Pelleg, D., & Moore, A. W. (2000).

X-means:

Extending K-means with

Efficient Estimation of the Number of Clusters.

In *ICML* (Vol. 1). Retrieved from http://cs.uef.fi/~zhao/Courses/Clustering2012/Xmeans.pdf

資料探勘之群集分析

X-means

改善K-means：X-means演算法

K-means的缺點

- 每一輪迭代的計算耗時
- 需要指定分群數量K，不利於探索性分析
- 資料存在離群值時，容易陷入局部最佳解

X-means的改進

- 使用kd-tree加速原本K-means的迭代效率
- 使用者只要指定K的最小值與最大值範圍，X-means會以BIC score選擇最佳K值
- 每一輪迭代只進行2-means，避免陷入局部最佳解

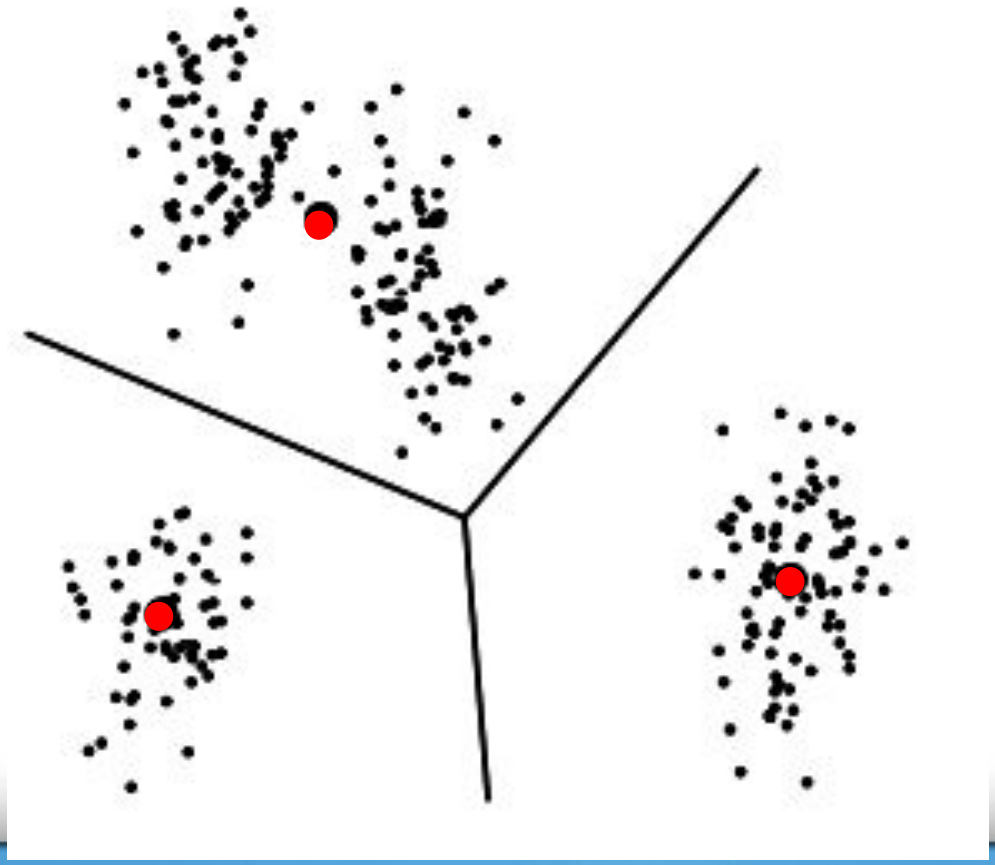
X-means的計算步驟

輸入：資料集D、指定群集數量最小值K_{min}、最大值K_{max}

1. 執行K_{min}-means
2. 在每個群集中執行2-means
   a. 分群前，計算k=1的BIC score
   b. 分群後，計算k=2的BIC score
   c. 如果BIC(k=2) 大於BIC(k=1)，則進行分群，K+1
   d. 反之則不分群
3. 如果K ＜ K_{max}，則繼續進行步驟2，否則返回結果

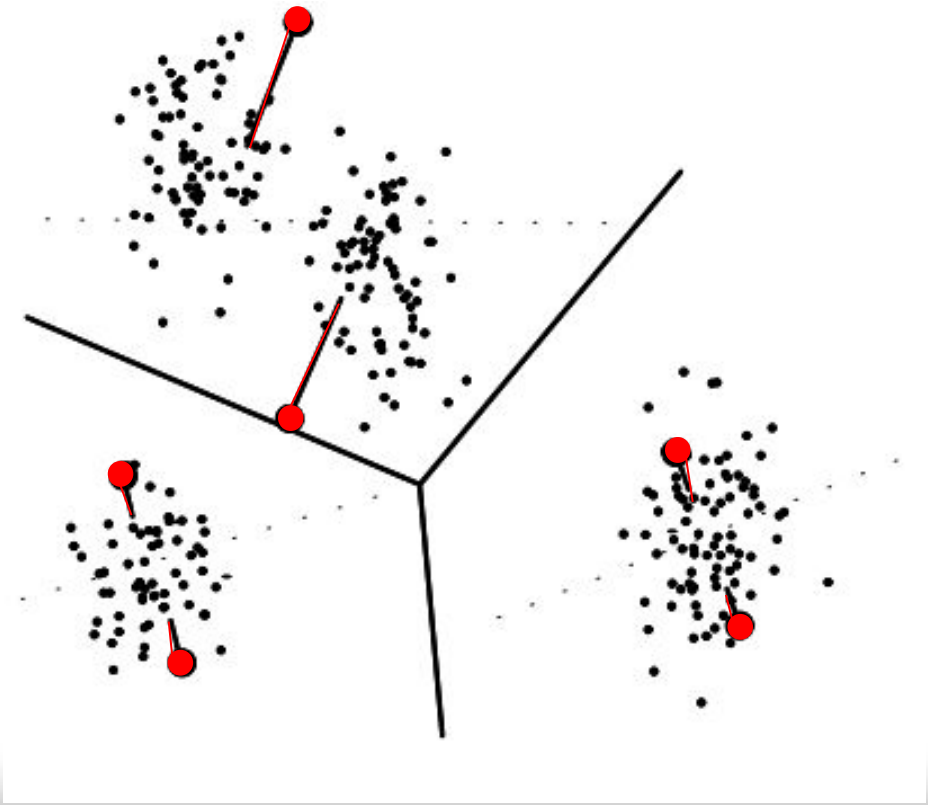
X-means運作示意圖 (K_{min}=3)

1. 首先將資料集D分成K_{min}=3群



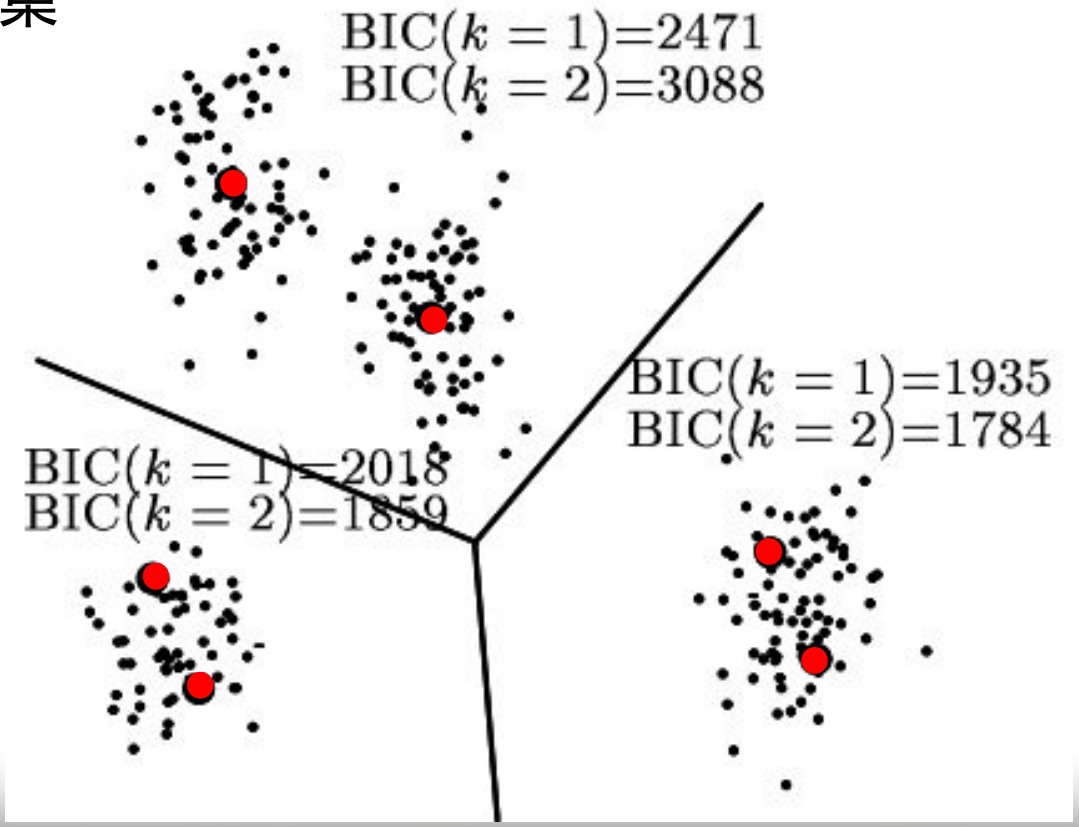
X-means運作示意圖 (K_{min}=3)

2. 為每一群集進行2-means分群，群集中心會逐漸往黑線方向移動



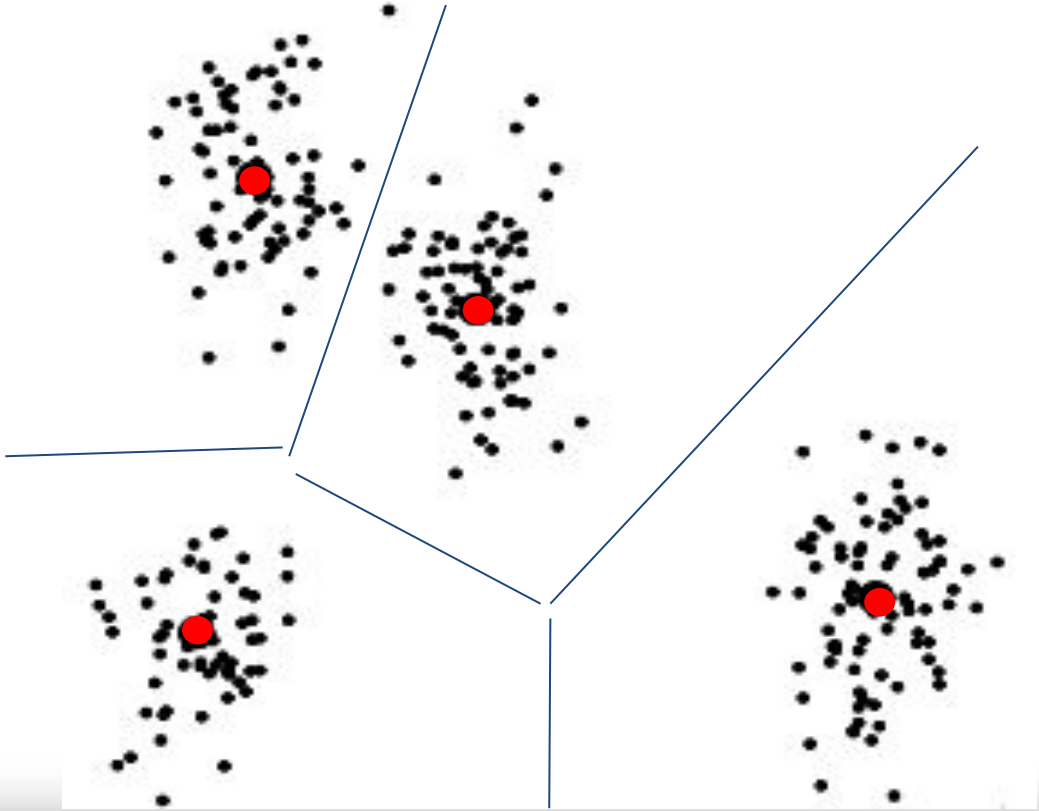
X-means運作示意圖 (K_{min}=3)

3. 當2-means完成後，計算各群在分群前後的BIC，決定是否切割群集



X-means運作示意圖 (K_{min}=3)

4. 判斷是否切割後，最後留下的群集中心



貝氏資訊準則
BIC Score

BIC是一種後驗機率，以最大相似估計法(maximun likelihood estimate)來計算不同分群結果的分數。

$$BIC(M_j) = \hat{l}_j(D) - \frac{p_j}{2} \cdot \log R$$

- M_j表示模型(分群的結果)
- $\hat{l}_j(D)$ 為likelihood
- P_j為模型的複雜度(自由參數個數)