第七章
分群與應用

內　容

7.1 前　言

分群 (clustering) 是將一組資料依據某種距離的量度將該組資料分割成若干群。

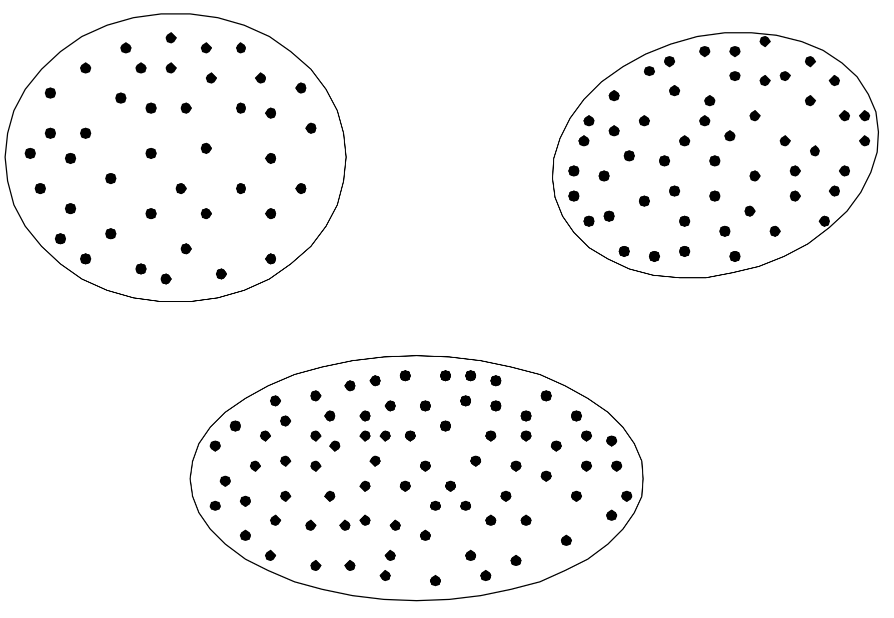

圖 7.1.1　分群示意圖

7.2 K-means 分群法

K-means 的 K 指的是分群數 。

範例 7.2.1：

給 10 筆資料，點 v_1 和點 v_8 為起始的兩個群心。

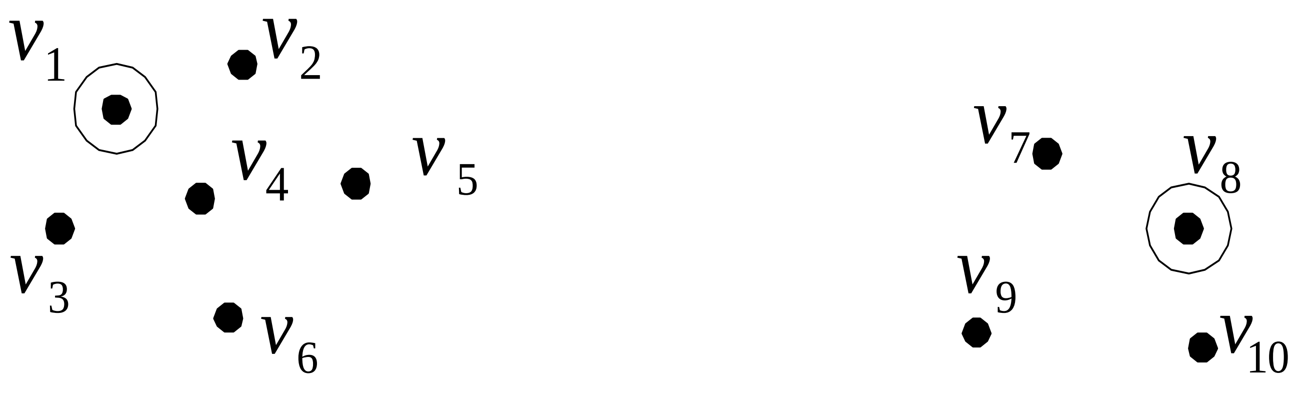

圖7.2.1 起始的兩個群心選定

圖7.2.2 各點的歸類

在下一次的疊代中，如果群心 $\overline{v_1}$ 和 $\overline{v_2}$ 不會再改變，那表示已經完成分群工作。

7.3 K-D 樹的分群法

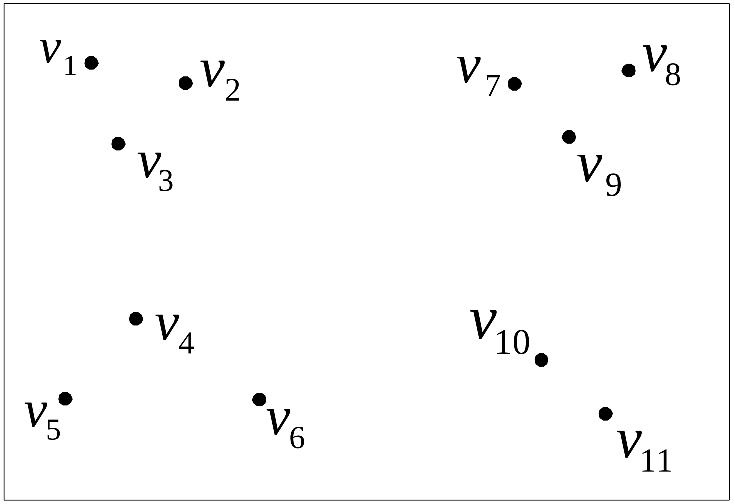

圖7.3.1　十一筆資料的分佈圖

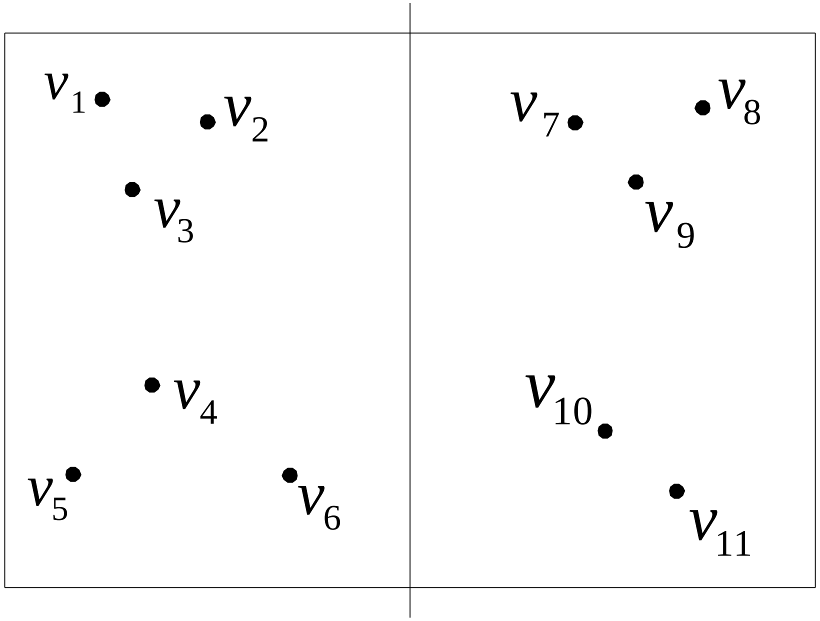

圖7.3.2 第一次分割

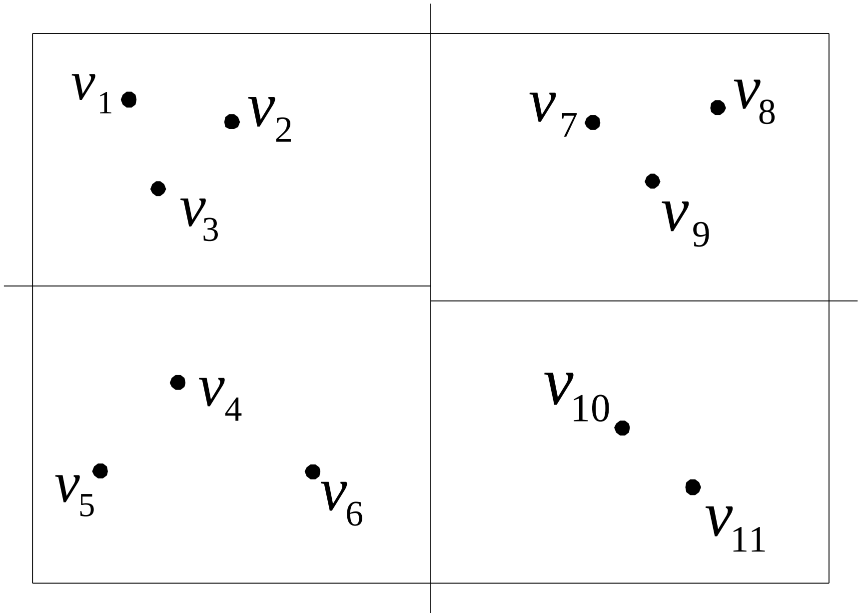

圖7.3.3 最後分割的結果

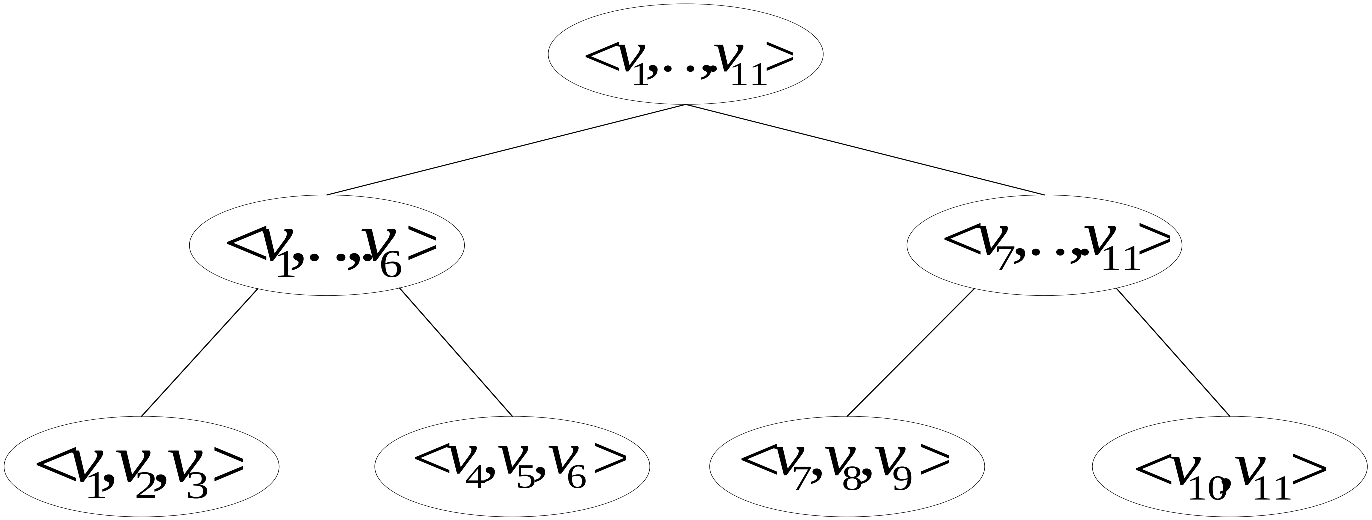

圖7.3.4　K-D 樹

實例：

有 32 個點，12 個起始群心被標示為 $C_1, C_2, \ldots$ 和 C_{12}。

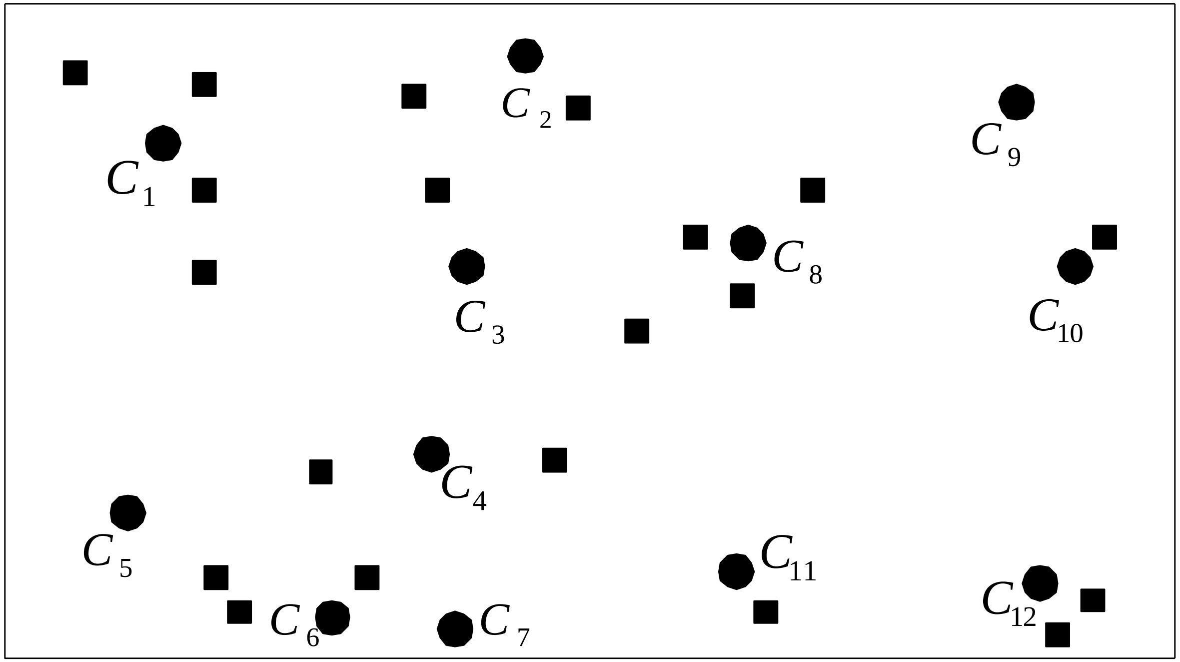

圖7.3.5　起始時的狀態

根據 K-D 樹的均等分割原理，經過第一次分割後，我們得到圖 7.3.6 的分割圖。左邊內的點 Z_1^* 代表左區的質心，這質心可幫助我們加快每一筆資料的群心歸屬。

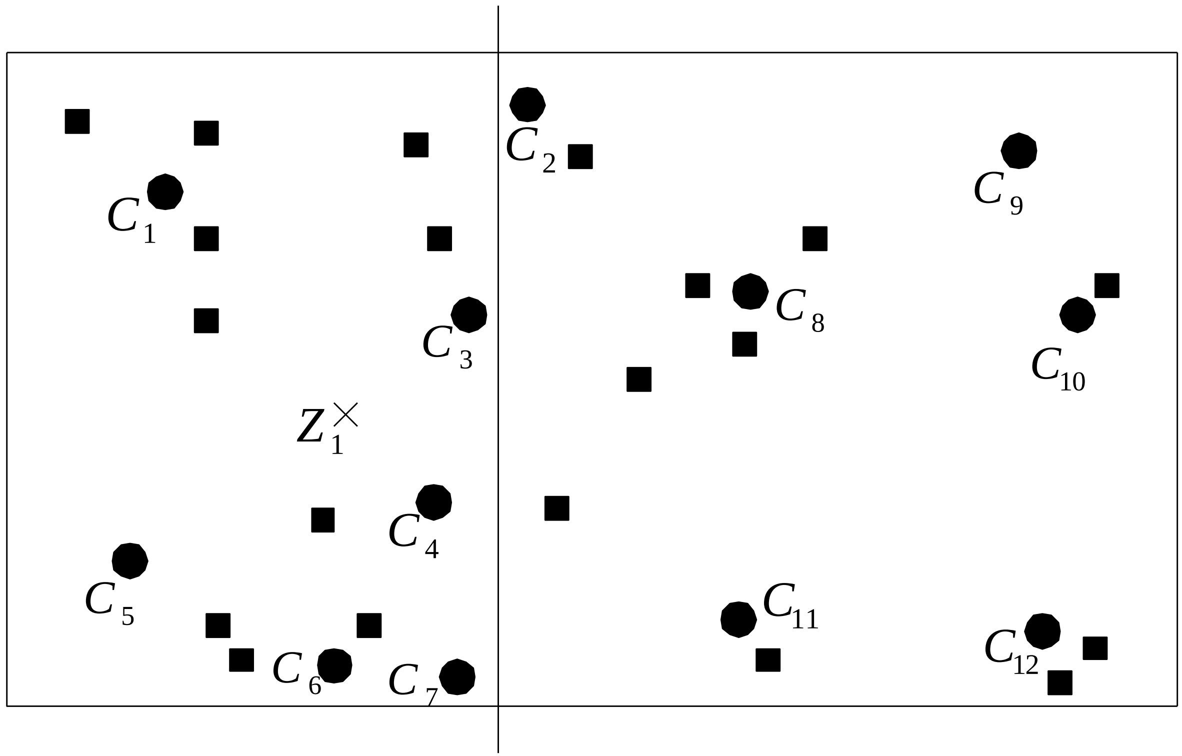

圖7.3.6　第一次分割圖

只需將點A的座標值代入中分線後得到負值就可知道A點距離Z^*較近，這時對點而言，群心Z^i是不必被納入考慮的。

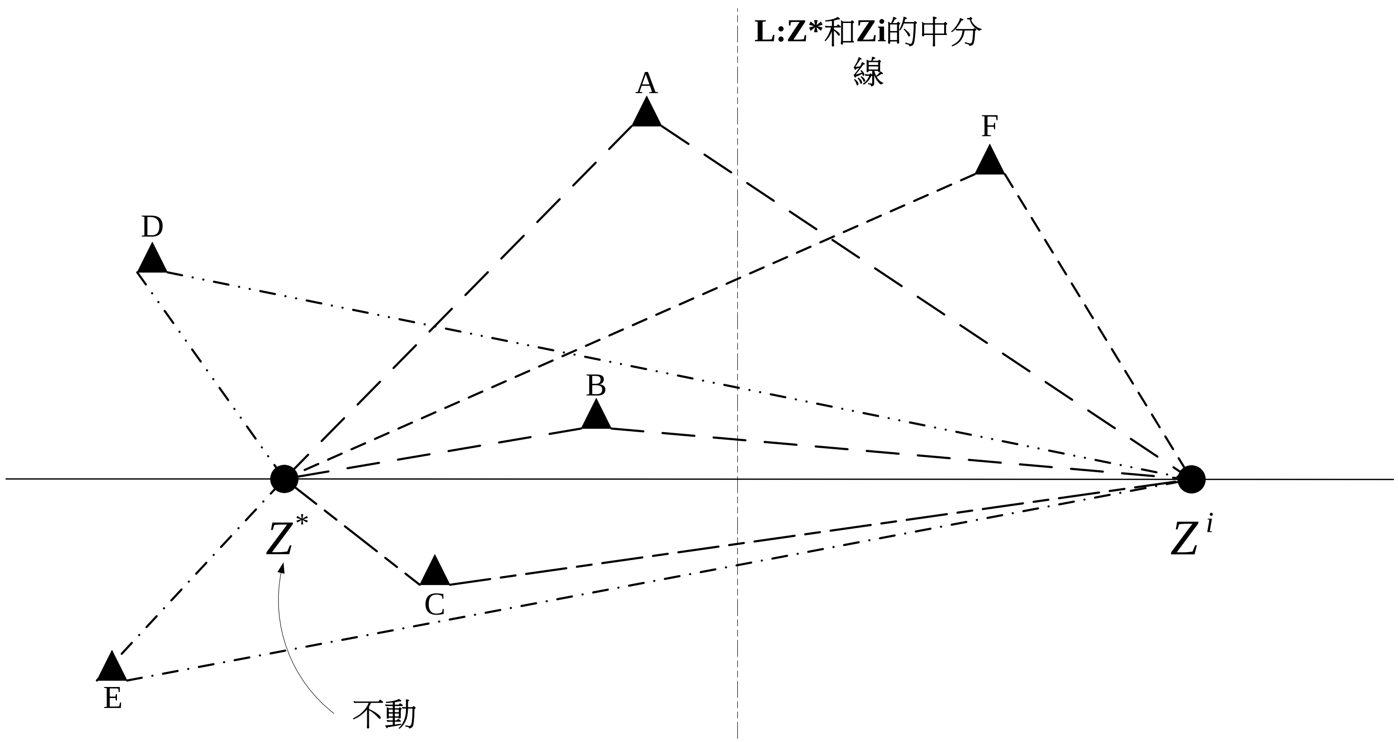

圖7.3.7　加快群心歸屬示意圖

7.4 模糊分群 (FCM) 法

資料點集為 $X=\{x_1,x_2,\ldots,x_n\}$，令這 C 群的群心集為 $V=\{v_1, v_2,\ldots, v_c\}$。令資料點 x_j 對群心 v_i 的隸屬函數值為 u_{ij}，則隸屬矩陣 U 可表示為

$$U=\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ u_{c1} & u_{c2} & \cdots & u_{cn} \end{bmatrix}$$

Lagrange Function (拉格蘭吉函數)

設需求解最大化或最小化的函數為$f(x_1, x_2, \cdots, x_n)$，其限制條件為 $g_j(x_1, x_2, \cdots, x_n) = b_j, j = 1,2, \cdots, m.$

可利用Lagrange function (定義如下)解此最佳化問題。
$L(x_1, x_2, \cdots, x_n, \lambda_1, \lambda_2, \cdots, \lambda_m) = f(x_1, x_2, \cdots, x_n) + \sum_{j=1}^{m} \lambda_j$ [$g_j(x_1, x_2, \cdots, x_n)$-$b_j$], 其中稱$\lambda_j$為Lagrange multiplier.

則求解$f(x_1, x_2, \cdots, x_n)$的最大值或最小值，可由下述方程式的聯立解。
$\frac{\partial L}{\partial x_1}, \frac{\partial L}{\partial x_2}, \cdots, \frac{\partial L}{\partial x_n}, \frac{\partial L}{\partial \lambda_1}, \frac{\partial L}{\partial \lambda_2}, \cdots, \frac{\partial L}{\partial \lambda_m}$ 。

群心集 V 和資料點集 X 的誤差為：

$$E(U,V:X)=\sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^{m}\left\|x_{j}-v_{i}\right\|^{2} \qquad (7.4.1)$$

$$\sum_{i=1}^{c}u_{ij}=1$$

$$L(U,\lambda)=\sum_{j=1}^{n}\sum_{i=1}^{c}(u_{ij})^m\left\|x_j-v_i\right\|^2-\sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{c}u_{ij}-1\right)$$

$$\frac{\partial L(U,\lambda)}{\partial \lambda_j}=0\Leftrightarrow\sum_{i=1}^{c}u_{ij}-1=0 \qquad (7.4.2)$$

$$\frac{\partial L(U,\lambda)}{\partial u_{ij}}=0\Leftrightarrow\left[m(u_{ij})^{m-1}\left\|x_j-v_i\right\|^2-\lambda_j\right]=0 \qquad (7.4.3)$$

由式 (7.4.3) 可解得

$$u_{ij}=\left(\frac{\lambda_j}{m\left\|x_j-v_i\right\|^2}\right)^{\frac{1}{m-1}} \qquad (7.4.4)$$

由式 (7.4.2) 和式 (7.4.4) 可得到

$$\sum_{i=1}^{c} u_{ij} = \sum_{i=1}^{c} \left(\frac{\lambda_j}{m \left\| x_j - v_i \right\|^2} \right)^{\frac{1}{m-1}} \tag{7.4.5}$$

$$= 1$$

從式 (7.4.5) 可得

$$\left(\frac{\lambda_j}{m} \right)^{\frac{1}{m-1}} = 1 \Bigg/ \sum_{i=1}^{c} \left(\frac{1}{m \left\| x_j - v_i \right\|^2} \right)^{\frac{1}{m-1}} \tag{7.4.6}$$

將式 (7.4.6) 代入式 (7.4.4)，得

$$u_{ij} = 1 \Bigg/ \sum_{k=1}^{c} \left(\frac{\left\| x_j - v_i \right\|}{\left\| x_j - v_k \right\|} \right)^{\frac{2}{m-1}} \qquad (7.4.7)$$

群心 v_i 可調整為

$$v_i = \frac{\sum_{j=1}^{n} \left(u_{ij}\right)^m x_j}{\sum_{j=1}^{n} \left(u_{ij}\right)^m}, 1 \le i \le c \qquad (7.4.8)$$

FCM法共分下列五個步驟：

步驟一：選定群數 C、次方 m、誤差容忍度 ε 和起始隸屬矩陣 U_0。

步驟二：根據資料點集和 U_0 算出起始的群心集。

步驟三：重新計算U_{ij}，$1 \le i \le c$ 和 $1 \le j \le n$。修正各個群心值。

步驟四：計算出誤差 $E = \sum_{i=1}^{c} \left\| v_i^{前} - v_i^{後} \right\|$ ，這裏 $v_i^{前}$ 和 $v_i^{後}$ 代表群心 v_i 連續兩個疊代回合的值。

步驟五：若 E 很小則停止；否則回到步驟三。

7.5 作　業

- 作業一：寫一 C 程式以完成 K-means 分群法的實作。

- 作業一：寫一 C 程式以完成 K-D 樹分群法的實作。

- 作業一：寫一 C 程式以完成 FCM 法的實作。