

The background is a gradient of red and purple with a starry texture. On the left side, there are several concentric circles and arcs, some with degree markings (40, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 260) and arrows, suggesting a circular or rotational theme.

資料探勘 DATA MINING

張家瑋 博士

助理教授

國立臺中科技大學資訊工程系

The background is a gradient from dark red at the top to dark blue at the bottom, speckled with small white dots. On the left side, there are several concentric circles and a large circular scale with degree markings from 140 to 260. Some circles have arrows indicating a clockwise direction.

分群練習 CLUSTERING EXERCISE

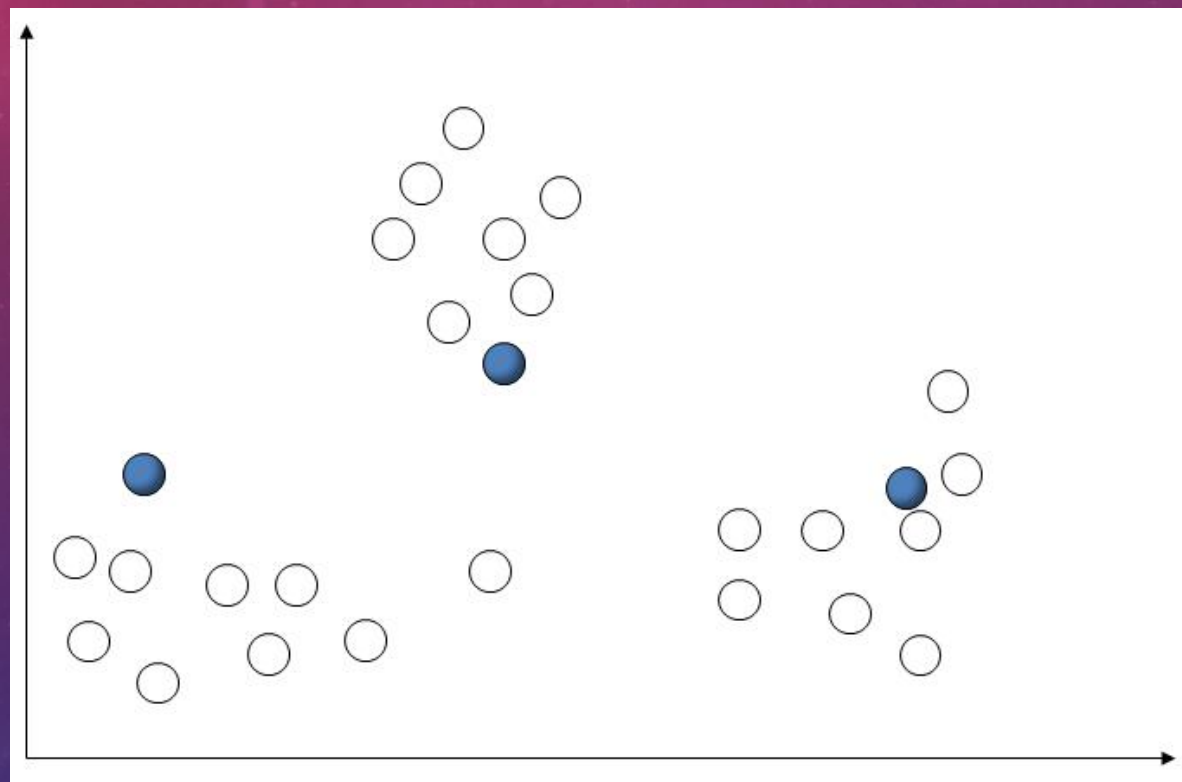
概念

- 把許多事物按照某種標準歸為數個類別，其中較為相近/類似的聚為一類，反之較不相近的則聚為不同類。
- 目的是企圖從一大堆雜亂無章的原始資料中，找出少數幾個較小的群體，使得群體內的分子在某些變項的測量值均很類似，而群體與群體間的分子在該測量值上差異較大。
- 同一組樣本會因不同目的、資料輸入方式、所選擇分群特徵或資料屬性，形成不同的分群結果

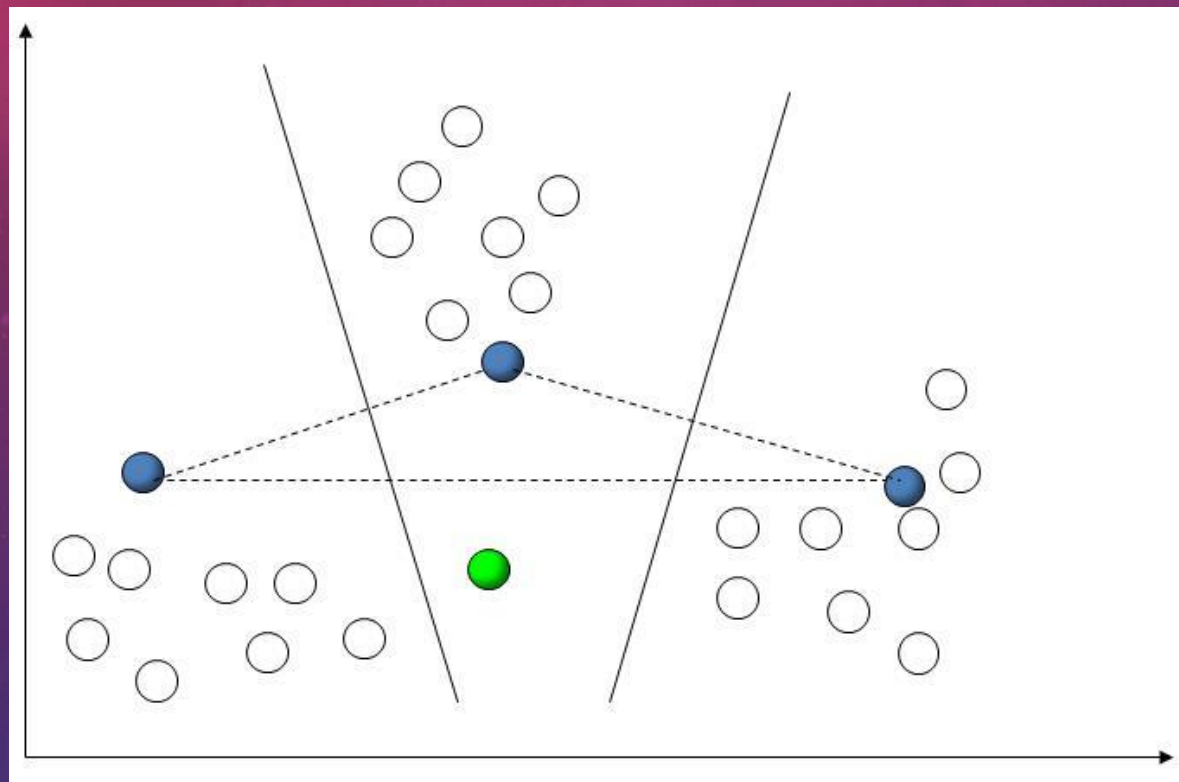
The background is a gradient from deep red at the top to dark blue at the bottom, speckled with white dots resembling stars. Overlaid on the left side are several concentric circles and arcs in white and light blue. Some of these circles have tick marks and numerical labels (40, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 260) along their perimeters, suggesting a circular scale or a data visualization related to angles or time. Arrows are also visible on some of the circular paths, indicating a direction of movement or flow.

K-MEANS

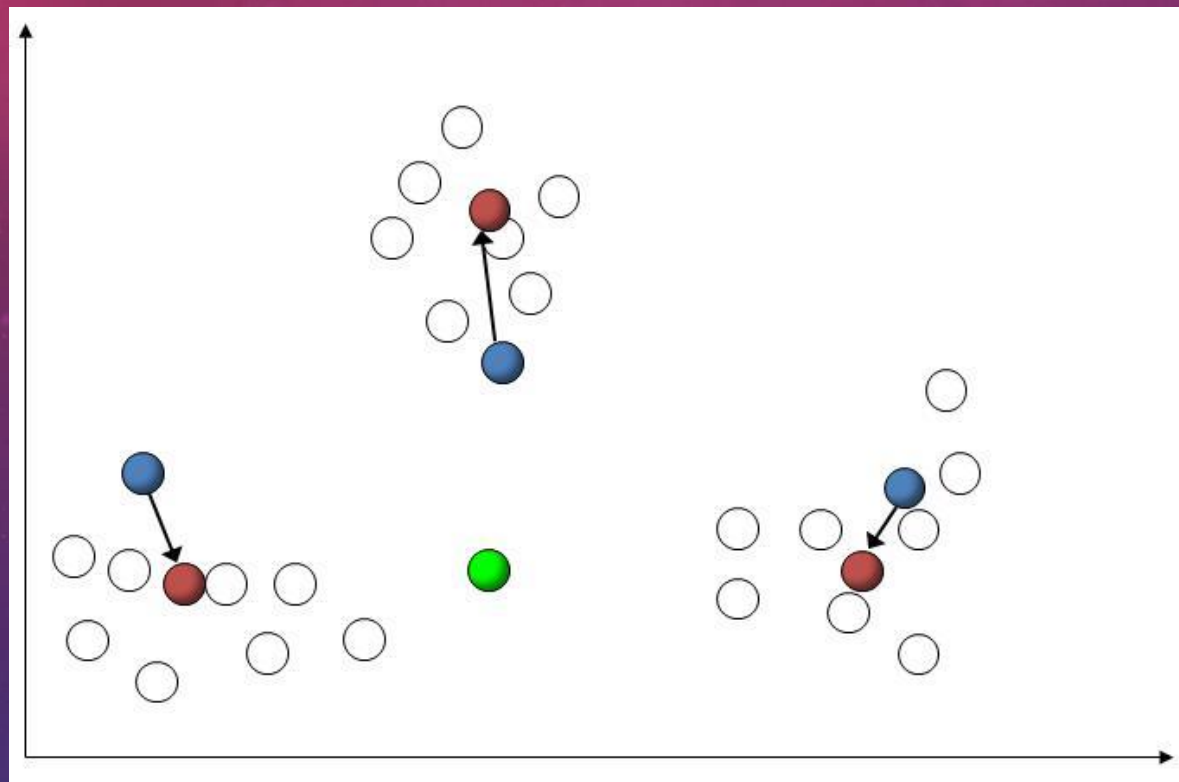
STEP 1. 隨機指派群集中心



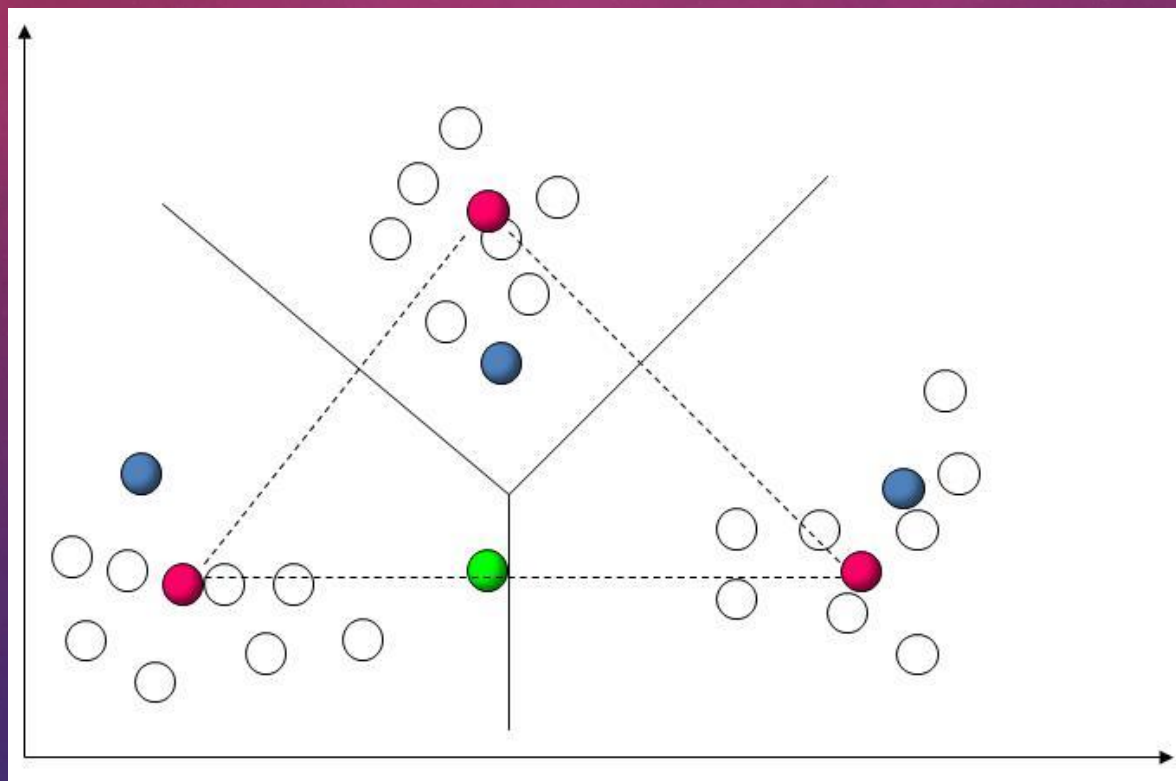
STEP 2. 產生初始群集



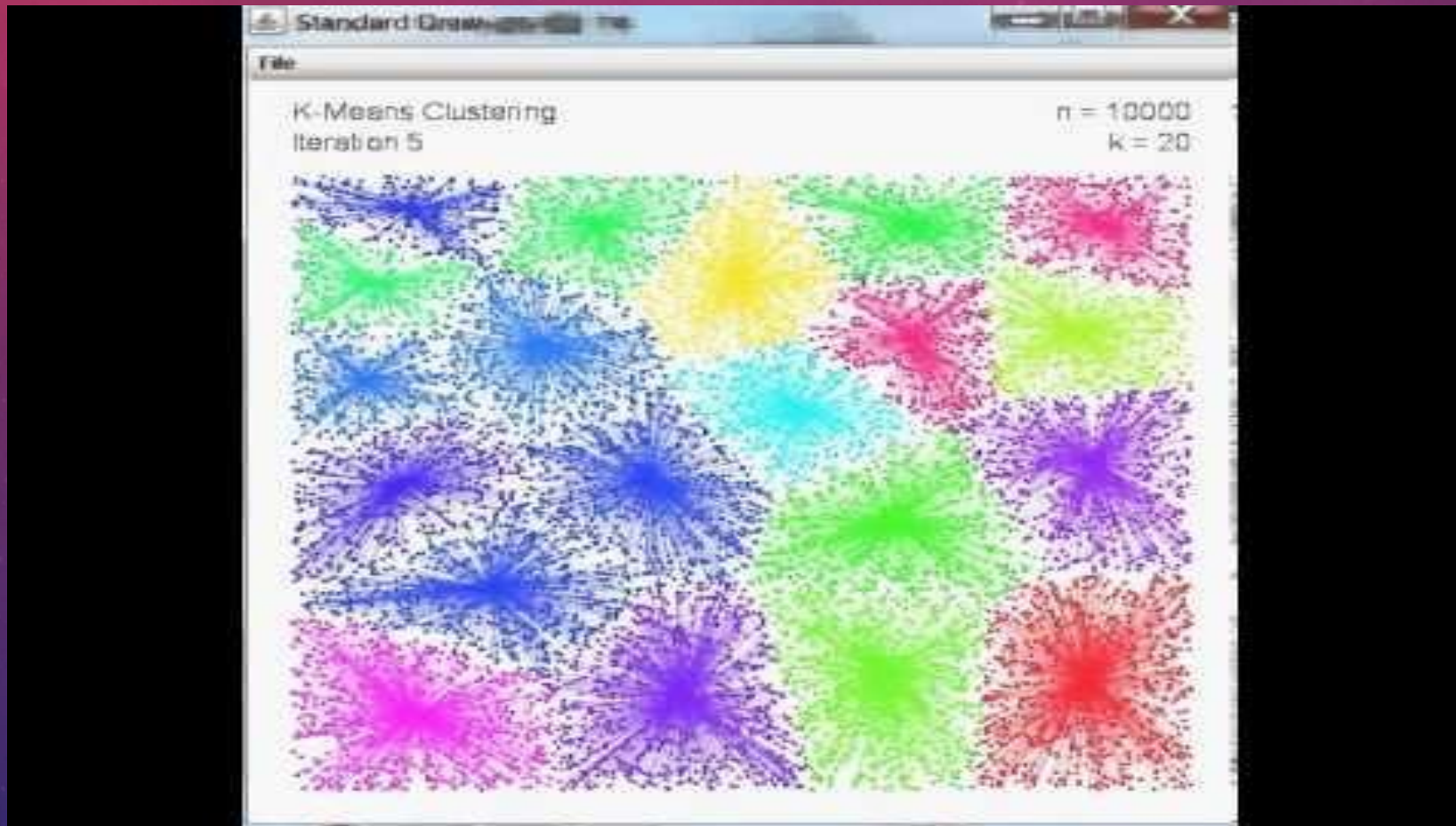
STEP 3. 產生新的質量中心



STEP 4. 變動群集邊界



EXAMPLE



<https://youtu.be/BVFG7fd1H30>

鳶尾花資料集

- 花瓣 (Petal) 的長
- 花瓣 (Petal) 的寬
- 花萼 (Sepal) 的長
- 花萼 (Sepal) 的寬

[5.1 3.5 1.4 0.2]

在設定某K的KMEANS

```
1  from sklearn import cluster, datasets
2
3  # 讀入鳶尾花資料
4  iris = datasets.load_iris()
5  iris_X = iris.data
6
7  # KMeans 演算法
8  kmeans_fit = cluster.KMeans(n_clusters = 3).fit(iris_X)
9
10 # 印出分群結果
11 cluster_labels = kmeans_fit.labels_
12 print("分群結果：")
13 print(cluster_labels)
14 print("---")
15
16 # 印出品種看看
17 iris_y = iris.target
18 print("真實品種：")
19 print(iris_y)
```


在設定某K的KMEANS

[5.1 3.5 1.4 0.2]

分群結果：

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 2 & 2 & 0 & 2 \\ 2 & 2 & 2 & 0 & 2 & 0 & 2 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 2 & 0 & 0 & 0 & 0 & 2 & 0 & 2 & 0 & 2 & 0 & 0 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 2 \end{bmatrix}$$

— — —

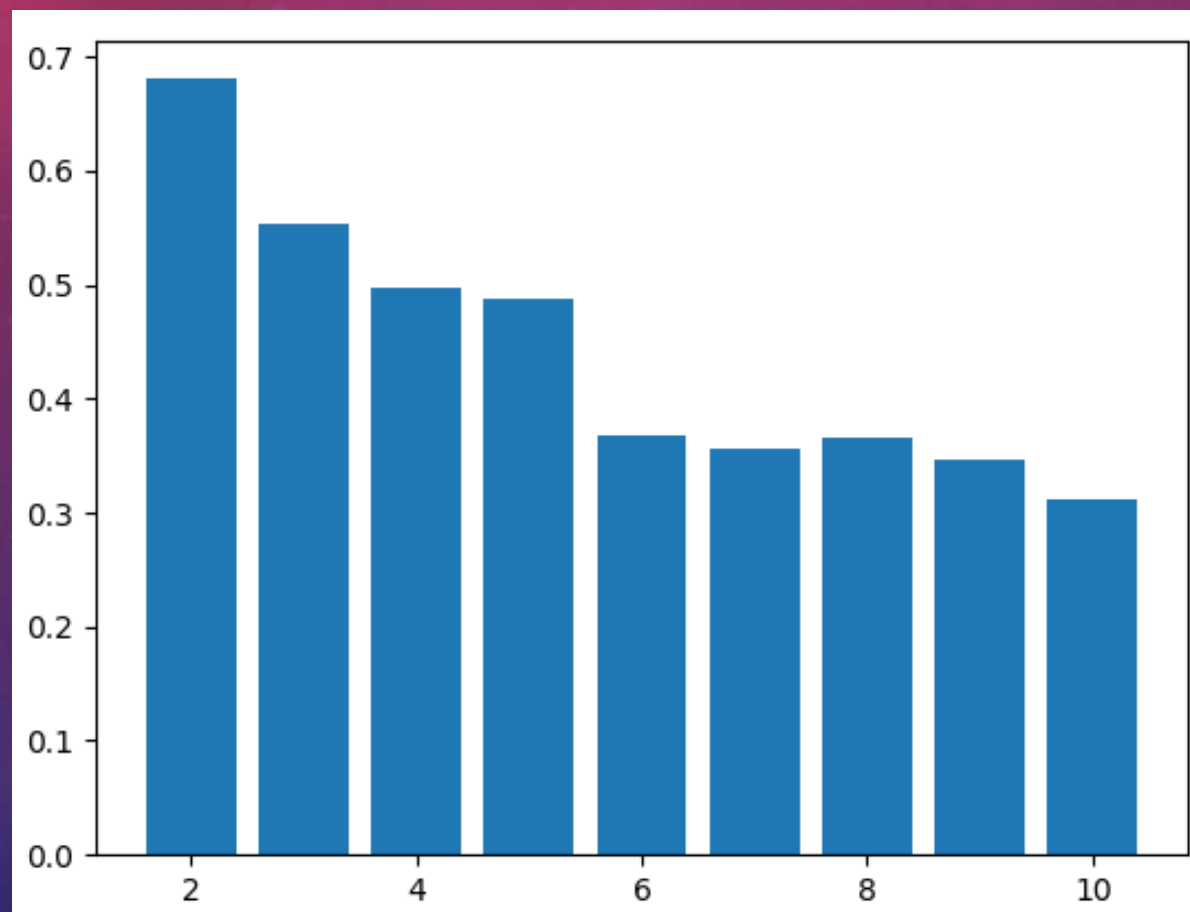
真實品種：

[illegible]

K從2到10的KMEANS效能

```
1 from sklearn import cluster, datasets, metrics
2 import matplotlib.pyplot as plt
3
4 # 讀入鸚尾花資料
5 iris = datasets.load_iris()
6 iris_X = iris.data
7
8 # 迴圈
9 silhouette_avgs = []
10 ks = range(2, 11)
11 ▼ for k in ks:
12     kmeans_fit = cluster.KMeans(n_clusters = k).fit(iris_X)
13     cluster_labels = kmeans_fit.labels_
14     silhouette_avg = metrics.silhouette_score(iris_X, cluster_labels)
15     silhouette_avgs.append(silhouette_avg)
16
17 # 作圖並印出 k = 2 到 10 的績效
18 plt.bar(ks, silhouette_avgs)
19 plt.show()
20 print(silhouette_avgs)
```

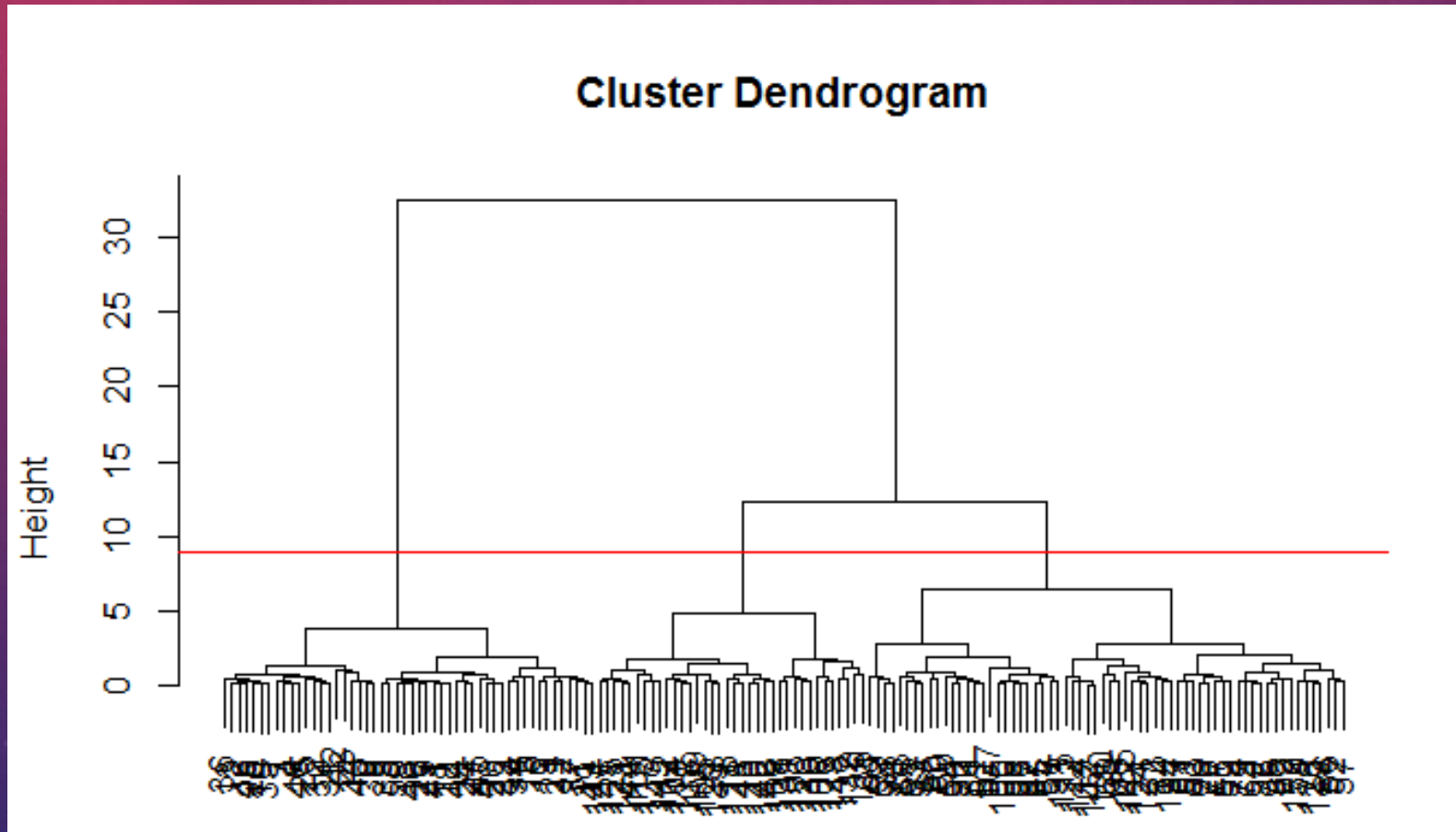
K從2到10的KMEANS效能



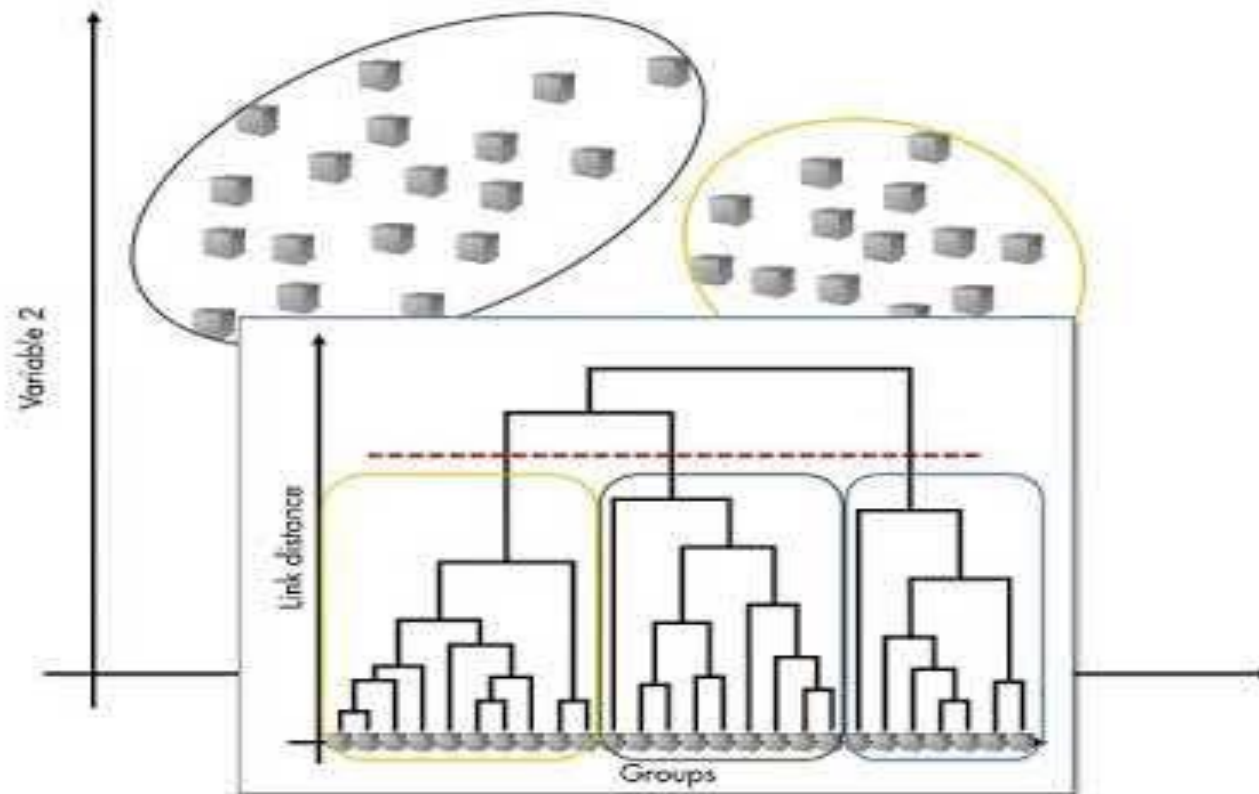
The background is a gradient from red at the top to blue at the bottom, with a starry space pattern. On the left side, there are several concentric circles and a large circular scale with degree markings from 140 to 260. Some circles have arrows indicating a clockwise direction.

階層式分群法 HIERARCHICAL CLUSTERING

PROCESSES



EXAMPLE



鳶尾花資料集

- 花瓣 (Petal) 的長
- 花瓣 (Petal) 的寬
- 花萼 (Sepal) 的長
- 花萼 (Sepal) 的寬

[5.1 3.5 1.4 0.2]

在設定某K的HIERARCHICAL CLUSTERING

```
1  from sklearn import cluster, datasets
2
3  # 讀入鳶尾花資料
4  iris = datasets.load_iris()
5  iris_X = iris.data
6
7  # 印出單筆測資
8  print(iris_X[0])
9
10 # Hierarchical Clustering 演算法
11 hclust = cluster.AgglomerativeClustering(linkage = 'ward', affinity = 'euclidean', n_clusters = 3)
12
13 # 印出分群結果
14 hclust.fit(iris_X)
15 cluster_labels = hclust.labels_
16 print(cluster_labels)
17 print("---")
18
19 # 印出品種看看
20 iris_y = iris.target
21 print(iris_y)
```

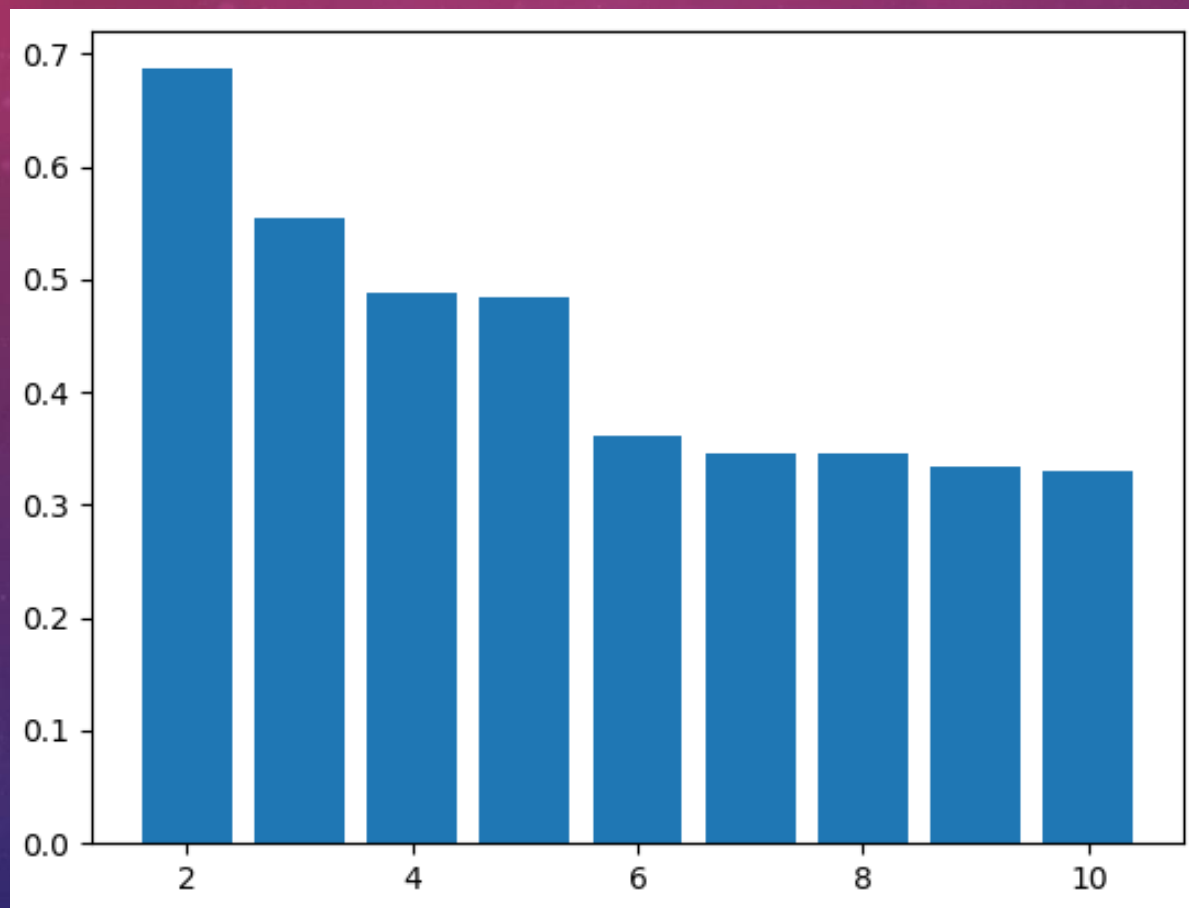
在設定某K的HIERARCHICAL CLUSTERING

[illegible]

K從2到10的效能

```
1 from sklearn import cluster, datasets, metrics
2 import matplotlib.pyplot as plt
3
4 # 讀入鳶尾花資料
5 iris = datasets.load_iris()
6 iris_X = iris.data
7
8 # 迴圈
9 silhouette_avgs = []
10 ks = range(2, 11)
11 for k in ks:
12     # Hierarchical Clustering 演算法
13     hclust_fit = cluster.AgglomerativeClustering(linkage = 'ward', affinity = 'euclidean', n_clusters = k).fit(iris_X)
14     cluster_labels = hclust_fit.labels_
15     silhouette_avg = metrics.silhouette_score(iris_X, cluster_labels)
16     silhouette_avgs.append(silhouette_avg)
17
18 # 作圖並印出 k = 2 到 10 的績效
19 plt.bar(ks, silhouette_avgs)
20 plt.show()
21 print(silhouette_avgs)
```

K從2到10的效能



The background is a gradient from deep red at the top to dark blue at the bottom, speckled with white stars. Overlaid on the left side are several concentric circular patterns. One large circle has a scale from 140 to 260 in increments of 10. Other circles have dashed lines and arrows indicating a clockwise direction.

自行練習

Wine Dataset

[1.207e+01, 2.160e+00, 2.170e+00, 2.100e+01, 8.500e+01, 2.600e+00, 2.650e+00, 3.700e-01, 1.350e+00, 2.760e+00, 8.600e-01, 3.280e+00, 3.780e+02]

(1) Alcohol → 1.207e+01

(3) Ash → 2.170e+00

(5) Magnesium → 8.500e+01

(7) Flavanoids → 2.650e+00

(9) Proanthocyanins → 1.350e+00

(11) Hue → 8.600e-01

(13) Proline → 3.780e+02

(2) Malic acid → 2.160e+00

(4) Alcalinity of ash → 2.100e+01

(6) Total phenols → 2.600e+00

(8) Nonflavanoid phenols → 3.700e-01

(10) Color intensity → 2.760e+00

(12) OD280/OD315 of diluted wines →

3.280e+00



THANK YOU