Data Science HW2

【資料集說明】

目標：透過 19 個 attributes 去分析各個 data 是否為同一群。

在此 dataset 中，每筆 id(rows)代表一張圖片，而一張圖片通常可以用多個屬性(columns)集合來表示，以下為各屬性所代表的意思，請利用這些屬性將資料做分群(也就是同一群內的圖片具有較相似的屬性)。

Feature1-2: 該圖片中物件的中心點座標值

Feature3: 將圖片預先分成九個區塊，故皆為 9

Feature4: line extraction 的結果中低對比且長度小於等於 5 的區域則為 1，並且正規化

Feature5: line extraction 的結果中低對比且長度大於 5 的區域則為 1，並且正規化

[註: line extraction algorithm 是在做圖像分析時會用到的特徵萃取演算法]

Feature6: 該區域在水平鄰近像素對比度的平均值

Feature7: 該區域在水平鄰近像素對比度的標準差

Feature8: 該區域在垂直鄰近像素對比度的平均值

Feature9: 該區域在垂直鄰近像素對比度的標準差

Feature10: 每個區域的密度平均值或所謂的灰階值，公式為(R+B+G)/3

Feature11-13: RGB 在該區域各自的平均值

Feature14-16: RGB 在該區域各自的 excess value，公式分別為(2R-G-B)、(2B-G-R)、(2G-R-B)

Feature17-19: 分別為明度、飽和度、色相的平均值

data.csv

id	feature1	feature2	feature3	feature4	feature5	feature6	feature7	feature8	feature9	feature10
0	97	123	9	0	0	1.5555553	1.0962937	13.222221	108.74074	52.814816
1	37	189	9	0	0	1.3888893	1.4851844	5.722223	23.885176	30
2	148	169	9	0.11111111	0	1.0555553	0.7740734	1.833333	2.6111097	19.592592
3	95	57	9	0	0	1.8333327	3.4111106	2.1111107	1.7185175	26.296297
4	200	189	9	0	0	1.1111113	1.0886619	3.1666667	1.0055408	49.666668
5	77	199	9	0	0	2.666668	1.8618976	4.666666	3.025815	57

id 為每項 data 的編號，每項 data 總共有 19 個 attributes，全部有 2100 筆資料。

test.csv

index	0	1
0	1303	1234
1	1710	878
2	1587	1637
3	892	119
4	83	940
5	120	463

在此文件內，顯示的是我們需要比較是否為相同 cluster 的 data 編號，例如 index 0 要比較的資料為 id 1303 以及 id 1234。

submit.csv

index	ans
0	
1	
2	
3	
4	
5	

需要將預測結果（0：不為同一群 / 1：同一群）寫進 submit.csv 內，也就是在 index 0 的 ans 需要把 id 1303 & id 1234 是否為同一群的結果寫入，同理 index 1 的 ans 需要寫入 id 1710 & id 878 是否為同一群的結果，以此類推，總共要預測 400 筆資料。