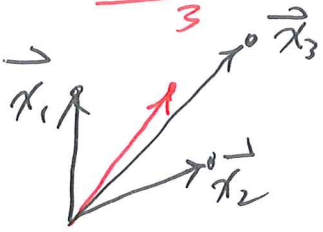
Linear algebra : k-mean clustering.

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

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

mean

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

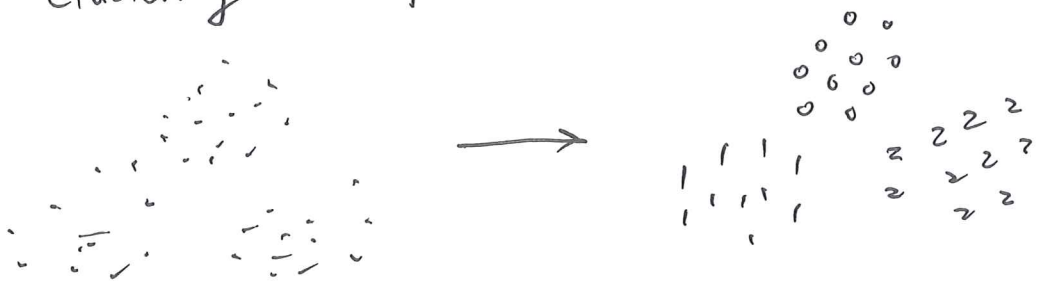


The distance between $\vec{x}$ and $\vec{y}$

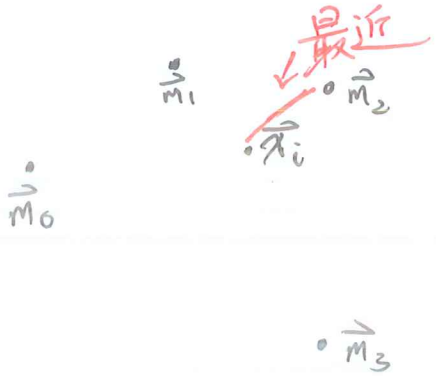
is $|\vec{x}-\vec{y}| = \sqrt{(\vec{x}-\vec{y})\cdot(\vec{x}-\vec{y})}$

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

clustering (unsupervised)



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



$\vec{x}_i$ is labeled by 2.

k-mean clustering algorithm

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

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

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

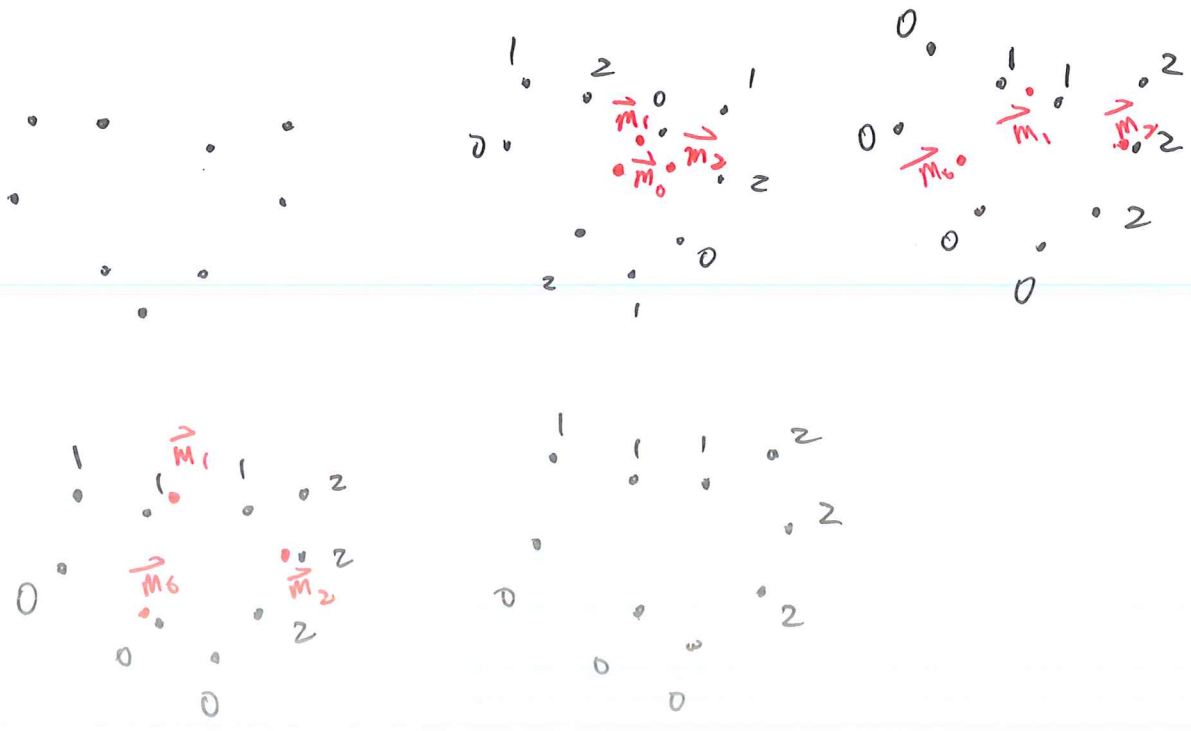
② For each group j $\{\vec{x}_i : y_i = j\}$,

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.



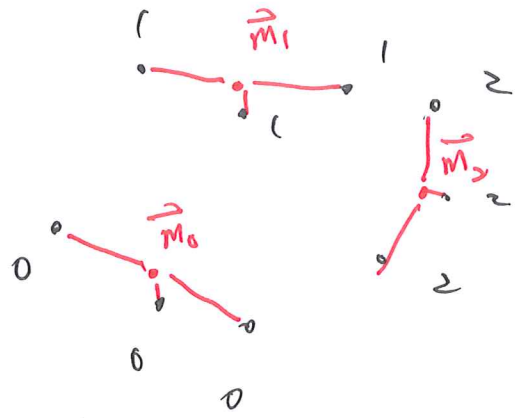
no more change.

Thm.

The k-mean clustering algorithm will always stop.

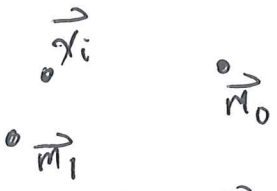
pf.

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



Claim: Err will strictly decrease.

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



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

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

Claim: There are only k^N ~~partiti~~ different labels.

$$\vec{y} = (y_1, \cdots, y_N) \quad k^N$$

$0 \sim k-1$

The algorithm will reach min Err and stop.