Fo DA . Clustering
L23 . Lloyd's Algorithm for k-means
R-Means Clustering Formulation

Input
\[ X \subseteq \mathbb{R}^d \]
\[ X = \{ x_1, x_2, \ldots, x_n \} \]
\[ k \in \mathbb{Z}^+ \]
\[ \mathcal{E} \subseteq \{ 1, 2, \ldots, k \} \]
\[ d: \times \times \times \rightarrow \mathbb{R}^n \]
\[ d(x_i, x_j) = \| x_i - x_j \| \]

Goal
Find set
\[ S = \{ s_1, s_2, \ldots, s_k \} \subseteq \mathbb{R}^d \]
minimize
\[ \sum_{i=1}^{\infty} d(x_i, \phi_S(x_i)) \]
\[ k = 3 \]

maps \( x_i \)
do
\[ \arg \min_{s_j \in S} \{ x_i - s_j \} \]
Lloyd's Algorithm

0. Initialize k points $S \subseteq \mathbb{X}$

1. Repeat
   a. For all $x_i \in \mathbb{X}$: assign $x_i$ to $X_j$ so \( \phi_s(x_i) = j \)
   b. For all $s_j \in S$: update \( s_j = \frac{1}{|X_j|} \sum_{x_i \in X_j} x_i \) until (the set $S$ unchanged or change is small)

\[ X = \bigcup_{j=1}^{k} X_j \]

(carbitrarily?)
Lloyd's Algo is stuck

Most of the time, Lloyd's works well (with a little help)

$b = 4$

optimal $S = \{s_1, s_2, s_3, s_4\}$

suboptimal $T = \{t_1, t_2, t_3, t_4\}$

$Lloyds$
Tricks to help Lloyd's

- Random Restarts
  a. Randomly initialize $S$.
  b. Run Lloyd's to compute cost (CSE).
  c. Repeat (a, b) say 10 times.
  d. Return final $S$ w/ lowest cost.

- Better Initialize
  a. Gonzalez Algo
  b. $k$-means++
Corner Cases

- Might be site w/ no points closest to it

→ Randomly assign \(s_q\) (w/ no points in) from \(X_j\), w/ most variance.

\[\rightarrow\] some \(x_i \in X\).
Most data looks like

No blobs!

Smear
Number of Clusters, \( k \)?

\[
\text{Cost}(X, S) = \sum_{i=1}^{n} \| x_i - \phi_s(x_i) \|^2
\]

\( k \) is chosen where the curve flattens out.
Why does Lloyd's Algorithm converge?

$\text{cost}(x, S) = \frac{1}{3} \sum_{x \in X} \| x - \phi(x) \|^2$

$= \frac{1}{3} \sum_{s \in S} \sum_{i \in S} \| x - \phi(s) \|^2$

(a) assignment

(b) centering