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 extractionalgorithm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excessvalue，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	86	140	9	0.000000	0.0	5.444444	4.768726	3.055556	2.678447	9.925926
1	34	93	9	0.000000	0.0	0.555556	0.272165	0.388889	0.389682	14.814815
2	134	148	9	0.000000	0.0	0.111111	0.172133	0.055556	0.136083	0.037037
3	199	144	9	0.000000	0.0	0.333333	0.516398	0.333333	0.365148	0.444444
4	197	236	9	0.000000	0.0	2.444444	6.829628	3.333333	7.599998	16.074074

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。