Assignment-based Clustering

Input: data $X$ and dist $d : X \times X \to \mathbb{R}$

Centers
$C = \{c_1, c_2, c_3, c_4\}$

$\phi_C(x) = \arg\min_{c \in C} d(x, c)$

Voronoi Diagram

$X \rightarrow S_1, S_2, \ldots, S_n$

$S_i \cap S_j = \emptyset$

$X = S_1 \cup S_2 \cup \ldots \cup S_n$

$S_j = \{ x \in X \mid \phi_C(x) = c_j \}$
\[ \text{Cost}_k(X, C) = \sum_{x \in X} d(\phi_c(x), x)^2 \]

\[ \mathcal{C} \leftarrow \text{minimize Cost}_k \]

\( k \)-means clustering formulation

\[ \text{Cost}_\infty(X, C) = \max_{x \in X} d(\phi_c(x), x) \]

\( k \)-center (Gonzalez)

\[ \text{Cost}_1(X, C) = \sum_{x \in X} d(\phi_c(x), x) \]

\( k \)-median

\( k \)-medioid

\[ \text{minimize Cost}_1(X, C) \]

s.t. \( C \subseteq \text{subset} \)
Gonzalez Algo. 12-Center outliers poorly w/ outliers

\[ \hat{C} \subset X, d \text{ metric} \]

\[ C^* \subset \text{optimal} = \arg\min \text{cost}_D(X, C) \]

D. Choose \( c, c \in X \) arbitrarily

1. \( \text{for } j = 2 \text{ to } n \)

\[ \text{Set } c_j = \arg\max_{x \in X} d(x, \phi_{c_{j-1}}(x)) \]

\[ \hat{C} \subset \tilde{C} \subset X \]

\[ C_1 = \{ c_1, \ldots, c_n \} \]

\[ C_{j-1} = \{ c_1, \ldots, c_j \} \]

Provides 2-approximation

\[ \text{cost}_D(X, \hat{C}) \leq 2 \cdot \text{cost}_D(X, C^*) \]
**K-means clustering**

Lloyd's Algorithm:

Choose **k** prototype vectors \( \mathbf{c}_1, \ldots, \mathbf{c}_k \) randomly from the data.

1. **Repeat**
   - **For all** \( x \in X \) assign \( \phi_c(x) \) to \( \mathbf{c}_j \) if \( d(x, \mathbf{c}_j) \leq d(x, \mathbf{c}_i) \) for all \( i \neq j \).
   - **Let** \( c_j = \text{average} \{ \phi_c(x) \} \)

2. Until \( C \) is fixed

Cost of assignment:

\[
\text{Cost}_k(X,C) = \sum_{x \in X} \min_{j} d(x, c_j) = \sum_{j} \left( \sum_{x \in S_j} d(x, c_j) \right)
\]
$k$-means++

$L$-initialize Lloyds Also

\[ C_j = \{ c_1, c_2, \ldots, c_j \} \]

0. Choose \( c_1 \in X \) arbitrarily \( C_1 = \{ c_1 \} \)

1. for \( j = 2 \) to \( k \)

Choose \( c_i \) from \( X \) w/ prob proportional to \( d(x, \phi_{c_{i-1}}(x))^2 \)

Partition of Unity

\[ \sum_{i=1}^{k} \omega_i = 1 \]

\[ \omega_i = \frac{w(x_i)}{\sum_{i=1}^{k} w(x_i)} \]

Choose \( \omega = 0.432 \) \( \omega_j \) \( \omega(x_j) \)

M4D Sec 2.41
Choosing $\hat{K}$

"elbow technique"