估算集群數量

設計說明：

1. 請使用DBSCAN 群聚演算法（Density-Based Spatial Clustering of Applications with Noise）撰寫程式，讀取data_perf.txt 檔中載入輸入資料，可得每個資料點的 x、y 值。

2. 設定 eps_grid = np.linspace(0.3, 1.2, num=10)，min_samples=5 的情況下，請找出最佳的 epsilon、最大的 silhouette score 及最佳集群數量。

估算集群數量

請依序回答下列問題：

1. 請填入最佳的epsilon（計算至小數點第四位後無條件捨去）？

2. 請填入最佳的epsilon對應之silhouette score最大值（四捨五入取至小數點後第四位）？

3. 請填入最佳集群數量（Estimated number of clusters）（整數）？

4. 讀取 data_perf_add.txt（由原data_perf.txt 加入下列五筆資料[1.65,1.91]、[2.77,4.98]、[5.82,2.56]、[7.24,5.24]、[-0.3,4.06]）。

   請填入 Best epsilon 對應之 silhouette score 最大值（四捨五入取至小數點後第四位）？

估算集群數量 原始檔

```
import numpy as np
# TODO

input_file = ('data_perf.txt')
# Load data 載入資料
# TODO

# Find the best epsilon
eps_grid = np.linspace(            )
silhouette_scores = []
# TODO

    # Train DBSCAN clustering model 訓練DBSCAN分群模型

    # Extract labels 提取標籤

    # Extract performance metric 提取性能指標

    print("Epsilon:", eps, " --> silhouette score:", silhouette_score)

    # TODO

# Best params
print("Best epsilon ="          )

# Associated model and labels for best epsilon
model =    # TODO
labels =  # TODO

# Check for unassigned datapoints in the labels
# TODO

# Number of clusters in the data
# TODO
print("Estimated number of clusters ="          )

# Extracts the core samples from the trained model
# TODO
```

```
import numpy as np
# TODO

input_file = ('data_perf.txt')
# Load data 載入資料
# TODO

# Find the best epsilon
eps_grid = np.linspace(          )
silhouette_scores = []
# TODO

    # Train DBSCAN clustering model 訓練DBSCAN分群模型

    # Extract labels 提取標籤

    # Extract performance metric 提取性能指標

    print("Epsilon:", eps, " --> silhouette score:", silhouette_score)

    # TODO

# Best params
print("Best epsilon ="          )

# Associated model and labels for best epsilon
model =   # TODO
labels =  # TODO

# Check for unassigned datapoints in the labels
# TODO

# Number of clusters in the data
# TODO
print("Estimated number of clusters ="          )

# Extracts the core samples from the trained model
# TODO
```

Reference - 輪廓係數法 Silhouette analysis

- 輪廓係數法的概念是「找出同群資料點內最近/不同群越分散」的值，也就是滿足 Cluster 的定義， a 為同群之間的距離，b 為不同群之間的點平均距離，S則越大越好，代表分得越清楚

$$S = \frac{b - a}{max(a, b))}$$

- 先用make_blobs(聚類數據生成器)來產生數據， scikit中的make_blobs() 函數常被用來生成聚類算法的測試數據，make_blobs會根據用戶指定的特徵數量、中心點數量、範圍等來生成幾類數據，這些數據可用於測試聚類算法的效果。
- 產生完後，可以使用 for 迴圈產生不同的 n_clusters 去看看何者輪廓係數較高。

Reference - 輪廓係數法 Silhouette analysis

- 分群演算法的績效可以使用 Silhouette 係數或 WSS（Within Cluster Sum of Squares）/BSS（Between Cluster Sum of Squares）。
- 使用 sklearn.metrics 的 silhouette_score() 方法，這個數值愈接近 1 表示績效愈好，反之愈接近 -1 表示績效愈差。
- Refer to ： sklearn.metrics.silhouette_score - scikit-learn 0.18.1 documentation

Compute the mean Silhouette Coefficient of all samples.
The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is (b - a) / max(a, b). The best value is 1 and the worst value is -1.

```
from sklearn import cluster, datasets, metrics

# 讀入鳶尾花資料
iris = datasets.load_iris()
iris_X = iris.data

# KMeans 演算法
kmeans_fit = cluster.KMeans(n_clusters = 3).fit(iris_X)
cluster_labels = kmeans_fit.labels_

# 印出績效
silhouette_avg = metrics.silhouette_score(iris_X, cluster_labels)
print(silhouette_avg)
```

Reference - 輪廓係數法 Silhouette analysis

- 輪廓係數（Silhouette Coefficient），是聚類效果好壞的一種評價方式。最早由 Peter J. Rousseeuw 在 1986 提出。它結合內聚度和分離度兩種因素。可以用來在相同原始資料的基礎上用來評價不同演算法、或者演算法不同執行方式對聚類結果所產生的影響。

Reference - 輪廓係數法 Silhouette analysis

方法：

1. 計算樣本i到同簇其他樣本的平均距離ai。ai 越小，說明樣本i越應該被聚類到該簇。將ai 稱為樣本i的簇內不相似度。簇C中所有樣本的a i 均值稱為簇C的簇不相似度。

2. 計算樣本i到其他某簇Cj 的所有樣本的平均距離bij，稱為樣本i與簇Cj 的不相似度。定義為樣本i的簇間不相似度：bi =min{bi1, bi2, ..., bik} bi越大，說明樣本i越不屬於其他簇。

3. 3，根據樣本i的簇內不相似度a i 和簇間不相似度b i ，定義樣本i的輪廓係數：

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$$

$$s(i)=\begin{cases}1-\dfrac{a(i)}{b(i)}, & a(i)<b(i)\\ 0, & a(i)=b(i)\\ \dfrac{b(i)}{a(i)}-1, & a(i)>b(i)\end{cases}$$

Reference - 輪廓係數法 Silhouette analysis

方法：

4. 判斷：

    1. si 接近 1，則說明樣本i聚類合理；

    2. si 接近 -1，則說明樣本i更應該分類到另外的簇；

    3. 若 si 近似為 0，則說明樣本i在兩個簇的邊界上。

    4. 所有樣本的 s i 的均值稱為聚類結果的輪廓係數，是該聚類是否合理、有效的度量。