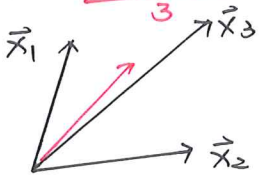
Linear algebra : k-means clustering

$\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_t \in \mathbb{R}^d$

The center (mean) is $\frac{1}{t}(\vec{x}_1 + \vec{x}_2 + \cdots + \vec{x}_t)$

重心 $\frac{\vec{x}_1+\vec{x}_2+\vec{x}_3}{3}$

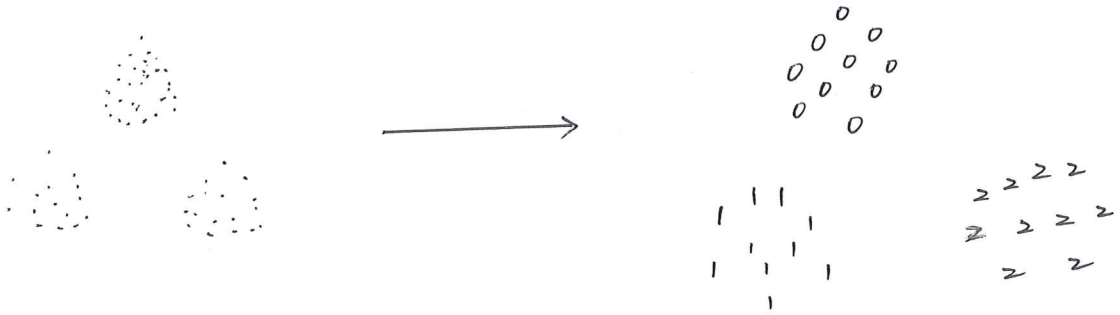


The distance between $\vec{x}$ and $\vec{y}$

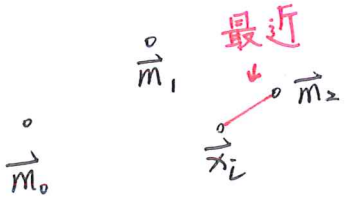
is $|\vec{x} - \vec{y}| = \sqrt{(\vec{x} - \vec{y}) \cdot (\vec{x} - \vec{y})}$

$$= \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_d - y_d)^2}$$

clustering (unsupervised)



一群資料 not labeled → (k) → k centers $\vec{m}_0, \cdots, \vec{m}_{k-1}$ → labeled data by nearest center



→ $\vec{x}_i$ is labeled by 2.

$\vec{m}_3$

k-means clustering algorithm

Input : unlabeled data $\vec{x}_1, \cdots, \vec{x}_N$, and integer k.
(number of groups)

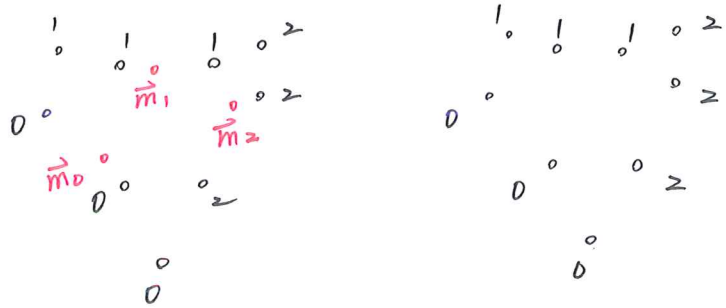
Output : label $\vec{y} = (y_1, \cdots, y_N)$, each $y_i \in \{0, 1, \cdots, k-1\}$

① Assign random $\vec{y}$ ($y_i \in \{0, \cdots, k-1\}$)

② For each group j $\{\vec{x}_i : y_i = j\}$,
compute the center $\vec{m}_j$.

③ For each $\vec{x}_i$, if $\vec{m}_j$ is the nearest center
assign $y_i = j$.

④ Repeat ②, ③ until $\vec{y}$ does not change.



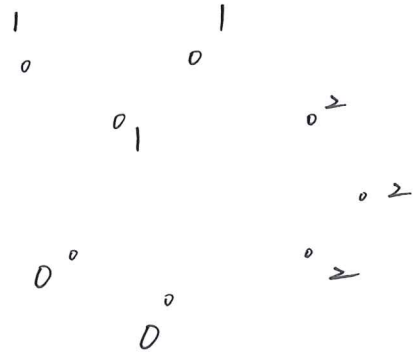
no more change

Thm.

The k-means clustering algorithm will always stop.

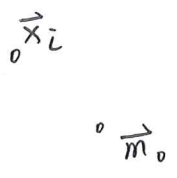
pf.

$$Err(\vec{m}_0, \cdots, \vec{m}_{k-1}) = \sum |\vec{x}_i - \vec{m}_{y_i}|^2$$



Claim: Err will strictly decrease.

e.g. If y_i is changed from 0 to 1,



$$\Rightarrow |\vec{x}_i - \vec{m}_1| < |\vec{x}_i - \vec{m}_0|$$

$$\Rightarrow \text{new } Err = Err - |\vec{x}_i - \vec{m}_0|^2 + |\vec{x}_i - \vec{m}_1|^2 < Err$$

Claim: There are only k^N different labels.

$$\vec{y} = (y_1, \cdots, y_N)$$

$0 \sim k-1$ $\qquad k^N$

The algorithm will reach min Err and stop.