ScikitLearn 操作記錄單 3

組別: Team16　　學號: 41071102H　　姓名: 徐敏皓

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/

	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()	功能: 根據資料點之間的距離（通常是歐幾里得距離），將資料點劃分到最近的群組。共分成 k 個群。 KMeans(n_clusters, init, n_init, max_iter, tol, random_state, algorithm) n_clusters: 指定要分的群數 (預設為 8)。 Init: 初始化質心的方法，預設為 k-means++'。 n_init: 初始聚類執行次數，預設為 10。

			max_iter: 最大迭代次數，預設為 300。 tol: 收斂的容忍度，預設為 1e-4。 random_state: 隨機種子，用於結果的可重現性。 algorithm: 演算法選擇，預設為'lloyd'(支持'auto','full','elkan'等)。
	Sklearn.cluster	AgglomerativeClustering()	功能: 從每個數據點都是單獨的 cluster，按照距離或相似度，逐步合併最相似的 cluster，直到達到指定的群數或距離閾值。會將數據逐步合併成層級結構的聚類樹（dendrogram）。 AgglomerativeClustering(n_clusters, metric, linkage, distance_threshold, compute_full_tree, compute_distances) n_clusters : 指定要生成的 cluster 數量（預設為 2）。若設定為 None，則根據距離臨界值

			（distance_threshold）進行分割。 metric:用於計算距離的指標，取代 affinity。 預設為 euclidean。 Linkage: 用於合併 cluster 的方式，選項為： • 'ward'：最小化總平方差（僅適用於歐幾里得距離）。 • 'complete'：最大距離（最遠點）。 • 'average'：平均距離。 • 'single'：最小距離（最近點）。 distance_threshold: 若設置，定義停止合併的距離臨界值（忽略 n_clusters）。 compute_full_tree: 是否計算整個階層式聚類樹。當 distance_threshold 設定時，會自動設

			為 True。 compute_distances: 是否在模型中計算兩個 cluster 之間的距離（False 為預設）。
	Sklearn.cluster	DBSCAN()	功能: 基於密度的分群算法，適合處理具有噪 聲和複雜分佈的資料集。 DBSCAN(eps, min_samples, metric, metric_params, algorithm, leaf_size, p) eps: 兩個樣本被認為是鄰居的最大距離（即 鄰域的半徑）。 min_samples: 形成一個核心點所需的最少鄰 居數量（包括核心點本身）。 metric: 距離計算方式（預設為 'euclidean'）。 可以設為 'manhattan'、'cosine' 等，或自定義函 數。

			metric_params: 距離計算的其他參數。 algorithm: 計算最近鄰的算法（可選 'auto'、'ball_tree'、'kd_tree' 或 'brute'）。 leaf_size: 用於 BallTree 或 KDTree 的葉節點大小，影響速度。 p: 用於 minkowski 距離的冪參數（例如 p=2 對應歐幾里得距離）。
Clustering Evaluation	Sklearn.metrics.cluster	normalized_mutual_info_score()	功能: 評估分群結果與真實標籤的一致性的指標。 normalized_mutual_info_score(labels_true, labels_pred, average_method) labels_true: 類別標籤的真實值 (ground truth labels)，即分群的真實分配。 labels_pred: 類別標籤的預測值，即分群演算

			法的分配結果。 average_method: 預設值為 'arithmetic'。指定 正規化的計算方式，選項包括： • 'min'：使用最小值正規化。 • 'geometric'：使用幾何平均正規化。 • 'arithmetic'：使用算術平均正規化。 • 'max'：使用最大值正規化。
	Sklearn.metrics.cluster	silhouette_score()	功能: 評估聚類結果的質量，其功能是衡量每 個數據點與其所屬群集的相似性（凝聚度） 與其與最近的其他群集的相似性（分離度） 之間的平衡。 silhouette_score(X, labels, metric, sample_size, random_state) X: 數組或類似數組的結構（如 DataFrame），

			表示特徵數據或距離矩陣。 labels: 數組或列表，群集標籤。 metric: 字串，計算相似度的度量方式（默認為'euclidean'）。包括： 'euclidean','cosine','manhattan','precomputed'等。 sample_size: 整數，指定用於計算的樣本大小（默認為 None，即使用全部數據）。 random_state: 整數或 np.random.RandomState，當使用 sample_size 時用於再現性。 kwargs: 傳遞給距離計算函數的其他參數。

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