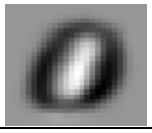
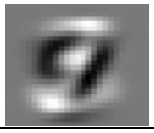
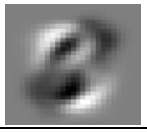
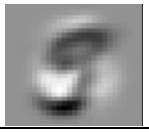
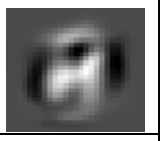
資料探勘 作業二

2025-10-21

Part I. 手寫數字影像特徵擷取

1. 使用 PCA 擷取特徵
   a. 畫出前十特徵（主成分）。
   b. 取出第 55, 1136, 2082, 2120 以及第 5422 張影像，以及其分別對應至前五特徵（PC1~PC5）的係數，完成以下表格。觀察這些數字影像，以及所對應的主成分係數，你是否可以看到他們之間的連結？請簡述。

	PC1 	PC2 	PC3 	PC4 	PC5 
55 	1062	331	-252	166	98
1136 					
2028 					
2120 					
5422 					

2. 使用 NMF 擷取特徵（基底）
   a. 設定特徵（基底）數為 25。任選十個特徵（基底），畫出這些特徵。注意：你必須先檢查 relative error 的趨勢，確認你得到的特徵滿足了你所定義的收斂條件。
   b. 觀察 2a. 中的 NMF 基底與 1a. 中的 PCA 主成分，在外觀與結構上的不同，以此說明他們在呈現資料的「組成方式」上有何差異。
   c. 針對 2b. 中觀察到的差異，請說明造成這些差異的關鍵因素。換句話說，NMF 與 PCA 在設定或限制條件上有哪些不同？這些差異如何導致 2b. 所看到的現象？
   d. 在 1a. 中我請你畫出前十個主成分，但在 2a. 中我請你畫出任選十個特徵，而不是前十特徵。請你推測，這是因為我想要向你強調 PCA 與 NMF 所找出的特徵組在哪一個性質上的差異？

3. 使用 t-SNE 擷取特徵（投影維度）
   a. 將 t-SNE 之投影維度設為三維。這三個維度是否能像 PCA 的主成分與 NMF 的基底，對應到可視化的影像，如 1a. 與 2a. 的作法？若可以，請畫出這三個維度；若不可以，請說明原因。

Part II. 手寫數字影像降維與分群

1. 視覺三維特徵空間中的群聚分布

   此處我們僅以三維特徵空間為基準，比較 PCA（擷取前三個主成分）、NMF（分解至三個基底） 與 t-SNE（投影至三維） 的表現，觀察它們是否能在新的特徵空間中清楚區分不同數字的群組。

   你可以使用 rgl:: plot3d(三維特徵空間的座標, col=影像的數字標籤) 來呈現影像於三維空間中的分佈。

2. 評估 k 維特徵空間中的群聚分布

   a. 對於 PCA，選擇 [解釋原資料約 75% 變異] 的主成分數量，作為特徵空間的維度。

   b. 對於 NMF，使用 50 個基底建構特徵空間。請確認這 50 個基底所產生的重建誤差，與使用 a. 中 [解釋 75% 變異量] 之主成分所產生的誤差相近。重建誤差定義為：

$$\sum_{i=1}^{28}\sum_{j=1}^{28}(x_{ij}-\tilde{x}_{ij})^2$$

   c. 對於 t-SNE，使用三維建構特徵空間。

   分別使用原始影像（784 維），以及原始影像於上述 a、b、c 三個特徵空間中的座標，以 K-Means 演算法（設定 k=10）對 MNIST 手寫數字影像進行分群，並以下方兩種指標評估分群表現：

   (i) 外在指標：Purity，以 [影像的數字標籤] 為標準。
   可使用 funtimes::purity( )

   (ii) 內在指標：Silhouette coefficient 的分布
   可使用 cluster:: silhouette( )

   對於內在指標，請提供兩種計算結果：使用原資料（784 維），以及使用特徵座標。

   根據上述指標，比較三種特徵擷取方法對於 MNIST 分群結果的影響與幫助。

Part III. 動物類別分群與特徵類別分析

資料 hw2_data_animals.csv 記錄了同學們以 [自己的觀點] 對動物做分類的結果。每一位同學的分類標準不同：有些同學可能著眼於物種屬性，有些可能是以棲息地分類，有些同學可能沒有具體訂標準，憑藉的是直覺。不論同學們採用何種視角做分類，我們都可以將每位同學的分類，視為動物的一種特徵。也就是說，此份資料記錄了每種動物的 24 項特徵。有些特徵（同學）描述的角度可能很相近，有些則可能很不同。以下請同學使用 Homogeneity Analysis 分析此資料。

1. 視覺化 [動物] 與 [變數類別] 於二維特徵空間中之相近度
   a. 以 HOMALS 二維特徵呈現動物的散佈圖。簡述你在圖中所看到的。
   b. 以 HOMALS 二維特徵呈現變數類別之散佈圖。請從圖中舉兩個例子。每個例子請呈現：哪兩位同學的哪兩個類別所在的位置十分相近？檢視原始資料，說明它們的位置為何相近。
   c. 以 HOMALS 二維特徵呈現
      (i) 鱷魚的位置，以及 [每位同學所指派鱷魚之類別] 的位置
      (ii) 鴕鳥的位置，以及 [每位同學所指派鴕鳥之類別] 的位置
      這些位置所呈現的遠近關係，與你所學到的 HOMALS 樣本與類別位置之解的設定，看起來是否相符合？簡單說明你的想法。

2. 動物的群聚分析
   a. 設定 HOMALS 投影維度為 30 維。畫出 eigenvalues 隨著維度數改變的趨勢。觀察 eigenvalues 是否在這其中已開始趨於平緩。
   b. 以投影 30 維為基準，計算前 20 維所解釋之變異百分比。
   c. 根據前 20 維之 HOMALS 座標，使用 Hierarchical Clustering 配合 Wald linkage 做動物的群聚分析，並且以 dendrogram 展現群聚結構。你喜歡這個 dendrogram 的模樣嗎？簡述你的想法。
   d. 基於這 20 維所對應的 eigenvalues 值是由大到小排列的性質，你認為 c. 的做法是否合理？
      (i) 檢查動物的 HOMALS 座標在每一個維度的變異是否相同。你所檢查的結果與 HOMALS 求解的設定是否相符合？
      (ii) 根據 c. 的檢查結果，你認為我們是否應該進一步調整動物的 HOMALS 座標於每個維度的變異，並且以調整後的座標做群聚

分析？說明你的看法。

(iii) 不論你在 c. 的答案為何，請你以 eigenvalues 調整動物的 HOMALS 座標，即：對於解釋變異越大的維度，我們給予該維度座標較大的權重：

```
hhx=hh$objscores[, 1:p]
hhx.w=sweep(hhx, 2, sqrt(e[1:p]), `*`)
```

使用調整後的座標，重新做一次 c. 中的分群。這次分群結果的 dendrogram 與 c. 中的 dendrogram 的差異在哪裡？你比較喜歡哪一個 dendrogram？說明你的原因。