ScikitLearn 操作記錄單 3

組別: _team15_____　　學號: _41147046S_____　　姓名: __楊子萱_____

Unsupervised Learning

1. 請根據以下教學資源操作: http://www.cse.msu.edu/~ptan/dmbook/tutorials/tutorial8/tutorial8.html

請自行查詢了解下列 scikit-learn 模組的功能作用 https://scikit-learn.org/stable/

程式碼連結: https://docs.google.com/document/d/1iLgXQBrAxYZiGHYVoEmOXQWWKFQQqkM1RpR087tqpJo/edit?usp=sharing

請根據需要的東西留下，其他部分註解後再跑

	Module	Function	試寫程式，實驗該函式所提供功能及主要參數設定效果
Clustering	Sklearn.cluster	KMeans()	KMeans(n_clusters=k, init='k-means++', random_state=42) 功能: 將數據分為若干群（n_clusters），以最小化群內樣本的距離平方和（inertia） **必要參數** 1. **n_clusters** ○ 指定要分成的群數。 ○ 選擇群數對聚類結果至關重要，肘部法則（Elbow Method）和輪廓係數（Silhouette Score）是常用方法。 2. **init** ○ 指定初始群心的選擇方式： ▪ 'k-means++'（預設）：優化群心初始位置，提升聚類效果和收斂速度。 ▪ 'random'：隨機選取群心位置。 ○ **效果**：'k-means++' 通常更穩定。

			3. **max_iter** ○ 設定最大迭代次數（默認為 300）。 ○ 如果數據量大，適當增大可能提升準確性。 4. **random_state** ○ 控制隨機性，用於保證結果可重現。 **3. 分群方法（其他聚類演算法）** 除了 KMeans，常用的分群演算法還包括： **(1) Hierarchical Clustering（階層式聚類）** 基於數據之間的距離，逐步合併或分裂群。 **(2) DBSCAN（基於密度的分群）** 識別高密度區域的數據為一群，適合處理不規則形狀群。 **分群效果評估** **(1) 肘部法則（Elbow Method）** 通過觀察不同群數對應的 inertia（群內樣本距離平方和）變化，選擇最佳群數。 • 當 n_clusters 過多時，inertia 減少趨勢變慢，該點即為最佳群數。 **(2) Silhouette Score（輪廓係數）** 衡量每個樣本的聚類質量，取值範圍為 [-1, 1]： • 越接近 1，聚類效果越好。 • 越接近 -1，樣本可能被分配錯誤群。 **(3) Davies-Bouldin Index** 衡量群內與群間距離的比例，數值越小越好。
	Sklearn.cluster	AgglomerativeClustering()	AgglomerativeClustering(n_clusters=optimal_clusters) 功能: 從每個樣本自身開始，逐步合併最相似的樣本或群，直到達到指定的群數（n_clusters）。不需要事先指定群數，而是可以根據距離閾值來進行分群。 **必要參數** 1. **n_clusters** ○ 指定最終要分成的群數。

			○ 類似於 KMeans，但 AgglomerativeClustering 也支持通過距離閾值（distance_threshold）來決定群數，而不必直接指定。 2. **affinity** ○ 定義相似度度量方法： ▪ 'euclidean'：歐氏距離（默認選項）。 ▪ 'manhattan'：曼哈頓距離。 ▪ 'cosine'：餘弦相似度。 ▪ 'precomputed'：使用事先計算的距離矩陣。 ○ **效果**：影響分群效果和計算速度。不同的相似度度量方法適用於不同的數據類型。 3. **linkage** ○ 定義如何計算兩個群之間的距離（聚合方式）： ▪ 'ward'：最小化群內方差（預設方法），計算兩群合併後的變異數。 ▪ 'complete'：最大化群內的最遠距離。 ▪ 'average'：使用群內的平均距離來衡量。 ▪ 'single'：最小化兩群中最短的距離。 ○ **效果**：不同的 linkage 方法會影響分群結果，特別是如何合併群。 4. **distance_threshold** ○ 指定一個距離閾值，如果兩群之間的距

			離小於該閾值，則合併它們。這是一種基於距離的停止標準。若指定，則可以不設置 n_clusters。 ○ **效果**：這個參數讓你能夠以更靈活的方式控制聚類結果。 **其他聚類演算法** **(1) KMeans** 基於距離的聚類方法，使用 n_clusters 指定群數，並且利用距離來最小化群內差異。 **(2) DBSCAN** 基於密度的分群方法，通過指定 eps（鄰域範圍）和 min_samples（最小樣本數量）來識別密度區域。 分群效果評估 **Calinski-Harabasz Index（Calinski-Harabasz 指數）** 這是一種基於群內和群間變異度來評估分群效果的指標，數值越大表示聚類效果越好 **Davies-Bouldin Index（Davies-Bouldin 指數）** 這是一種衡量分群結果的指標，數值越小越好，表示群間的分隔越清晰。它考慮群內的緊密度和群間的分隔。
	Sklearn.cluster	DBSCAN()	DBSCAN(eps=0.5, min_samples=5) 功能: 不需要指定預先確定的群數，它根據數據點的密度自動識別群集。這使得 DBSCAN 能夠處理形狀不規則的群集，並且能夠識別噪聲點 **主要參數** 1. **eps**： ○ 定義每個點的鄰域範圍。當兩個點之間的距離小於 eps 時，這兩個點被視為相鄰。 ○ 設定過小的 eps 可能會將許多點視為噪聲，設置過大則可能會將不同的群集

			合併為一個群集。 2. **min_samples**： ○ 形成一個群集所需的最小點數。 ○ 設置過小的 min_samples 可能會產生更多的群集，而設置過大的 min_samples 則可能導致某些群集無法形成。 3. **metric**： ○ 定義距離度量，默認是歐式距離（'euclidean'）。也可以設置為其他度量方式，如曼哈頓距離（'manhattan'）等。 4. **algorithm**： ○ 用來計算鄰域點的算法。選項包括 'auto', 'ball_tree', 'kd_tree', 和 'brute'。'auto' 會根據數據選擇最佳算法。 5. **leaf_size**： ○ 用於 ball_tree 和 kd_tree 算法中，控制樹的葉子節點大小，這會影響計算的速度。 **評估方法比較：** 1. **Silhouette Score**： ○ 用來衡量每個樣本是否被正確分配到聚類中。值範圍從 -1 到 1，越接近 1 越好，越接近 -1 表示樣本被錯誤分配到聚類中。 ○ DBSCAN 的 Silhouette Score 可能較低，因為它可能將許多點標記為噪聲（即標記為 -1）。 2. **Davies-Bouldin Index (DBI)**： ○ 衡量群集間的分離度和群內的緊密度，數值越小表示聚類效果越好。DBSCAN

			的 DBI 可能會較高，尤其在存在大量噪聲點的情況下，因為噪聲點不屬於任何群集，可能會影響分隔性。
Clustering Evaluation	Sklearn.metrics.cluster	normalized_mutual_info_score()	功能: 計算兩個聚類結果之間的互信息（Mutual Information，簡稱 MI），並將其標準化。其值範圍從 0 到 1，1 表示兩個聚類結果完全一致，0 表示兩個聚類結果完全不同。 **主要參數設定效果** 1. **labels_true**： ○ 這是代表真實標籤的數據，通常是已知的分類結果，用來作為評估基準。 2. **labels_pred**： ○ 這是算法產生的聚類結果，即預測標籤。這些標籤是聚類算法給出的群集標識符。 3. **average_method**： ○ 這是用來計算互信息的平均方法。常見選項包括： ▪ 'arithmetic'（算術平均，預設）：返回計算的平均值。 ▪ 'geometric'（幾何平均）：使用幾何平均來計算互信息。 ▪ 'max'（最大值）：返回最大的互信息。 4. **beta**（在某些版本中提供）： ○ 用於控制對標準化的影響的參數。這主要涉及對不同的聚類結果進行加權，這

			不是經常使用的參數。 **分群的效果評估方法比較** 1. **Silhouette Score**： ○ 這個指標衡量每個樣本的聚類效果，範圍從 -1 到 1，值越大表示聚類效果越好。若聚類結果與真實標籤相符，Silhouette Score 會接近 1。DBSCAN 可能會因為有噪聲點（標記為 -1）而導致較低的分數。 2. **Normalized Mutual Information (NMI)**： ○ 這個指標用來評估兩個聚類結果之間的一致性，與真實標籤的匹配度。NMI 的範圍是 0 到 1，1 表示完全一致，0 表示完全無關。當真實標籤已知時，這個方法能夠有效評估不同聚類算法的表現。
	Sklearn.metrics.cluster	silhouette_score()	sklearn.metrics.silhouette_score(X, labels, metric='euclidean', sample_size=None, random_state=None) 功能: 衡量樣本與其所屬群組的相似度與與其他群組的區別度。該指標的值範圍為 [-1, 1] **參數說明：** 1. **X**： ○ 輸入的數據集，通常是經過標準化處理的數值數據（如 KMeans 或 DBSCAN 聚類後的數據）。 ○ 其形狀應該是 (n_samples, n_features)，

			即每行為一個樣本，列為該樣本的特徵。 2. **labels**： ○ 聚類結果的標籤（每個樣本的群組標籤），形狀應該為 (n_samples,)。 ○ 如果使用 DBSCAN，這裡會包含標註為 -1 的噪音點。 3. **metric**： ○ 用來計算距離的度量方法，默認為 'euclidean'（歐式距離）。也可以使用 'manhattan'（曼哈頓距離）等其他距離度量方法。 4. **sample_size**： ○ 如果數據集很大，可以指定樣本數量以減少計算時間。默認情況下，使用全部數據來計算。 5. **random_state**： ○ 隨機狀態，用於確保結果的可重現性（特別是在隨機初始化的聚類方法中）。

其他參考資源:

- machine learning 參考書: "Introduction to Machine Learning with Python" 之 github code

github code of the books “Introduction to Machine Learning with Python”
https://github.com/amueller/introduction_to_ml_with_python/blob/master/03-unsupervised-learning.ipynb

Scikit Learn documentation(http://scikit-learn.org/stable/index.html)

- 尋搜尋其他可信網路資源