資料科學導論 HW2

【資料集說明】

目標：透過每首歌曲的特徵資料，將歌曲資料進行分群。

資料集中的每個 row 代表一首歌曲，每首歌曲有自己獨立的 song_id，並附有 13 個特徵，請利用這些特徵將歌曲分群，盡量將相同種類的歌曲分在同一群中。歌曲種類可能有電子樂、嘻哈、搖滾等類型，而同一種類型的歌曲，通常具有較相似的特徵。

以下為各特徵所代表的意思：

Feature 1：歌曲的原聲程度。越接近 1，表示歌曲所包含的電子音樂成份越少。

Feature 2：歌曲表現的強度。強度較高的歌曲，會讓人感到有活力、響亮、甚至吵雜。

Feature 3：根據歌曲的節奏與穩定性，來評斷歌曲是否適合作為舞曲。越接近 0 的值，越不適合；越接近 1 的值越適合。

Feature 4：歌曲的流行度，數值是根據播放次數作為依據，數值越高表示越流行。

Feature 5：歌曲的速度，以每分鐘節拍數 (BPM) 為單位。

Feature 6：歌曲在有現場觀眾的情況下進行錄製的機率。數值越高表示歌曲越有可能是現場即時錄製，或是演唱會版本。

Feature 7：歌曲傳達的情緒。數值越高，表示歌曲聽起來越積極 (快樂、輕快)；數值越低，表示歌曲聽起來越消極 (憤怒、悲傷)。

Feature 8：歌曲的持續時間，單位為毫秒。

Feature 9：歌曲的響度，以分貝為單位 (dB)。

Feature 10：歌曲中是否存在「口語」。若數值偏高，可能是脫口秀、Podcast；若數值偏低，則可能是純音樂。

Feature 11：歌曲的調性 (大調=0，小調=1)。

Feature 12：歌曲中的樂器演奏所佔比例，數值越高，表示歌曲的樂器演奏佔比越大。

Feature 13：歌曲的音高，內容是依據標準的 Pitch class 來進行映射。

【檔案說明】

Train.csv

song_id	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	Feature 10	Feature 11	Feature 12	Feature 13
0	0.147	0.798	0.745	46	111.016	0.197	0.388	240000	-5.436	0.0384	1	0.651	F
1	0.0658	0.804	0.521	66	143.952	0.0521	0.553	224700	-4.395	0.0569	0	0	C
2	0.0395	0.96	0.755	67	99.023	0.332	0.661	170440	-3.189	0.123	1	2.43E-05	E
3	0.359	0.769	0.592	40	171.94	0.122	0.223	226520	-7.127	0.19	1	0.0143	D
4	0.16	0.838	0.769	83	93.996	0.0935	0.602	249609	-5.238	0.0633	0	0	D
5	0.209	0.632	0.511	39	135.971	0.145	0.342	225813	-5.462	0.0286	0	0	A
6	0.0368	0.403	0.226	40	163.959	0.201	0.294	265200	-16.024	0.0445	1	0.205	F
7	0.934	0.127	0.31	10	135.423	0.109	0.298	349533	-23.284	0.0498	0	0.387	A#
8	0.697	0.37	0.41	55	76.926	0.211	0.33	318846	-15.757	0.0766	1	0.000144	E
9	0.0441	0.737	0.809	77	80.025	0.341	0.367	226938	-5.186	0.108	0	0	C#
10	0.0486	0.717	0.842	51	81.495	0.292	0.552	204387	-4.158	0.241	1	0	B

song_id 為每首歌曲的編號，每首歌曲共有 13 個特徵，全部有 40,114 首歌曲。

Test.csv

id	col_1	col_2
0	16868	39362
1	7661	30499
2	35255	18705
3	23946	31785
4	13606	25069
5	27050	25981
6	25157	3309
7	20049	4075
8	25019	6067
9	13643	10546
10	32135	23058

在測試資料中，需要判斷 col_1 和 col_2 的歌曲編號(song_id)，是否屬於同一群。

例如：

id=0，要判斷歌曲編號 16,868 和 39,362 的兩首歌曲是否屬於同一群。

Submit.csv

id	ans
0	
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

在繳交檔案中，需要將判斷結果寫入 Submit.csv，若兩首歌曲是屬於同一群，則寫入 1；若不屬於同一群，則寫入 0。

例如：

id=0 的 ans 欄位，需要寫入歌曲編號 16,868 和 39,362，是否為同一種類的判斷結果。依此類推，共預測 4000 首歌曲。

Kaggl 網址:https://www.kaggle.com/t/c7258dc752a6479989c8e28c161678ff