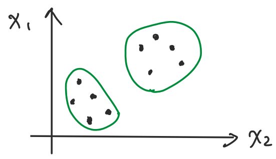
- Unsupervised learning introduction
  - Unsupervised learning



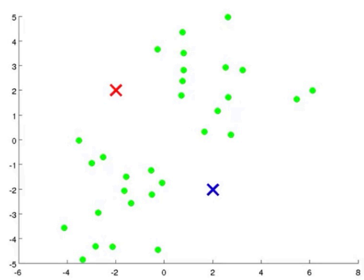
Clustering algorithm

  - Applications of clustering

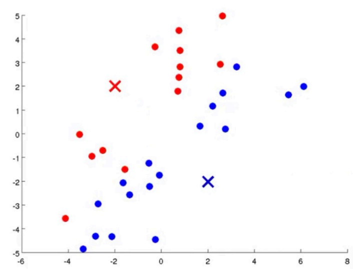
    e.g. Market segmentation

    Social network analysis

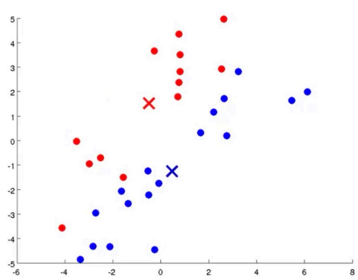
- K-means algorithm



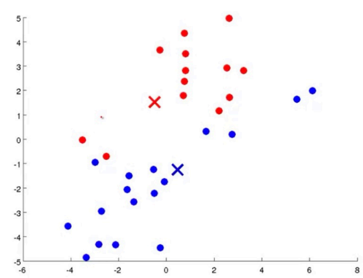
→



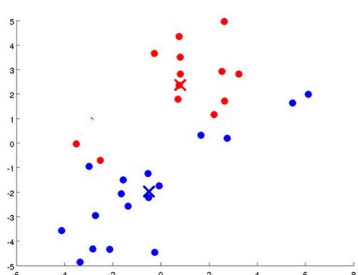
隨機取兩點(cluster centroids)　　距離紅點近染為紅色，反之



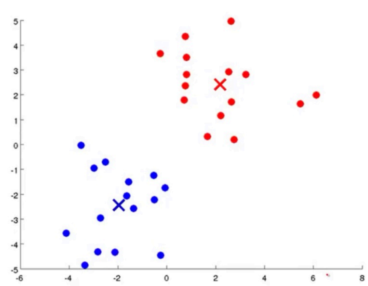
→



紅色圓點計算 cluster centroids 新的位置　　距離紅點近染為紅色，反之



→



紅色圓點計算 cluster centroids 新的位置　　距離紅點近染為紅色，反之 收斂

- K-means alogrithm

input : K (number of clusters)

Training Set $\{x^{(1)}, x^{(2)}, \dots x^{(m)}\}$

$x^{(i)} \in R^n$ (drop $x_0 = 1$ convention)

Randomly initialize K cluster centroids $\mu_1, \mu_2, \dots, \mu_k \in R^n$

Repeat {

$\boxed{\begin{array}{l} \text{for } i = 1 \text{ to } m \\ \quad c^{(i)} = \text{index (for 1 to k) of centroid closest to } x^{(i)} \end{array}}$

cluster assignment step
minimize $J(\dots)$ cost $c^{(1)}, \dots, c^{(m)}$
(holding $\mu_1, \dots, \mu_k$ fixed)

$\min_k \| x^{(i)} - \mu_k \|^2$

for k = 1 to K

move centroid

μ_k = average(mean) of points assigned to cluster k

$x^{(1)}, x^{(5)}, x^{(6)}, x^{(10)} \rightarrow c^{(1)} = 2,\ c^{(5)} = 2,\ c^{(6)} = 2,\ c^{(10)} = 2$

}

$\mu_2 = \frac{1}{4}[x^{(1)} + x^{(5)} + x^{(6)} + x^{(10)}] \in R^n \rightarrow$ location of cluster centroid

- Optimization objective
  - K-means optimization objective

$\mu_{c^{(i)}}$ = cluster centroid of cluster to which example $x^{(i)}$ has been assigned

$x^{(i)} \rightarrow 5 \quad c^{(i)} = 5 \quad \mu_{c^{(i)}} = \mu_5$

Optimization objective :

$$J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots \mu_k) = \frac{1}{m} \sum_{i=1}^{m} \| x^{(i)} - \mu_{c^{(i)}} \|^2$$

$$\min_{\substack{c^{(1)} \dots c^{(m)} \\ \mu_1, \dots, \mu_k}} J(c^{(1)}, \dots c^{(m)}, \mu_1, \dots, \mu_k)$$

- Random initialization 如何避開局部最優 (local optima)

解決方法：多次隨機初始化

For i = 1 to $\underbrace{100}_{50-1000}$ {

Randomly initialize K-means.

Run K-means. Get $c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_k$

* Pick k distinct random integers

$i_1, \dots, i_k$ from $\{1, \dots, m\}$

Set $\mu_1 = x^{(i_1)}, \dots, \mu_k = x^{(i_k)}$

Compute cost function (distortion)

$J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_k)$

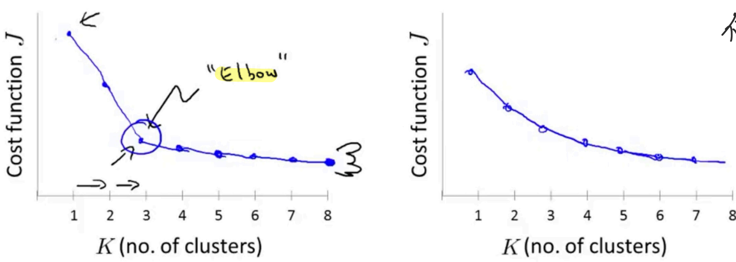
}

Pick clustering that gave lowest cost $J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_k)$

$k = 2-10$

- Choosing the number of clusters
  - Elbow method

大部份的情況很難用Elbow method做出判斷



  - Sometimes, you're running K-means to get clusters to use for some later / downstream purpose. Evaluate K-means based on a metric for how well it performs for that later purpose.

    → 由人工決定