{c}_V \|^2 \,|\mathbf{x} \in V\}.
\end{aligned}$$

The centroid of W is given by:

$$\mathbf{c}_W = \frac{1}{|W|} \sum\{\mathbf{x}|\mathbf{x} \in W\} = \frac{|U|}{|W|}\mathbf{c}_U + \frac{|V|}{|W|}\mathbf{c}_V.$$

This allows us to evaluate the variation of the sum of squared errors:

$$\begin{aligned}
&\text{sse}(W) - \text{sse}(U) - \text{sse}(V) \\
&= \sum\{\| \mathbf{x} - \mathbf{c}_W \|^2 \mid \mathbf{x} \in U \cup V\} \\
&\quad - \sum\{\| \mathbf{x} - \mathbf{c}_U \|^2 \mid \mathbf{x} \in U\} - \sum\{\| \mathbf{x} - \mathbf{c}_V \|^2 \mid \mathbf{x} \in V\} \\
&= \sum\{\| \mathbf{x} - \mathbf{c}_W \|^2 - \| \mathbf{x} - \mathbf{c}_U \|^2 \mid \mathbf{x} \in U\} \\
&\quad + \sum\{\| \mathbf{x} - \mathbf{c}_W \|^2 - \| \mathbf{x} - \mathbf{c}_V \|^2 \mid \mathbf{x} \in V\}.
\end{aligned}$$

Observe that:

$$\begin{aligned}
&\sum\{\| \mathbf{x} - \mathbf{c}_W \|^2 - \| \mathbf{x} - \mathbf{c}_U \|^2 \mid \mathbf{x} \in U\} \\
&= \sum_{\mathbf{x}\in U} \left((\mathbf{x} - \mathbf{c}_W)'(\mathbf{x} - \mathbf{c}_W) - (\mathbf{x} - \mathbf{c}_U)'(\mathbf{x} - \mathbf{c}_U)\right) \\
&= |U|(\mathbf{c}'_W\mathbf{c}_W - \mathbf{c}'_U\mathbf{c}_U) + 2(\mathbf{c}'_U - \mathbf{c}'_W)\sum_{\mathbf{x}\in U}\mathbf{x} \\
&= |U|(\mathbf{c}'_W\mathbf{c}_W - \mathbf{c}'_U\mathbf{c}_U) + 2|U|(\mathbf{c}'_U - \mathbf{c}'_W)\mathbf{c}_U \\
&= |U|(\| \mathbf{c}_W \|^2 - \| c_U \|^2 + 2 \| \mathbf{c}_U \|^2 - 2\mathbf{c}'_W\mathbf{c}_U) \\
&= |U|(\| \mathbf{c}_W \|^2 + \| \mathbf{c}_U \|^2 - 2\mathbf{c}'_W\mathbf{c}_U) \\
&= |U| \| \mathbf{c}_W - \mathbf{c}_U \|^2 .
\end{aligned}$$

Using the equality

$$\mathbf{c}_W - \mathbf{c}_U = \frac{|U|}{|W|}\mathbf{c}_U + \frac{|V|}{|W|}\mathbf{c}_V - \mathbf{c}_U = \frac{|V|}{|W|}(\mathbf{c}_V - \mathbf{c}_U),$$

we obtain

$$\sum\{\| \mathbf{x} - \mathbf{c}_W \|^2 - \| \mathbf{x} - \mathbf{c}_U \|^2 \mid \mathbf{x} \in U\} = \frac{|U||V|^2}{|W|^2} \| \mathbf{c}_V - \mathbf{c}_U \|^2 .$$

In a similar manner we have:

$$\sum\{\| \mathbf{x} - \mathbf{c}_W \|^2 - \| \mathbf{x} - \mathbf{c}_V \|^2 \mid \mathbf{x} \in V\} = \frac{|U|^2|V|}{|W|^2} \| \mathbf{c}_V - \mathbf{c}_U \|^2,$$

so,

$$\text{sse}(W) - \text{sse}(U) - \text{sse}(V) = \frac{|U||V|}{|W|} \| \mathbf{c}_V - \mathbf{c}_U \|^2,$$

Theorem

Let X be a finite set. The function sse : $PART(X) \longrightarrow \mathbb{R}_{\geqslant 0}$ between the posets $(PART(X), \leqslant)$ and $(\mathbb{R}_{\geqslant 0}, \leqslant)$ is monotonic.

Proof: It suffices to show that for $\pi, \pi' \in \text{PART}(X)$, if $\pi \prec \pi'$, then $\text{sse}(\pi) \leqslant \text{sse}(\pi')$. If two blocks U and V of a partition π are fused into a new block W to yield a new partition π' that covers π then, by Lemma on Slide 13 the variation of the sum of squared errors is given by

$$\text{sse}(\pi') - \text{sse}(\pi) = \text{sse}(W) - \text{sse}(U) - \text{sse}(V) = \frac{|U|\,|V|}{|W|} \parallel \mathbf{c}_U - \mathbf{c}_V \parallel^2 \geqslant 0.$$

- The k-means algorithm is a partitional algorithm that requires the specification of the number of clusters k as an input.
- The set of objects to be clustered $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is a subset of $\mathbb{R}^m$.

The Starting Point

- The k-means algorithm begins with a randomly chosen collection of k centroids $\mathbf{c}^1, \ldots, \mathbf{c}^k$ in $\mathbb{R}^m$.
- An initial partition of the set S of objects is computed by assigning each object $\mathbf{x}_i$ to its closest centroid $\mathbf{c}^j$. Let U_j be the set of points assigned to the centroid $\mathbf{c}^j$.
- The assignments of objects to centroids are expressed by a matrix (b_{ij}), where

$$b_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_i \in U_j, \\ 0 & \text{otherwise.} \end{cases}$$

Since each object is assigned to exactly one cluster, we have $\sum_{j=1}^{k} b_{ij} = 1$. Also, $\sum_{i=1}^{n} b_{ij}$ equals the number of objects assigned to the centroid $\mathbf{c}^j$.

Recomputing the Centroids

After these assignments, expressed by the matrix (b_{ij}), the centroids $\mathbf{c}^j$ must be re-computed using the formula:

$$\mathbf{c}^j = \frac{\sum_{i=1}^{n} b_{ij}\mathbf{x}_i}{\sum_{i=1}^{n} b_{ij}} \tag{1}$$

for $1 \leqslant j \leqslant k$.
The sum of squared errors of a partition $\pi = \{U_1, \ldots, U_k\}$ of a set of objects S was defined as

$$sse(\pi) = \sum_{j=1}^{k} \sum_{\mathbf{x} \in U_j} d^2(\mathbf{x}, \mathbf{c}^j),$$

where $\mathbf{c}^j$ is the centroid of U_j for $1 \leqslant j \leqslant k$. The error of such an assignment is the sum of squared errors of the partition $\pi = \{U_1, \ldots, U_k\}$ defined as

$$sse(\pi) \quad = \quad \sum_{i=1}^{n} \sum_{j=1}^{k} b_{ij} ||\mathbf{x}_i - \mathbf{c}^j||^2$$

The mk necessary conditions for a local minimum of this function,

$$\frac{\partial sse(\pi)}{\partial c_p^j} = \sum_{i=1}^{n} b_{ij} \left(-2(\mathbf{x}_p^i - c_p^j)\right) = 0,$$

for $1 \leqslant p \leqslant m$ and $1 \leqslant j \leqslant k$, can be written as

$$\sum_{i=1}^{n} b_{ij} \mathbf{x}_p^i = \sum_{i=1}^{n} b_{ij} c_p^j = c_p^j \sum_{i=1}^{n} b_{ij},$$

or as

$$c_p^j = \frac{\sum_{i=1}^{n} b_{ij} \mathbf{x}_p^i}{\sum_{i=1}^{n} b_{ij}}$$

for $1 \leqslant p \leqslant m$.

In vectorial form, these conditions amount to

$$\mathbf{c}^j = \frac{\sum_{i=1}^{n} b_{ij}\mathbf{x}_i}{\sum_{i=1}^{n} b_{ij}},$$

which is exactly the formula that is used to update the centroids. Thus, the choice of the centroids can be justified by the goal of obtaining local minima of the sum of squared errors of the clusterings.

Since we have new centroids, objects must be reassigned, which means that the values of b_{ij} must be recomputed, which, in turn, affects the values of the centroids, etc.
The halting criterion of the algorithm depends on particular implementations and may involve:

- performing a certain number of iterations;
- lowering the sum of squared errors $sse(\pi)$ below a certain limit;
- the current partition coinciding with the previous partition.

Forgy's Algorithm

Algorithm 1: The k-means Forgy's Algorithm

Data: the set of objects to be clustered $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and the number of clusters k

Result: collection of k clusters

1 extract a randomly chosen collection of k vectors $\mathbf{c}_1, \ldots, \mathbf{c}_k$ in $\mathbb{R}^n$;

2 assign each object $\mathbf{x}_i$ to the closest centroid $\mathbf{c}^j$;

3 let $\pi = \{U_1, \ldots, U_k\}$ be the partition defined by $\mathbf{c}^1, \ldots, \mathbf{c}^k$;

4 recompute the centroids of the clusters $U_1, \ldots, U_k$;

5 **while** *halting criterion is not met* **do**

6 compute the new value of the partition π using the current centroids;

7 recompute the centroids of the blocks of π;

Theorem

The function sse(π) does not increase as the k-means through successive iterations of the Lloyd-Forgy Algorithm.

Proof

Let $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be the set of objects in $\mathbb{R}^m$ to be clustered. Suppose that the partition $\pi = \{C_1, \ldots, C_p, \ldots, C_q, \ldots C_k\}$ was built at a certain stage of the algorithm and let $\pi' = \{C'_1, \ldots, C'_p, \ldots, C'_q, \ldots C'_k\}$ be the partition of X obtained by reassigning an object $\mathbf{x}_r$ from C_p to C_q. We have:

$$C'_i = \begin{cases} C_i & \text{if } i \notin \{p, q\}, \\ C_p - \{\mathbf{x}\} & \text{if } i = p, \\ C_q \cup \{\mathbf{x}\} & \text{if } i = q. \end{cases}$$

Proof cont'd

This reassignment may take place only if $\| \mathbf{x}_r - \mathbf{c}_p \| \geqslant \| \mathbf{x}_r - \mathbf{c}_q \|$. Since

$$\sum \{ \| \mathbf{x} - \mathbf{c}_p \|^2 \mid \mathbf{x} \in C_p \} + \sum \{ \| \mathbf{x} - \mathbf{c}_q \|^2 \mid \mathbf{x} \in C_q \}$$
$$\geqslant \sum \{ \| \mathbf{x} - \mathbf{c}_p \|^2 \mid \mathbf{x} \in C_p - \{\mathbf{x}_r\} \} + \sum \{ \| \mathbf{x} - \mathbf{c}_q \|^2 \mid \mathbf{x} \in C_q \cup \{\mathbf{x}_r\}.$$

We have:

$$\begin{aligned}
\text{sse}(\pi) &= \sum_{j=1}^{k} \sum \{ \| \mathbf{x} - \mathbf{c}_j \|^2 \mid \mathbf{x} \in C_j \} \\
&= \sum \left\{ \sum \{ \| \mathbf{x} - \mathbf{c}_j \|^2 \mid \mathbf{x} \in C_j \} \middle| j \in \{1, \ldots, k\} - \{p, q\} \right\} \\
&\quad + \sum \{ \| \mathbf{x} - \mathbf{c}_p \|^2 \mid \mathbf{x} \in C_p \} + \sum \{ \| \mathbf{x} - \mathbf{c}_q \|^2 \mid \mathbf{x} \in C_q \} \\
&\geqslant \sum \left\{ \sum \{ \| \mathbf{x} - \mathbf{c}_j \|^2 \mid \mathbf{x} \in C_j \} \mid j \in \{1, \ldots, k\} - \{p, q\} \right\} \\
&\quad + \sum \{ \| \mathbf{x} - \mathbf{c}_p \|^2 \mid \mathbf{x} \in C_p - \{\mathbf{x}_r\} \} \\
&\quad + \sum \{ \| \mathbf{x} - \mathbf{c}_q \|^2 \mid \mathbf{x} \in C_q \cup \{\mathbf{x}_r\} \} = \text{sse}(\pi').
\end{aligned}$$

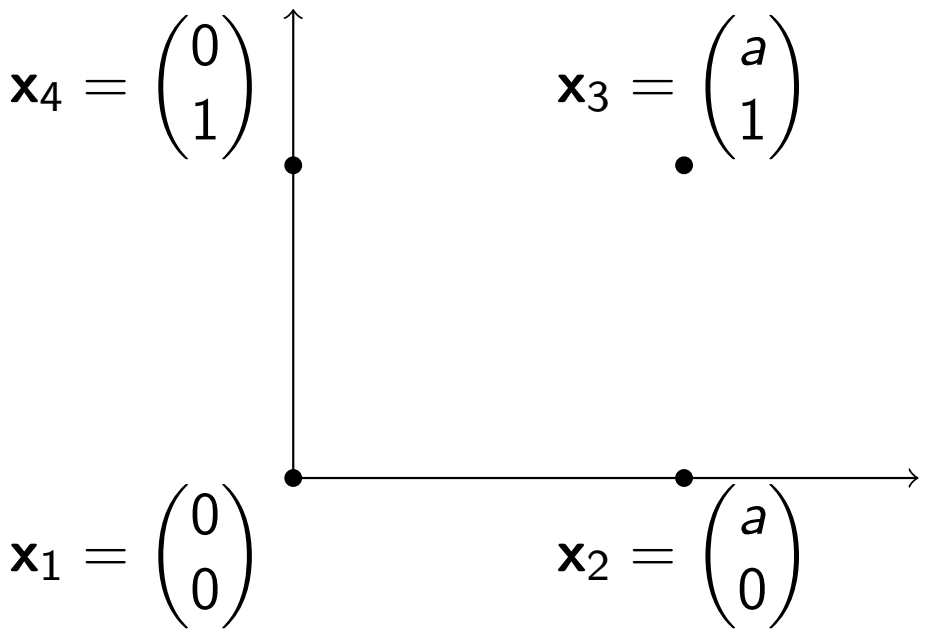
Thus, $\text{sse}(\pi)$ does not increase when $\mathbf{x}_r$ is reassigned.

Example

Consider the set $S = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$ in $\mathbb{R}^n$ given by

$$\mathbf{x}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} a \\ 0 \end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix} a \\ 1 \end{pmatrix}, \mathbf{x}_4 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

shown below.



There are 7 distinct partitions having two blocks on a 4-element set, so there exist seven modalities to cluster these four objects, shown below:

Clusters		centroids		$\text{sse}(\pi)$
C_1	C_2	$\mathbf{c}_1$	$\mathbf{c}_2$	
$\{\mathbf{x}_1\}$	$\{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$	$\mathbf{x}_1$	$\begin{pmatrix} 2a/3 \\ 2/3 \end{pmatrix}$	$\frac{2}{3}(a^2+1)$
$\{\mathbf{x}_2\}$	$\{\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_4\}$	$\mathbf{x}_2$	$\begin{pmatrix} a/3 \\ 2/3 \end{pmatrix}$	$\frac{2}{3}(a^2+1)$
$\{\mathbf{x}_3\}$	$\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4\}$	$\mathbf{x}_3$	$\begin{pmatrix} a/3 \\ 1/3 \end{pmatrix}$	$\frac{2}{3}(a^2+1)$
$\{\mathbf{x}_4\}$	$\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$	$\mathbf{x}_4$	$\begin{pmatrix} 2a/3 \\ 1/3 \end{pmatrix}$	$\frac{2}{3}(a^2+1)$
$\{\mathbf{x}_1, \mathbf{x}_2\}$	$\{\mathbf{x}_3, \mathbf{x}_4\}$	$\begin{pmatrix} a/2 \\ 0 \end{pmatrix}$	$\begin{pmatrix} a/2 \\ 1 \end{pmatrix}$	a^2
$\{\mathbf{x}_1, \mathbf{x}_3\}$	$\{\mathbf{x}_2, \mathbf{x}_4\}$	$\begin{pmatrix} a/2 \\ 1/2 \end{pmatrix}$	$\begin{pmatrix} a/2 \\ 1/2 \end{pmatrix}$	a^2+1
$\{\mathbf{x}_1, \mathbf{x}_4\}$	$\{\mathbf{x}_2, \mathbf{x}_3\}$	$\begin{pmatrix} 0 \\ 1/2 \end{pmatrix}$	$\begin{pmatrix} a \\ 1/2 \end{pmatrix}$	1

If $a \leqslant 1$, the least value of $\text{sse}(\pi)$ is a^2; for $a > 1$, the least value is 1.

- If $a < 1$ and the centroids are $\begin{pmatrix} 0 \\ \frac{1}{2} \end{pmatrix}$ and $\begin{pmatrix} a \\ 1/2 \end{pmatrix}$, then the k-means algorithm will return the clustering $\{\{\mathbf{x}_1, \mathbf{x}_4\}, \{\mathbf{x}_2, \mathbf{x}_3\}\}$ whose $\text{sse}(\pi)$ value is 1 instead of the minimal value a^2.
- If $a > 1$ and the centroids are $\begin{pmatrix} a/2 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} a/2 \\ 1 \end{pmatrix}$, the algorithm returns the partition $\{\{\mathbf{x}_1, \mathbf{x}_2\}, \{\mathbf{x}_3, \mathbf{x}_4\}\}$ and the value of $\text{sse}(\pi)$ for this partition is a^2 instead of the least value of 1.
- These observations show that we may have gaps between the sum-of-squares value of the partition returned by the k-means algorithm and the minimum value of the objective function.

The next theorem shows a limitation of the k-means algorithm because this algorithm produces only clusters whose convex closures may intersect only at the points of S.

Theorem

Let $S = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^m$ be a set of n vectors. If $C_1, \ldots, C_k$ is the set of clusters computed by the k-means algorithm in any step, then the convex closure of each cluster C_i, $K_{\text{conv}}(C_i)$ is included in a polytope P_i that contains c_i for $1 \leq i \leq k$.

Proof:
Suppose that the centroids of the partition $\{C_1, \ldots, C_k\}$ are $\mathbf{c}_1, \ldots, \mathbf{c}_k$. Let $\mathbf{m}_{ij} = \frac{1}{2}(\mathbf{c}_i + \mathbf{c}_j)$ be the midpoint of the segment $\overline{\mathbf{c}_i\mathbf{c}_j}$ and let H_{ij} be the hyperplane $(\mathbf{c}_i - \mathbf{c}_j)'(\mathbf{x} - \mathbf{m}_{ij}) = 0$ that is the perpendicular bisector of the segment $\overline{\mathbf{c}_i\mathbf{c}_j}$.
Equivalently,

$$H_{ij} = \{\mathbf{x} \in \mathbb{R}^m \mid (\mathbf{c}_i - \mathbf{c}_j)'\mathbf{x} = \frac{1}{2}(\mathbf{c}_i - \mathbf{c}_j)'(\mathbf{c}_i + \mathbf{c}_j)\}.$$

The halfspaces determined by H_{ij} are described by the inequalities:

$$\begin{aligned} H_{ij}^+ : (\mathbf{c}_i - \mathbf{c}_j)'\mathbf{x} &\leq \frac{1}{2}(\| \mathbf{c}_i \|_2^2 - \| \mathbf{c}_j \|_2^2) \\ H_{ij}^- : (\mathbf{c}_i - \mathbf{c}_j)'\mathbf{x} &\geq \frac{1}{2}(\| \mathbf{c}_i \|_2^2 - \| \mathbf{c}_j \|_2^2). \end{aligned}$$

Proof cont'd

It is easy to see that $\mathbf{c}_i \in H_{ij}^+$ and $\mathbf{c}_j \in H_{ij}^-$.
Moreover, if $d_2(\mathbf{c}_i, \mathbf{x}) < d_2(\mathbf{c}_j, \mathbf{x})$, then $\mathbf{x} \in H_{ij}^+$, and if $d_2(\mathbf{c}_i, \mathbf{x}) > d_2(\mathbf{c}_j, \mathbf{x})$, then $\mathbf{x} \in H_{ij}^-$. Indeed, suppose that $d_2(\mathbf{c}_i, \mathbf{x}) < d_2(\mathbf{c}_j, \mathbf{x})$, which amounts to $\| \mathbf{c}_i - \mathbf{x} \|_2^2 < \| \mathbf{c}_j - \mathbf{x} \|_2^2$. This is equivalent to

$$(\mathbf{c}_i - \mathbf{x})'(\mathbf{c}_i - \mathbf{x}) < (\mathbf{c}_j - \mathbf{x})'(\mathbf{c}_j - \mathbf{x}).$$

The last inequality is equivalent to

$$\| \mathbf{c}_i \|_2^2 - 2\mathbf{c}_i'\mathbf{x} < \| \mathbf{c}_j \|_2^2 - 2\mathbf{c}_j'\mathbf{x},$$

which implies that $\mathbf{x} \in H_{ij}^+$. In other words, $\mathbf{x}$ is located in the same half-space as the closest centroid of the set $\{\mathbf{c}_i, \mathbf{c}_j\}$. Note also that if $d_2(\mathbf{c}_i, \mathbf{x}) = d_2(\mathbf{c}_j, \mathbf{x})$, then $\mathbf{x}$ is located in $H_{ij}^+ \cap H_{ij}^- = H_{ij}$, that is, on the hyperplane shared by P_i and P_j.

Proof cont'd

Let P_i be the closed polytope defined by

$$P_i = \bigcap \{H_{ij}^+ \mid j \in \{1, \ldots, k\} - \{i\}\}$$

Objects that are closer to $\mathbf{c}_i$ than to any other centroid $\mathbf{c}_j$ are located in the closed polytope P_i. Thus, $C_i \subseteq P_i$ and this implies $\mathbf{K}_{\text{conv}}(C_i) \subseteq P_i$.

A bit of linear algebra recall:
The Frobenius norm of a matrix $A \in \mathbb{C}^{m \times n}$ is

$$\| A \|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}^2}.$$

It is easy to see that for $A \in \mathbb{C}^{m \times n}$ we have

$$\| A \|_F^2 = \text{trace}(AA') = \text{trace}(A'A)$$

because

$$\begin{aligned} \text{trace}(AA') &= \sum_{i=1}^{n} (AA')_{ii} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}^2 = \| A \|_F^2 . \end{aligned}$$

Definition

Let $f : \mathbb{R}^{m\times n} \longrightarrow \mathbb{R}$ be a function. The derivative of f with respect to the matrix $X \in \mathbb{R}^{m\times n}$ is the function $\frac{\partial f}{\partial X} : \mathbb{R}^{m\times n} \longrightarrow \mathbb{R}^{m\times n}$ given by

$$\frac{\partial f}{\partial X}(X) = \begin{pmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{12}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \frac{\partial f}{\partial x_{m2}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{pmatrix}.$$

Example

Let $f : \mathbb{R}^{m\times n} \longrightarrow \mathbb{R}$ be defined by $f(X) = \text{trace}(XAX')$, where $X \in \mathbb{R}^{m\times n}$ and $A \in \mathbb{R}^{n\times n}$. Since

$$f(X) = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{n} x_{ij}a_{jk}x_{ik},$$

we have:

$$\begin{aligned}\frac{\partial f}{\partial x_{pq}} &= \sum_{k=1}^{n} a_{qk}x_{pk} + \sum_{j=1}^{n} x_{pj}a_{jq} \\ &= (XA')_{pq} + (XA)_{pq} = (X(A + A'))_{pq},\end{aligned}$$

which implies

$$\frac{\partial f}{\partial X} = X(A + A').$$

Example

Let $f : \mathbb{R}^{m\times m} \longrightarrow \mathbb{R}$ be the function defined by $f(X) = \text{trace}(AXB)$, where $A \in \mathbb{R}^{m\times p}$, $X \in \mathbb{R}^{p\times n}$, and $B \in \mathbb{R}^{n\times m}$. Note that

$$f(X) = \sum_{i=1}^{m}(AXB)_{ii} = \sum_{i=1}^{m}\sum_{j=1}^{p}\sum_{k=1}^{n} a_{ij}x_{jk}b_{ki},$$

hence

$$\frac{\partial f}{\partial x_{jk}} = \sum_{i=m}^{n} a_{ij}b_{ki} = (BA)_{kj} = (A'B')_{jk}.$$

Therefore, $\frac{\partial}{\partial X}\text{trace}(AXB) = A'B'$.

Example

For $g(X) = \text{trace}(AX'B)$ we have

$$g(X) = \sum_{i=1}^{n} (AX'B)_{ii} = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} a_{ij} x_{kj} b_{ki},$$

which implies $\frac{\partial g}{\partial x_{kj}} = \frac{\partial f}{\partial x_{jk}}$. Therefore, we have:

$$\frac{\partial \text{trace}(AX'B)}{\partial X} = (A'B')' = BA.$$

Example

Let $f : \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}$ be the function defined by $f(X) = \text{trace}(B'X'XB)$, where $B, X \in \mathbb{R}^{n\times n}$. Since

$$(B'X'XB)_{ij} = \sum_{p=1}^{n}\sum_{q=1}^{n}\sum_{r=1}^{n} b_{pi}x_{qp}x_{qr}b_{rj}$$

we have:

$$\begin{aligned}\text{trace}(B'X'XB) &= \sum_{i=1}^{n}(B'X'X'B)_{ii} \\ &= \sum_{i=1}^{n}\sum_{p=1}^{n}\sum_{q=1}^{n}\sum_{r=1}^{n} b_{pi}x_{qp}x_{qr}b_{ri}.\end{aligned}$$

Example cont'd

Thus, the partial derivative $\frac{\partial f}{\partial x_{uv}}$ can be written as

$$\begin{aligned}
\frac{\partial f}{\partial x_{uv}} &= \sum_{i=1}^{n}\sum_{r=1}^{n} b_{vi}x_{ur}b_{ri} + \sum_{i=1}^{n}\sum_{p=1}^{n} b_{pi}x_{uv}b_{vi} \\
&= \sum_{i=1}^{n}\sum_{r=1}^{n} b_{vi}x_{ur}b_{ri} + \sum_{i=1}^{n}\sum_{p=1}^{n} b_{pi}x_{uv}b_{vi} \\
&= \sum_{i=1}^{n}\sum_{r=1}^{n} b_{vi}x_{ur}b_{ri} + \sum_{i=1}^{n}\sum_{r=1}^{n} b_{ri}x_{uv}b_{vi} \\
&\quad \text{(by changing the summation index } p \text{ in the second sum to } r\text{)} \\
&= \sum_{i=1}^{n}\sum_{r=1}^{n}(x_{ur}b_{ri}b_{vi} + x_{uv}b_{vi}b_{ri}).
\end{aligned}$$

This allows us to write $\frac{\partial f}{\partial X} = 2XBB'$.

Example

Let $f : \mathbb{R}^{m\times k}$ be the function $f(X) = \| A - XB \|_F^2$, where $A \in \mathbb{R}^{m\times n}$, $X \in \mathbb{R}^{m\times k}$, and $B \in \mathbb{R}^{k\times n}$. We have:

$$\begin{aligned} f(X) &= \| A - XB \|_F^2 = \text{trace}((A - XB)'(A - XB)) \\ &= \text{trace}(A'A) - 2\text{trace}(A'XB) + \text{trace}(B'X'XB). \end{aligned}$$

By Example on Slide 40 we have $\frac{\partial A'XB}{\partial X} = AB'$; by Example on Slide 42 we have $\frac{\partial B'X'XB}{\partial X} = 2XBB'$, hence

$$\frac{\partial f(M)}{\partial M} = 2(XBB' - AB').$$

Thus, to minimize f we must have $X = AB'(BB')^{-1}$.

Let $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, where $S \subseteq \mathbb{R}^m$ be the set of objects to be clustered by the k-means algorithm, and let $C = \{\mathbf{c}_1, \ldots, \mathbf{c}_k\}$ be a subset of $\mathbb{R}^m$. For $1 \leqslant i \leqslant$ define the subset C_i of S as consisting of those members of S for which the closest point in C is $\mathbf{c}_i$ (such that ties between distances $d(\mathbf{x}, \mathbf{c}_i)$ and $d(\mathbf{x}, \mathbf{c}_j)$ are broken arbitrarily). The collection $\{C_1, \ldots, C_k\}$ is a partition π_C of S.

The k-means algorithm entails choosing the elements of C, to accomplish the minimization of the objective function

$$\text{sse}(\pi_C) = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} \| \mathbf{x}_j - \mathbf{c}_i \|^2 = \sum_{i=1}^{k} \sum_{j=1}^{n} z_{ij} \| \mathbf{x}_j - \mathbf{c}_i \|^2,$$

where

$$z_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_j \in C_i, \\ 0 & \text{otherwise} \end{cases}$$

are binary variables that indicate whether or not a data point $\mathbf{x}_j$ belongs to C_i. $Z = (z_{ij})$ is a binary matrix that belongs to $\{0, 1\}^{k \times n}$. The first index i in z_{ij} designates the cluster; the second designates the object.

To express the fact that a given cluster C_i contains n_i objects we write $\sum_{j=1}^{n} z_{ij} = n_i$ for $1 \leqslant i \leqslant k$. On other hand, every object belongs to exactly one cluster, so $\sum_{i=1}^{k} z_{ij} = 1$ for $1 \leqslant j \leqslant n$. In matrix form these conditions amount to

$$Z\mathbf{1}_n = \begin{pmatrix} n_1 \\ \vdots \\ n_k \end{pmatrix},$$

and $Z'\mathbf{1}_k = \mathbf{1}_n$. The matrix Z describes completely the assignment of objects to clusters.

The rows of Z are pairwise orthogonal due to the fact that each object $\mathbf{x}_j$ belongs exactly to one cluster. Therefore, for $i \neq i'$ we have $z_{i'j} z_{ij} = 0$ for every j, $1 \leqslant j \leqslant n$. In turn, this implies that $ZZ' \in \mathbb{R}^{k \times k}$ is a diagonal matrix where

$$(ZZ')_{ii'} = \sum_j (Z)_{ij} (Z')_{ji'} = \sum_j z_{ij} z_{i'j} = \begin{cases} n_i & \text{if } i = i', \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Therefore,

$$(ZZ')^{-1} = \begin{pmatrix} \frac{1}{n_1} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{n_k}. \end{pmatrix}$$

Let $Y = Z'(ZZ')^{-1} \in \mathbb{R}^{n\times k}$. The columns of the matrix Y correspond to the clusters $C_1, \ldots, C_k$ and $\sum_{j=1}^{n} y_{ij} = 1$ for $1 \leqslant i \leqslant n$. Since

$$y_{ji} = \sum_{\ell=1}^{k} (Z')_{j\ell}((ZZ')^{-1})_{\ell i} = \sum_{\ell=1}^{k} z_{\ell j}((ZZ')^{-1})_{\ell i} = z_{ij}\frac{1}{n_i},$$

it follows that $\sum_{j=1}^{n} y_{ji} = 1$. In other words, the components of each column $\mathbf{y}_i$ of Y are non-negative numbers that sum up to 1, so they can be regarded as probability distributions.

Let $X = (\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_n) \in \mathbb{R}^{m\times n}$ be a matrix whose columns are the data points of the set S. The set C is represented by the matrix

$$M = (\mathbf{c}_1\ \mathbf{c}_2\ \cdots\ \mathbf{c}_k) \in \mathbb{R}^{m\times k}.$$

The Frobenius norm of the matrix X is given by:

$$\| X \|^2 = \sum_{j=1}^{n} \| \mathbf{x}_j \|^2 = \sum_{j=1}^{n} \mathbf{x}_j'\mathbf{x}_j = \sum_{j=1}^{n} (X'X)_{jj} = \text{trace}(X'X).$$

The next theorem shows that to minimize $\text{sse}(\pi_C)$ amounts to minimizing the norm of the matrix $X - MZ$, where $M \in \mathbb{R}^{m \times k}$ and $Z \in \mathbb{R}^{k \times n}$, that is, to find the best approximation of X as product MZ.

Theorem

(Baukhage's Factorization Theorem) *The following equality holds:*

$$\sum_{i=1}^{k} \sum_{j=1}^{n} z_{ij} \parallel \mathrm{x}_j - \mathrm{c}_i \parallel^2 = \parallel X - MZ \parallel^2 .$$

Proof

The left-hand member of the equality of the theorem can be written as

$$\begin{aligned}
&\sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij} \parallel \mathbf{x}_j - \mathbf{c}_i \parallel^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij}(\mathbf{x}_j - \mathbf{c}_i)'(\mathbf{x}_j - \mathbf{c}_i) \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij}(\mathbf{x}_j'\mathbf{x}_j - 2\mathbf{x}_j'\mathbf{c}_i + \mathbf{c}_i'\mathbf{c}_i) \\
&= T_1 - 2T_2 + T_3,
\end{aligned}$$

where

$$T_1 = \sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij}\mathbf{x}_j'\mathbf{x}_j,\ T_2 = \sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij}\mathbf{x}_j'\mathbf{c}_i,$$

and $T_3 = \sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij}(\mathbf{c}_i'\mathbf{c}_i)$.

We can further write

$$\begin{aligned} T_1 &= \sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij}\mathbf{x}_j'\mathbf{x}_j = \sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij} \parallel \mathbf{x}_j \parallel^2 = \sum_{j=1}^{n} \parallel \mathbf{x}_j \parallel^2 \sum_{i=1}^{k} z_{ij} \\ &= \sum_{j=1}^{n} \parallel \mathbf{x}_j \parallel^2 = \text{trace}(X'X). \end{aligned}$$

Next, we have:

$$\begin{aligned} T_2 &= \sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij}\mathbf{x}_j'\mathbf{c}_i \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij}\sum_{\ell=1}^{m} x_{\ell j}c_{\ell i} = \sum_{j=1}^{n}\sum_{\ell=1}^{m} x_{\ell j}\sum_{i=1}^{k} z_{ij}c_{\ell i} \\ &= \sum_{j=1}^{n}\sum_{\ell=1}^{m} x_{\ell j}(MZ)_{\ell j} = \sum_{j=1}^{n}\sum_{\ell=1}^{m}(X')_{j\ell}(MZ)_{\ell j} \\ &= \sum_{j=1}^{n}(X'MX)_{jj} = \text{trace}(X'MZ). \end{aligned}$$

Finally,

$$\begin{aligned} T_3 &= \sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij}\mathbf{c}_i'\mathbf{c}_i = \sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij} \parallel \mathbf{c}_i \parallel^2 \\ &= \sum_{i=1}^{k} \parallel \mathbf{c}_i \parallel^2 \sum_{j=1}^{n} z_{ij} = \sum_{i=1}^{k} \parallel \mathbf{c}_i \parallel^2 n_i, \end{aligned}$$

where $n_i = |C_i|$.
For the right-hand member of the equality of the theorem we have

$$\begin{aligned} \parallel X - MZ \parallel^2 &= \text{trace}((X - MZ)'(X - MZ)) \\ &= \text{trace}(X'X) - 2\text{trace}(X'MZ) + \text{trace}(Z'M'MZ) \\ &= T_1 - 2T_2 + T_4, \end{aligned}$$

where $T_4 = \text{trace}(Z'M'MZ)$.

For the right-hand member of the equality of the theorem we have:

$$\begin{aligned} \| X - MZ \|^2 &= \text{trace}((X - MZ)'(X - MZ)) \\ &= \text{trace}(X'X) - 2\text{trace}(X'MZ) + \text{trace}(Z'M'MZ) \\ &= T_1 - 2T_2 + T_4, \end{aligned}$$

where $T_4 = \text{trace}(Z'M'MZ)$. Now, we have

$$\begin{aligned} T_4 &= \text{trace}(Z'M'MZ) = \text{trace}(M'MZZ') \\ &\quad \text{(due to the cyclic permutation invariance of the trace)} \\ &= \sum_{i=1}^{k} (M'MZZ')_{ii} = \sum_{i=1}^{k} \sum_{\ell=1}^{m} (M'M)_{i\ell} (ZZ')_{\ell i} \\ &= \sum_{i=1}^{k} (M'M)_{ii} (ZZ')_{ii} = \sum_{i=1}^{k} \| \mathbf{c}_i \|^2 \, n_i. \end{aligned}$$

Thus, $T_4 = T_3$, and this completes the argument.

The centroid matrix $M = (\mathbf{c}_1\ \mathbf{c}_2\ \cdots\ \mathbf{c}_k)$ that minimizes the objective function

$$F(M) = \| X - MZ \|^2$$

is obtained, by Example on Slide 44 as

$$M = XZ'(ZZ')^{-1} = XY,$$

where Y is the matrix $Y = Z'(ZZ')^{-1} \in \mathbb{R}^{n \times k}$ previously introduced.