**Michine Learning-Homework1**

資訊碩一 11177016 董承宥

1. 目的
2. 隨機尋找K個初始中心點。
3. 確認最佳K值為何。
4. 解決方法
5. 利用sample函數在資料中挑選初始中心點，確保初始中心點一定位於資料當中。
6. 利用elbow找出最佳K值為何。
7. 方法實驗
8. 利用sample函數在資料中挑選初始中心點，如圖 1，確保初始中心點一定位於資料當中，如圖 2。

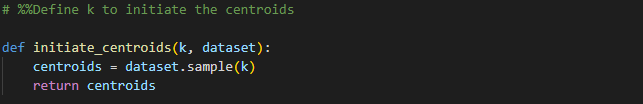


圖 1. 利用sample函數挑選初始中心點。

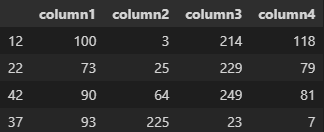
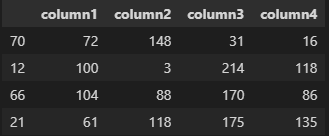


圖 2. 左邊和右邊分別為2次隨機挑選初始中心點，左圖第一列70 12 66 21表示在原始資料中的編號，右圖第一列 12 22 42 37表示在原始資料中的編號。

1. 利用elbow找出最佳K值為何，如圖 3，並且視覺化，如圖 4。



圖 3. 利用err\_total陣列儲存kmeans回傳回來的err總和，並輸出X軸為Number of clusters Y軸為Total error的折線圖，如圖 4。

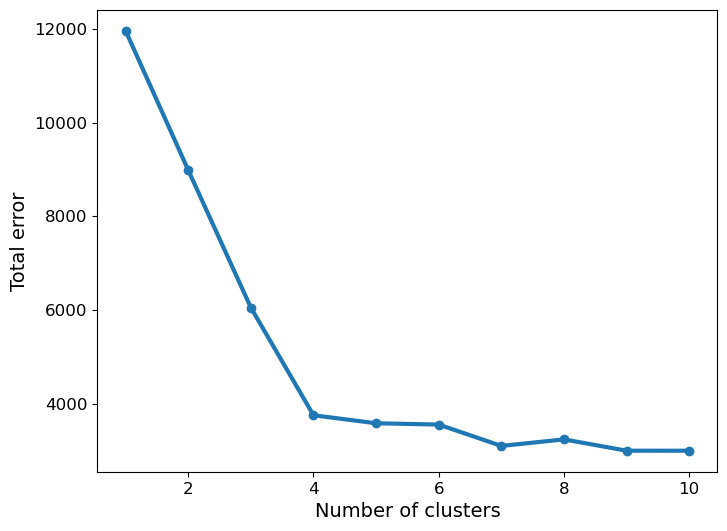


圖 4. 顯示K值在等於4的時候Total error值下降幅度沒有那麼明顯。

1. Source code

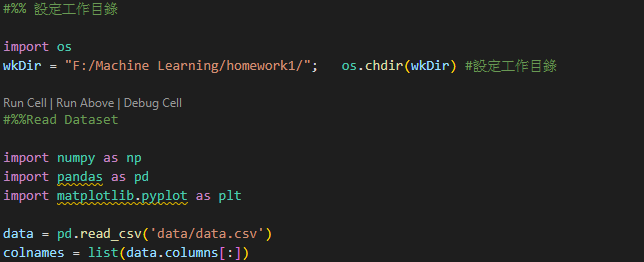


圖 5. 設定工作目錄，並且import所需模組，將所需資料集以變數data儲存。

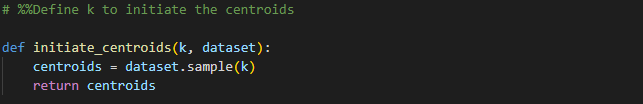


圖 6. 利用sample函數挑選初始中心點。

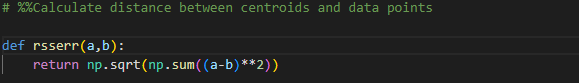


圖 7. 計算每個中心點和數據之間的距離。

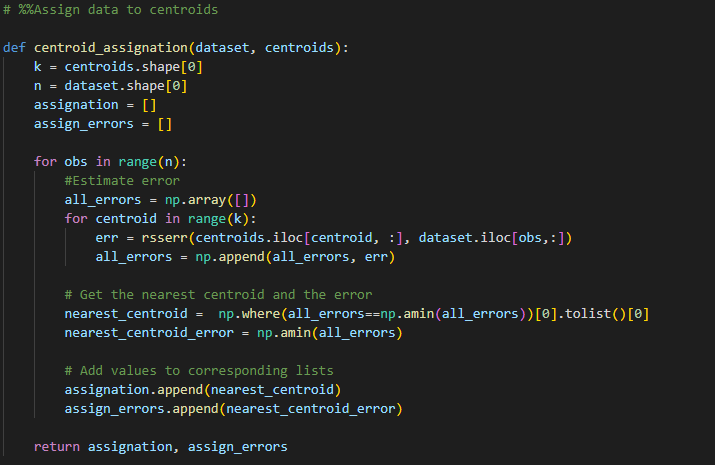


圖 8. 估計每個中心和數據之間的誤差，並把誤差最小的數據分配給該中心點。

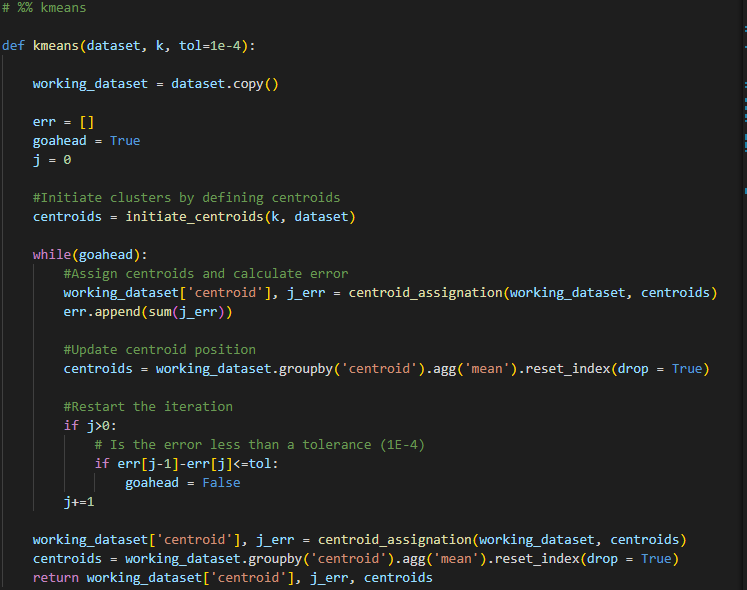


圖 9. 建立一個while迴圈來重複估計每個中心和數據之間的誤差，並把誤差最小的數據分配給該中心點，再透過平均該群數據與其中心點的距離來重新更新該群中心點的位置，同時設置一個if迴圈，當前該群數據與其中心點的誤差總和，若與更新前中心點的誤差總和小於0.0004，則跳出while迴圈不再更新中心點位置。

1. Clustering result(聚類結果)

column1~4表示該點的4D座標，centroid表示該點屬於哪群，error表示該點與其所屬中心點的error值。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| column1 | column2 | column3 | column4 | centroid | error |
| 21 | 49 | 31 | 251 | 3 | 32.27833 |
| 21 | 53 | 25 | 244 | 3 | 28.0872 |
| 21 | 105 | 41 | 267 | 2 | 47.21334 |
| 81 | 66 | 239 | 97 | 0 | 112.4715 |
| 69 | 3 | 132 | 73 | 0 | 141.4546 |
| 228 | 29 | 193 | 207 | 0 | 88.81916 |
| 22 | 132 | 63 | 198 | 2 | 39.18614 |
| 15 | 158 | 29 | 167 | 2 | 70.2742 |
| 25 | 127 | 55 | 232 | 2 | 15.50686 |
| 16 | 118 | 40 | 277 | 2 | 48.91653 |
| 18 | 159 | 61 | 275 | 2 | 51.36243 |
| 24 | 144 | 62 | 205 | 2 | 33.54385 |
| 100 | 3 | 214 | 118 | 0 | 89.46606 |
| 75 | 50 | 221 | 123 | 0 | 99.27939 |
| 105 | 108 | 211 | 81 | 0 | 110.1552 |
| 95 | 82 | 248 | 90 | 0 | 114.3744 |
| 223 | 52 | 87 | 216 | 0 | 133.8893 |
| 198 | 28 | 148 | 188 | 0 | 68.08498 |
| 266 | 24 | 198 | 213 | 0 | 123.2967 |
| 211 | 55 | 214 | 183 | 0 | 64.09679 |
| 275 | 61 | 134 | 234 | 0 | 149.6919 |
| 61 | 118 | 175 | 135 | 0 | 125.567 |
| 73 | 25 | 229 | 79 | 0 | 124.9334 |
| 61 | 52 | 220 | 106 | 0 | 116.8682 |
| 75 | 102 | 134 | 128 | 0 | 118.3637 |
| 14 | 104 | 17 | 256 | 2 | 48.97039 |
| 12 | 79 | 44 | 211 | 3 | 20.94375 |
| 25 | 160 | 58 | 181 | 2 | 58.13627 |
| 286 | 29 | 177 | 231 | 0 | 148.4694 |
| 260 | 45 | 146 | 240 | 0 | 138.3673 |
| 257 | 16 | 126 | 192 | 0 | 123.4277 |
| 80 | 150 | 42 | 14 | 1 | 36.87428 |
| 90 | 187 | 41 | 15 | 1 | 6.149187 |
| 79 | 173 | 62 | 18 | 1 | 28.28449 |
| 70 | 192 | 44 | 13 | 1 | 17.39289 |
| 85 | 210 | 25 | 24 | 1 | 28.14094 |
| 108 | 179 | 33 | 11 | 1 | 25.18358 |
| 93 | 225 | 23 | 7 | 1 | 43.21125 |
| 110 | 137 | 16 | 17 | 1 | 59.42905 |
| 323 | 50 | 182 | 175 | 0 | 160.6314 |
| 65 | 37 | 170 | 67 | 0 | 130.4273 |
| 101 | 84 | 208 | 95 | 0 | 92.52212 |
| 90 | 64 | 249 | 81 | 0 | 119.2132 |
| 14 | 140 | 37 | 315 | 2 | 83.54318 |
| 20 | 99 | 24 | 229 | 3 | 37.74441 |
| 21 | 49 | 54 | 158 | 3 | 71.16102 |
| 15 | 154 | 24 | 179 | 2 | 59.53005 |
| 26 | 64 | 50 | 311 | 3 | 85.73879 |
| 14 | 80 | 54 | 183 | 3 | 46.99086 |
| 12 | 55 | 54 | 221 | 3 | 18.18353 |
| 80 | 167 | 18 | 16 | 1 | 28.04661 |
| 64 | 218 | 45 | 18 | 1 | 38.73903 |
| 105 | 159 | 38 | 25 | 1 | 35.0373 |
| 77 | 200 | 21 | 10 | 1 | 23.86865 |
| 120 | 196 | 45 | 8 | 1 | 37.98569 |
| 68 | 208 | 64 | 21 | 1 | 38.34205 |
| 285 | 50 | 154 | 203 | 0 | 136.127 |
| 268 | 48 | 200 | 177 | 0 | 108.0291 |
| 300 | 9 | 115 | 242 | 0 | 183.9816 |
| 306 | 26 | 212 | 201 | 0 | 154.659 |
| 81 | 87 | 273 | 113 | 0 | 130.7657 |
| 60 | 36 | 168 | 91 | 0 | 121.624 |
| 85 | 114 | 250 | 105 | 0 | 128.0686 |
| 90 | 160 | 90 | 30 | 1 | 60.29853 |
| 56 | 57 | 227 | 85 | 0 | 132.3302 |
| 93 | 8 | 169 | 82 | 0 | 107.659 |
| 104 | 88 | 170 | 86 | 0 | 96.73623 |
| 63 | 189 | 34 | 8 | 1 | 23.78261 |
| 82 | 218 | 28 | 14 | 1 | 33.41426 |
| 78 | 223 | 16 | 20 | 1 | 43.46737 |
| 72 | 148 | 31 | 16 | 1 | 40.93547 |
| 84 | 186 | 40 | 15 | 1 | 2.590849 |
| 247 | 20 | 143 | 185 | 0 | 104.3927 |
| 191 | 31 | 196 | 194 | 0 | 56.4477 |
| 273 | 22 | 193 | 231 | 0 | 139.0462 |

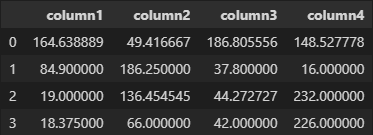


圖 10. 中心點座標

1. 觀察結果

就算用手肘法選出最佳群數，也會因為初始中心點的位置導致整體的error有所差距，如圖 11。

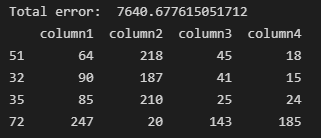
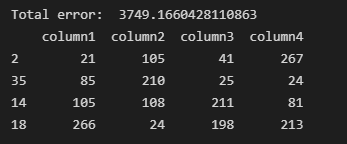


圖 11. 初始中心點座標和Total error，雖然左右都是分成4群，但是因為初始中心點的選擇，導致最終的Total error有接近4000的差距。