Final Project Proposal

Team member 周伯宇 林子鵑 呂宸漢

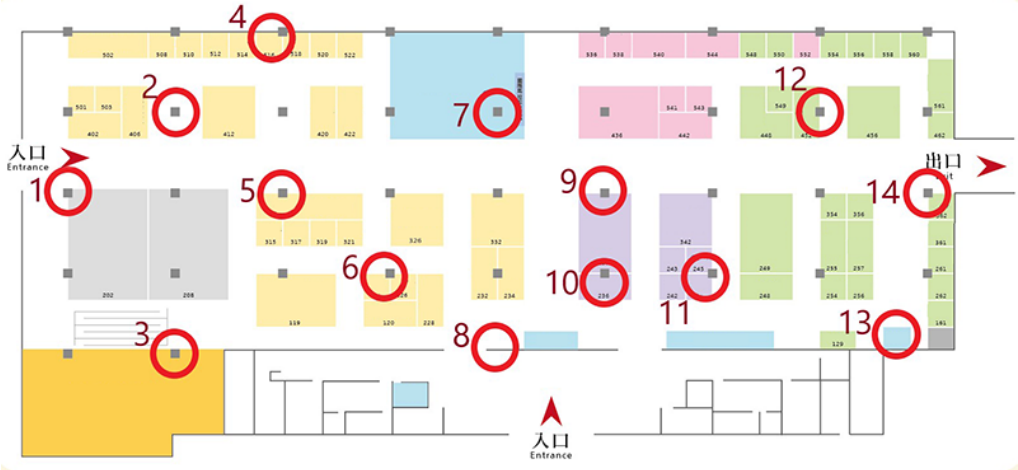
Project Title

Behavior Classification of Exposition Visitors（馬拉松博覽會參訪動線類別預測）

Dataset

本次題目給了四個 csv 檔，以及一個展場配置圖，說明如下：

1. train.csv：訓練所需的樣本數據資料（CSV），共有 3 個特徵，41,640 筆。
   mac_hash：樣本代號。
   sniffer_loc：樣本資料採集地點。
   created_time：樣本資料採集時間。
2. training-label.csv：訓練所需的樣本數據資料標記（CSV）。
   mac_hash：樣本代號。
   label：樣本類別標記。
3. test.csv：測試所需的樣本數據資料（CSV），樣本特徵如同 train.csv 但無標記變項，一共有 20,419 筆。
4. submit_samples.csv：模型預測結果上傳範例格式檔。
5. marathon-map.png：展場群眾收集器配置圖。



Methodology

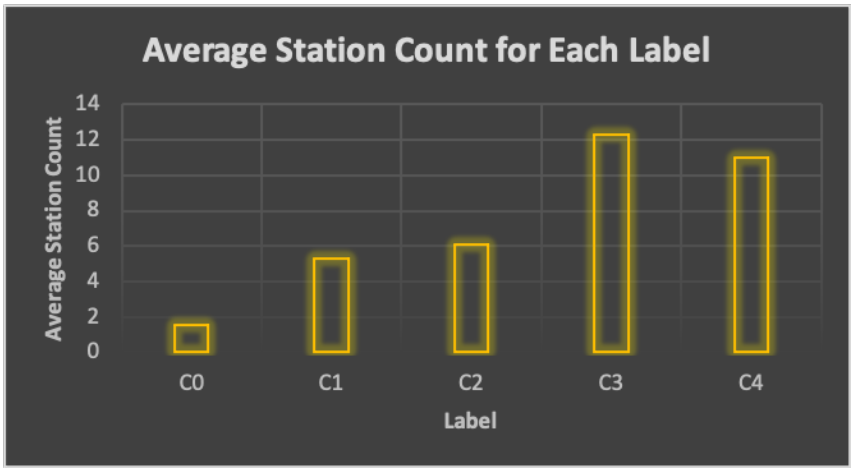
根據 training data 以及 training label 給定的資料，我們可以得知每一個樣本在特定時間經過哪幾個 sniffer location，而我們可以以樣本代號找到其相對應的類別標記。

1. 資料觀察

透過 train.csv 以及 training-label.csv，我們可以將資料做統計數據，根據經過的 sniffer_loc 以及停留的站點數量等等資訊作為觀察目標。

- 每一個 Label 停留的平均站點數量：

Label	C0	C1	C2	C3	C4
Station Cnt	1.52583	5.31506	6.05625	12.29480	10.96780



- 每個 Label 在 14 個站點停留的機率值

我們將每個 Label 經過 14 個 sniffer_loc 的機率值計算出來。不難從下面表格中觀察到，每個 label 都有他比較常出現的 sniffer_loc 編號，舉例來說：像是 C0, C1 同時經過 2 和 3 的幾率蠻高的，以及同時經過 9, 10, 11 的話很可能是屬於 C2 等等推測。

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
C0	1.93	12.89	17.24	5.85	8.83	4.4	3.99	4.23	5.37	2.78	4.74	13.88	6.63	7.23
C1	3.06	17.87	17.11	11.32	11.82	10.19	2.49	5.05	5.4	2.61	3.68	4.66	2.27	2.47
C2	0.19	4.73	3.49	6.98	1.8	6.67	5.63	9.27	13.56	11.81	14.1	9.82	6.93	5.04
C3	7.75	8.01	7.69	7.59	7.61	7.91	7.48	7.76	7.87	7.17	7.75	7.07	5.76	2.58
C4	0	8.82	8.08	8.58	7.77	8.4	4.84	8.13	8.67	7.94	8.4	8.34	7.21	4.82

- 每個 Label 在 7 個分類站點停留的機率值
  從下方表格中我們將地圖中相同顏色區塊（商品類型相同）的 sniffer_loc 分成七大類，再透過統計得到五種 label 經過七大類的機率值。也可以從中得到一些資訊：像 label 為 C4 的樣本可能都不是從左側入口進入，因為他們都不會經過 P1 這個類別；或是在 P2 經過的機率最高的很有可能是 label 為 C1 的樣本。

	P1(1)	P2(2, 4, 5, 6)	P3(3)	P4(8)	P5(7, 13)	P6(9, 10, 11)	P7(12, 14)
C0	1.93	31.97	17.24	4.23	10.62	12.89	21.11
C1	3.06	51.2	17.11	5.05	4.76	11.68	7.13
C2	0.19	20.17	3.49	9.27	12.57	39.46	14.86
C3	7.75	31.12	7.69	7.76	13.24	22.79	9.64
C4	0	33.58	8.08	8.13	12.05	25.01	13.16

- 每個 Label 的停留時間統計
  根據 train.csv 中的第三個 feature，我們可以知道某些 label 的樣本會在每幾個 sniffer_loc 停留較長的時間，亦或是在某些 sniffer_loc 停留較短暫的時間等等資訊，也可以作為判斷依據。

2. Feature 前處理
   經過上面的觀察圖表，我們可以統整出一些樣本的行為模式，但由於 train.csv 只提供了三個 features 供我們作為 training 的資訊，我們認為在這個問題當中，應該將 14 個 sniffer_loc 分開來考量，因此我們將針對 train.csv 做一些簡單的前處理當作我們的 feature。以下提出三種可能的處理方法：

- Method 1 : 14 sniffer_loc with stay time consideration
  14 個 sniffer_loc，以停留的時間長短給定一個 weight 值，把每一個樣本的數據做統計過後，停留最久與最短的時間切成 5 等分，對於一個樣本，分別給予每一個 sniffer_loc 一個介於[0, 5]之間的 weight 值。

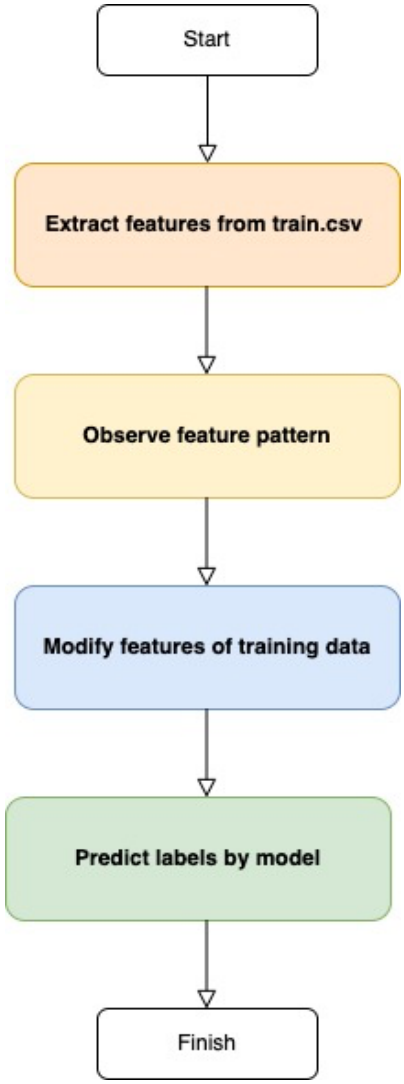
- Method 2 : 7 sniffer_group with average stay time consideration
  與 Method1 前處理的方式差不多，但只考量以 map 中同顏色區塊為一組的 sniffer_group，時間以其中包含的 sniffer_loc 所停留的時間加總平均作為考量。

- Method 3 : sniffer_loc/group without stay time consideration
  只考量有沒有經過該 sniffer_loc，有經過給 1 沒經過給 0，忽略 time_stamp 資訊作為 feature 考量。

3. 預測 Label

   將前處理完成的 train.csv 拿進來作為 features，加入一些我們在觀察 dataset 所得到的結論，像是某些比較常被參訪到的 sniffer_loc，或是參訪的 sniffer_loc 特別多，我們就會將它可能的 Label 機率值乘上一定的權重來增加預測的準確度，預計使用的方法可能是以 scikit learn 中 decision tree 作為基礎來預測 label。

4. Flow Chart



5. 可能遇到的困難

   前處理後 Feature 數量的上升可能導致 model 的複雜度過高，且整體訓練出的可能是一個非線性的問題，且對於 output 需要輸出每個 label 的機率值也是一個需要考量的因素，針對這些困難點我們仍然在尋找適當的 model，使得訓練的準確率更好。