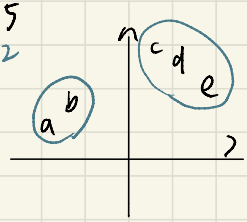
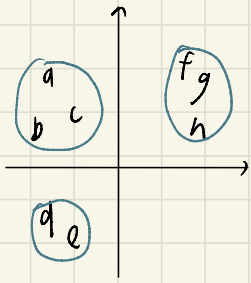
Clustering

把data分群

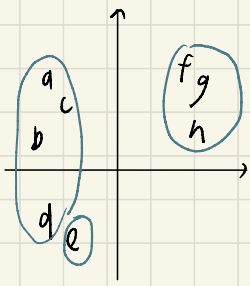
$n=5$

$p=2$





v.s



哪個好？

Goal of Clustering

① Group similar observations

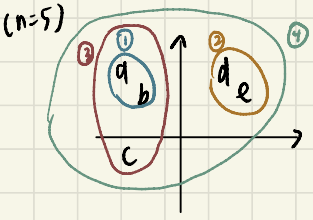
② Ideally, each group is very different.

組內差異小，組間差異大

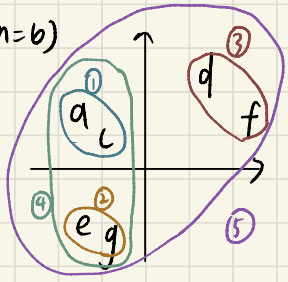
How to generalize to $p>2$?

1. Hierarchical clustering

$(n=5)$



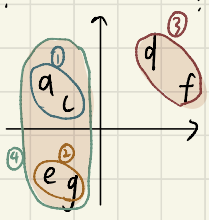
$(n=6)$



想分2組 不要做第5步

3：

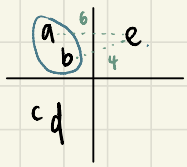
4：



- Algorithm: Bottom up

repeat ① Calculate pairwise distance
② Fuse the two nearest points
③ repeat ①, ② until every observation is in one group.

- How to calculate distance between groups?

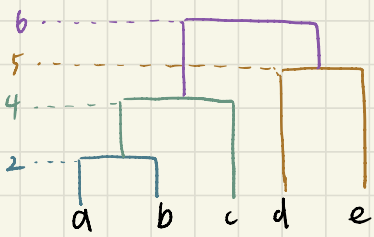


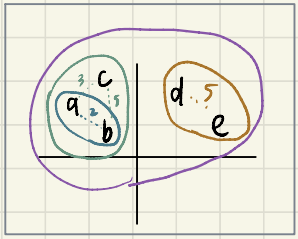
① Average of all distances
dist between (a_b), e is 5

② maximal distance
dist between (a_b), e is 6

- Dendrogram

y = distance between points fused





observation

2. K-means clustering

- Goal = find clusterings $C_1, C_2, \cdots C_K$ (K clusters)

$C_1, C_2, \cdots C_K$ is a partition of $\{1, 2, \cdots, n\}$ (eg. $C_1 = \{1, 3, 5\}$)

- $C_K \cap C_{K'} = \emptyset$
- $C_1 \cup C_2 \cup \cdots C_K = \{1, 2, \cdots n\}$

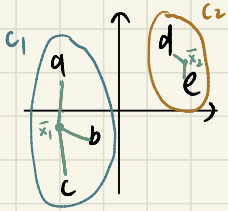
$\Rightarrow$ every observation is assigned and only assigned to one group

- k-means clustering find optimal $C_1, C_2, \cdots C_K$

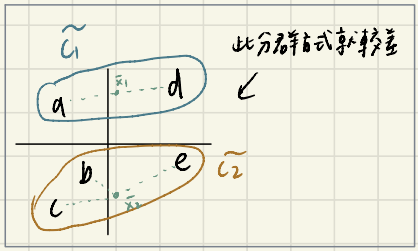
$$\min \sum_{k=1}^{K} \sum_{x_i \in C_k} \| x_i - \bar{x}_k \|^2$$

, where $\bar{x}_k = \frac{1}{n_k} \sum_{x_i \in C_k} x_i$ (centroids)

n_k = #observation in C_k



Find clustering s.t within group variation is the smallest.



Solution space = all possible partition $C_1, C_2, \cdots C_k$

Quiz: How many possibilities?

① n^K

② K^n ✓ (∵ every observation has K choices) n個東K個選擇

③ nK

NP-Hard.

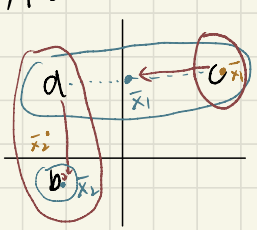
Hence, This is a combinatorial problem (NP Hard)

> **Remark**
> optimization is general hard outside calculus / linear programs

- K-means Algorithm
  1. Randomly assign points to cluster
  2. Calculate centroid $\bar{x}_1, \bar{x}_2, \cdots \bar{x}_K$
  3. Re-assign points to nearest centroids.

) repeat until convergence.

$n=3, K=2$



① random assign, calculate $\bar{x}_i$

② reassign

③ calculate $\bar{x}_i$

> **Remark:**
> ① it always converges
> ② But does not guarantee to the optimum
>
> So, try many time with different initialization