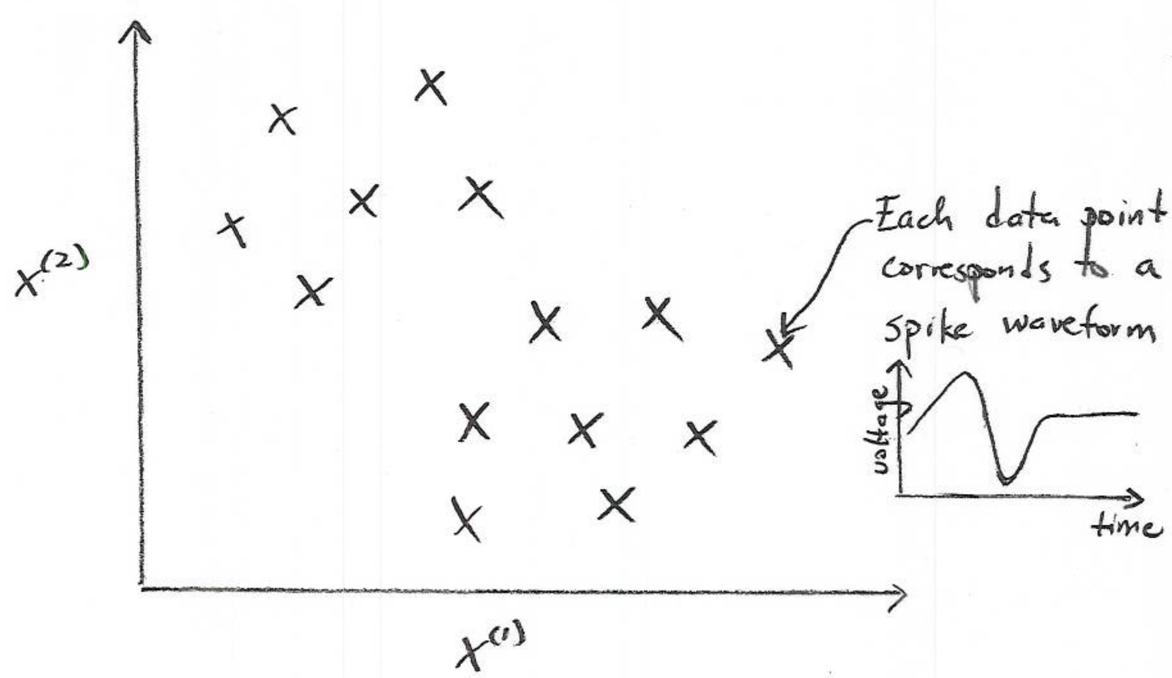
A) K-means Clustering

Suppose we have a data set $\underline{x}_n \in \mathbb{R}^D$ where $n = 1, \ldots, N$.

Goal: Partition the data set into some number K of clusters.

For now, assume K is given.

Picture to have in mind:



How would you partition this dataset into $K = 2$ clusters?

Intuitive definition of a cluster:

A group of data points whose inter-point distances are small compared with the distances to points outside the cluster.

Let's formalize this notion:

Let $\underline{\mu}_k \in \mathbb{R}^D$ where $k = 1, \ldots, K$ be the "prototype" associated with the kth cluster.

$$r_{nk} = \begin{cases} 1 & \text{if } \underline{x}_n \text{ belongs to the } k\text{th cluster} \\ 0 & \text{else} \end{cases}$$

objective function

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \| \underline{x}_n - \underline{\mu}_k \|^2$$

Find $\{r_{nk}\}$ and $\{\underline{\mu}_k\}$ such that J is minimized.

This can be solved by iterating:

- Minimize J w.r.t. r_{nk}, keeping $\underline{\mu}_k$ fixed ("E-step")
- Minimize J w.r.t. $\underline{\mu}_k$, keeping r_{nk} fixed ("M-step")

A.1) E-step for K-means

Constraint: $\{r_{n1}, \ldots, r_{nK}\}$ is a set of $(K-1)$ zeros and a single 1.

Can optimize J w.r.t. r_{nk} for each n separately.

For each n, assign:

$$r_{nk} = \begin{cases} 1 & \text{if } k = \underset{j}{\operatorname{argmin}} \|\underline{x}_n - \underline{\mu}_j\|^2 \\ 0 & \text{else} \end{cases}$$

In words: Assign each data point to the closest cluster center.

A.2) M-step for K-means

$$\frac{\partial J}{\partial \underline{\mu}_k} = 2\sum_{n=1}^{N} r_{nk}(\underline{x}_n - \underline{\mu}_k) = 0$$

$$\boxed{\underline{\mu}_k = \frac{\sum_{n=1}^{N} r_{nk}\underline{x}_n}{\sum_{n=1}^{N} r_{nk}}}$$

In words: set $\underline{\mu}_k$ equal to the mean of all data points assigned to cluster k.

A.3) Convergence of the K-means algorithm.

- The E-step and M-step should be iterated until there is no further change in cluster assignments, or until some maximum number of iterations is exceeded.

- Each iteration reduces the objective function J, so the algorithm is <u>guaranteed to converge</u>. (We will prove this later in the context of the EM algorithm.)

- Convergence is guaranteed to a <u>local</u> (rather than a global) minimum of J.

- If there are multiple local optima, the particular local optimum reached depends on the <u>initialization</u> of the $\{\underline{\mu}_k\}$.