Clustering

K-means

- Random cluster centroid
- Each iteration assigning points to each centroid
- Each iteration moving the centroid to the centre of its group

K = number of clusters

Training set = $\{x^1, x^2 \dots x^m\}$; $x_i \in \mathbb{R}^n$

Randomly initialize K cluster centroids $\mu_1, \mu_2 \dots \mu_K$, $k \in \mathbb{R}^n$

For $i = [1, m]$

$c^i :=$ index (from 1 to K) of cluster centroid closest to $x^i \rightarrow \min_k \|x^i - \mu_k\|^2$

For $k = 1$ to K:

$\mu_k :=$ mean of points assigned to cluster k

- Optimization objective (distortion) $\rightarrow J(c^1 \dots c^m, \mu_1 \dots \mu_K) = \frac{1}{m} \sum_{i=1}^{m} \|x^i - \mu_{c^i}\|^2$

minimize it

- c^i = index of cluster $[1, K]$ to which example x^i is currently assigned
- μ^k = cluster centroid k, $\mu_k \in \mathbb{R}^n$
- μ_{c^i} = cluster centroid of cluster to which example x^i has been assigned

- Random initialization

  - should have $k < m$
  - randomly pick K training examples
  - set $\mu_1 \dots \mu_k = x^{(i)} \dots x^{(j)}$

- choosing the value of k (the elbow)