# More details of K-means clustering

**K-means clustering algorithm**

0. Start with initial guesses for cluster centers (centroids)
1. For each data point, find closest cluster center (partitioning step)
2. Replace each centroid by average of data points in its partition
3. Iterate 1+2 until convergence

(See Fig 14.4, 14.6)

Write $x_i = (x_{i1}, ...x_{ip})$:

If centroids are $m_1, m_2, ...m_k$, and partitions are

$c_1, c_2, ...c_k$, then one can show that K-means converges to a *local* minimum of

$$\sum_{k=1}^{K} \sum_{i \in c_k} ||x_i - m_k||^2 \qquad \text{Euclidean distance}$$

(within cluster sum of squares)

**In practice:**

- Try many random starting centroids (observations) and choose solution with smallest of squares

**How to choose K?**

- Difficult – details later

## Stepping back

- All clustering algorithms start with a dissimilarity measure for $j^{th}$ feature

$$d_j(x_{ij}, x_{i'j}) \text{ and define}$$

$$D(x_i, x_{i'}) = \sum_{j=1}^{P} d_j(x_{ij}, x_{i'j})$$

Usually $d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2$

**Other possibilities:**

- Correlation

$$\rho(x_i, x_{i'}) = \frac{\sum_j (x_{ij} - \overline{x}_i)(x_{i'j} - \overline{x}_{i'})}{\sqrt{\sum_j (x_{ij} - \overline{x}_i)^2 \sum_j (x_{i'j} - \overline{x}_{i'})^2}}$$

$\overline{x}_i$ = mean of observation $i$

- If observations are standardized:

$$x_{ij} \leftarrow \frac{x_{ij} - \bar{x}_i}{\sqrt{\sum_j (x_{ij} - \bar{x}_i)^2}}$$

then $2(1 - \rho(x_i, x_{i'})) = \sum_j (x_{ij} - x_{i,j'})^2$

So clustering via correlation $\equiv$ clustering via Euclidean distance with standardized features

- **Categorical** features - usually coded as dummy variables

$$
\begin{aligned}
\text{e.g. } \mathrm{X}_1 = 1, 2 \text{ or } 3 \quad &\rightarrow \quad (1\ 0\ 0) \\
& \quad\quad\; (0\ 1\ 0) \\
& or \quad (0\ 0\ 1)
\end{aligned}
$$

- **Weighting** is also possible (see chapter 14)

## Partitioning (Clustering) Algorithms

- Group assignment function ("encoder") $C(i)$

  $C : 1, 2, ...N \rightarrow (1, 2, ...K)$

- **Criterion**: choose $C$ to minimize

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} d(x_i, x_{i'})$$

(within cluster scatter)

**Fact**:

- $K$-means minimizes $W(C)$ when $D = ||x_i - x_{i'}||^2$

$$
\begin{aligned}
W(C) &= \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k} ||x_i - x_{i'}||^2 \\
&= \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - \overline{x}_k||^2
\end{aligned}
$$

- K-means solves *enlarged* problem:

$$
\min_{C, m_1 \ldots m_k} \sum_{k} \sum_{C(i)=k} ||x_i - m_k||^2
$$

to find assignment function $C$