Clustering is unsupervised learning.

Training set $S = \{ x^{(i)} \}_{i=1}^n$.

"Cluster Data": Group examples that are similar in one cluster and keep examples that are dissimilar in different clusters.

Clustering setup:

Input: $S = \{ x^{(i)} \}_{i=1}^n$ and distance metric $d$.

$p: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, $k$ clusters.

Output: $C_1, C_2, C_3 \ldots C_k$ s.t. $C_i \cap C_j = \emptyset \quad \forall i \neq j$.

$C_1 \cup C_2 \cup \ldots \cup C_k = S$.

"$C_1, C_2, C_3 \ldots C_k$ is a partition of $S$".
**k-means Clustering**: On input $S$, $p$, $k$.

Goal is find $C_1, C_2, \ldots, C_k$ s.t.

$$\min \sum_{j=1}^{k} \sum_{x \in C_j} p(x^{(i)}, M_j)$$

$$M_j = \frac{\sum_{x \in C_j} x^{(i)}}{|C_j|}$$

NP-hard.

**k-means Clustering Algorithm**:

1. Identify centers $M_1, \ldots, M_k$.
2. $y^{(i)}$ identifies the cluster to which $x^{(i)}$ belongs.

Input: $S$, $p$, $k$.

Initially: Let $M_1, M_2, \ldots, M_k$ be some centers.

While ($y \neq y^{(i)}$) do

- $y^{(i)} = \frac{1}{k} \sum_{j=1}^{k} \frac{\sum_{x^{(i)} \in C_j} x^{(i)}}{|C_j|}$

- $y^{(i)} = \text{argmin}_{j \in \{1, \ldots, k\}} p(x^{(i)}, M_j)$

- $y^{(i)} = \frac{1}{k} \sum_{i=1}^{n} 1[y^{(i)} = j] x^{(i)}$

Output $y^{(i)}$

# clusters $= n = |S|$. (Put each point in its own cluster.)
When the data consists of well separated spherical groups of the same size, this algorithm does well.

Linkage based clustering

Input: \( S, p \).
Initialize: \( C = \{ x^{(i)} \mid i \in \{1, 2, \ldots, n\} \} \) → every point in its own cluster.

Repeat:

\[ C^*, D^* = \text{argmin}_{C, D \in E} \text{distance}(C, D) \]

\[ \ell = (C - \{C^*, D^*\}) \cup \{C^*, D^*\} \]

until

\[ |C^*| = k. \]
\[ |C^*| = 1 \]

Linkage clustering:

- Single linkage clustering
  \[ \text{distance}(C, D) = \min_{x \in C, y \in D} p(x, y) \]
- Average linkage clustering
  \[ \text{distance}(C, D) = \frac{1}{|C| \cdot |D|} \sum_{x \in C} \sum_{y \in D} p(x, y) \]
DENDROGRAM.