機器學習期末專題報告

0416211黃紹瑜、0416033黃品任、0416094 黃兆宇、04160206李明峻

**主題 : 國道一號車速預測**

一、實作想法 :

利用交通部高速公路局網路提供的**eTag**資料庫和其他網路資料來預測某一時間點、一特定位置的車速以判斷交通狀況。

二、資料處理：

1,使用資料 :

1. 交通部高速公路局ETC資料庫M05A

2. 中央氣象局地面觀測站資料

3. ETC測站之經緯度

4. 中央氣象局各測站之經緯度

5. 高速公路車速速限

2,分析資料：



Direction (N, S) ：南、北向(ETC測站資料)

Miles (km) ：高速公路的里程數，由測站編號分析後得出

(ex: 01F0413S = 南向041.3km )

Latitude, longitude ：測站的經緯度，因里程數更能代表高速公路上的位置，不加入分析。

Speed\_limit (km/hr) ：該測站之速限。

Day ：以一周7天來表示(0=星期日，1=星期一……)。

Time (min) ：M05A資料每5分鐘更新。

Rainfall (mm) ：降雨量。

Hpa ：氣壓。

Speed (km/hr) ：平均車速。我們將車速分為7個區間：

<60 = -1, [60,70) = 0, [70,80) = 1, [80,90) = 2, [90,100) = 3, [100,110) = 4, >=110 = 5

加入天氣資料的方法：

Step1. 修改門架編號格式與門架經緯度資料相符

Step2. 找到門架經緯度

Step3. 根據上下游測站門架算出中心點經緯度

Step4. 找出與該經緯度最接近的氣象站(調整時間格式)

Step5. 用時間做線性內插(因為氣象資料間格為1hr、etag間格5min)

三、分析結果：

1.環境、library：

Python3.6.3

Goepy(加入天氣資料), Sklearn, numpy, matplotlib(輸出圖片)

2.使用model：

Random forest

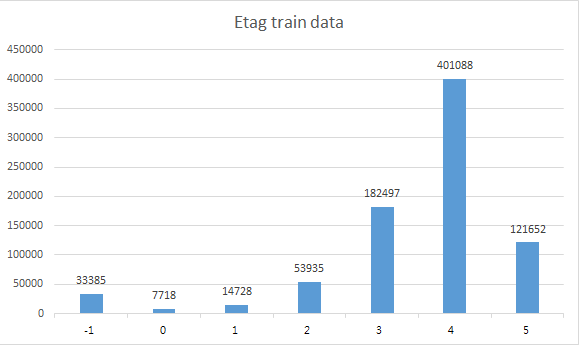
K-nn k-d tree

Naive bayes

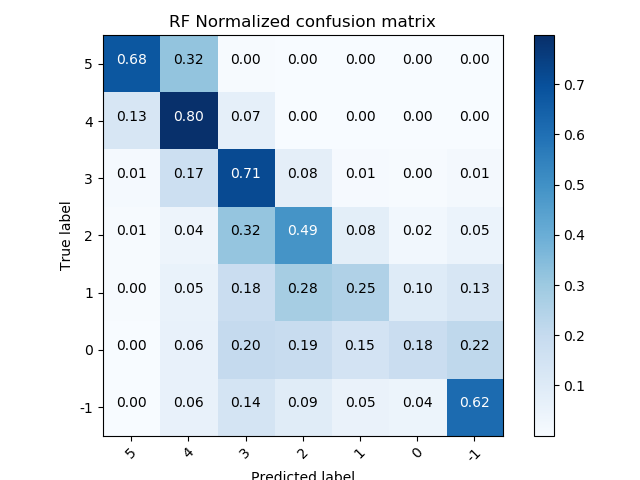
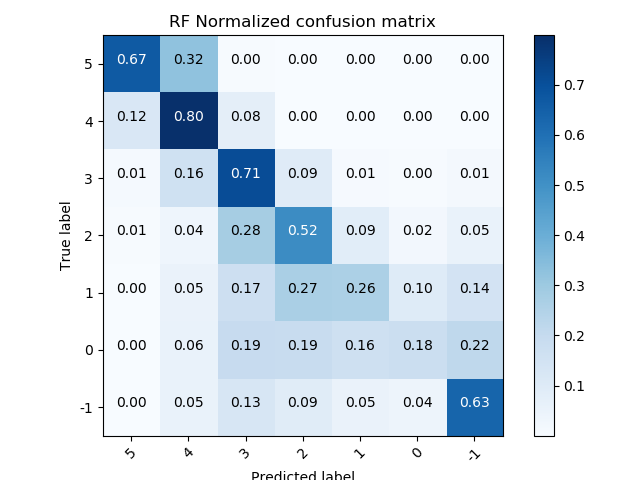
SVM

3.結果：

* Trainning data: 11 / 26 0: 00 ~ 12/16 23:55 共885736筆
* Test data: 12 /17 0:00 ~ 12 /23 23:55 共295187筆



1. Random forest:

**含天氣accuracy : 0.715709025126 不含天氣accuracy : 0.716711779313**

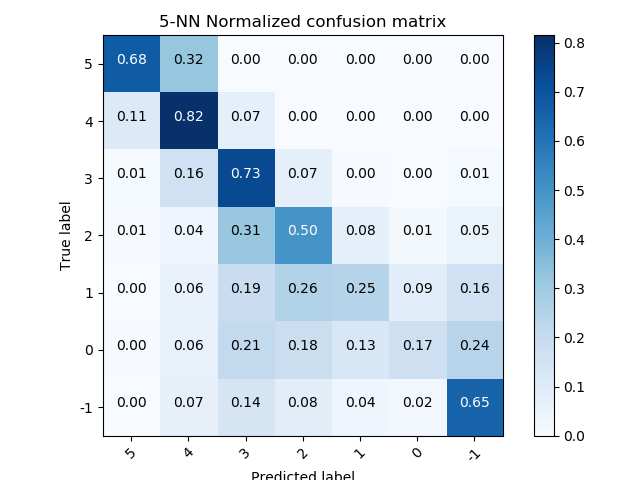
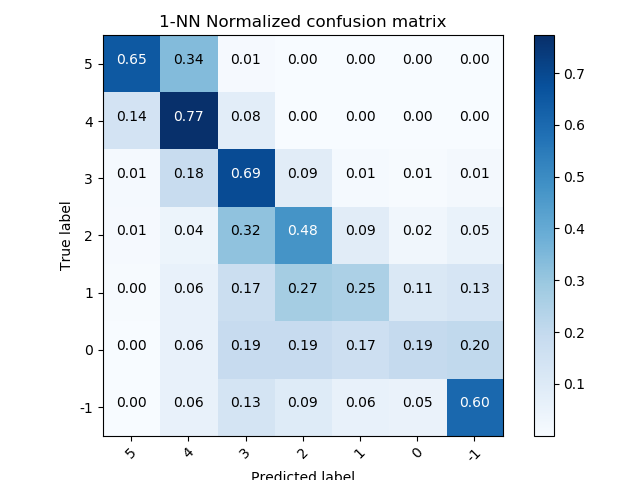
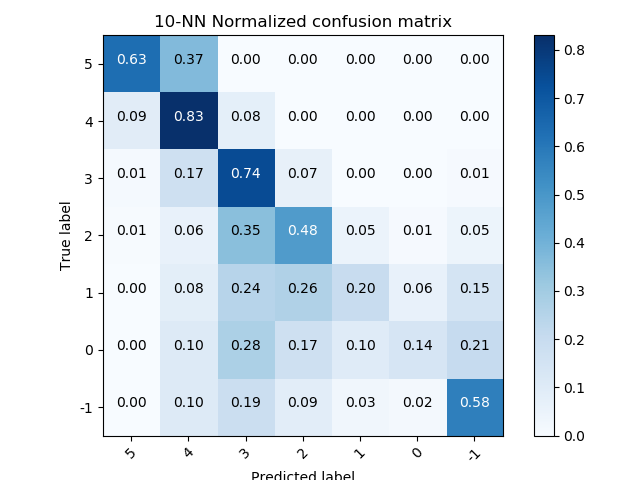
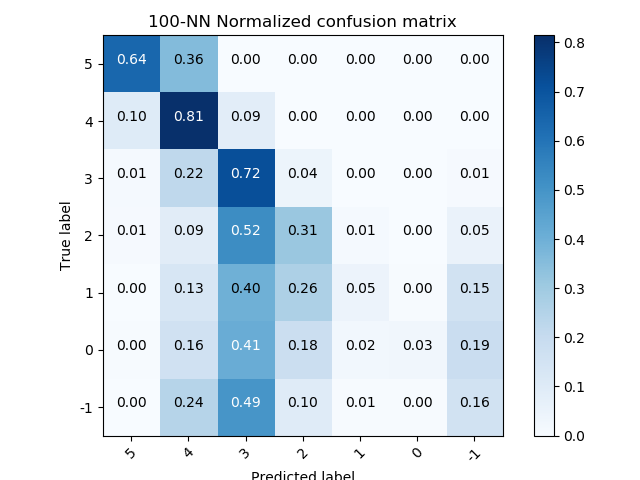
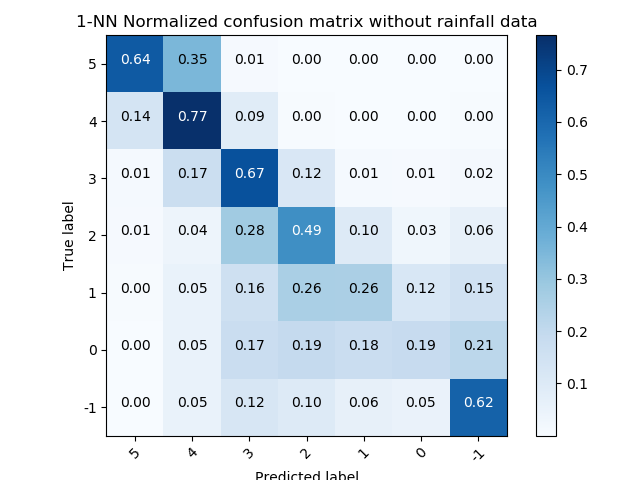
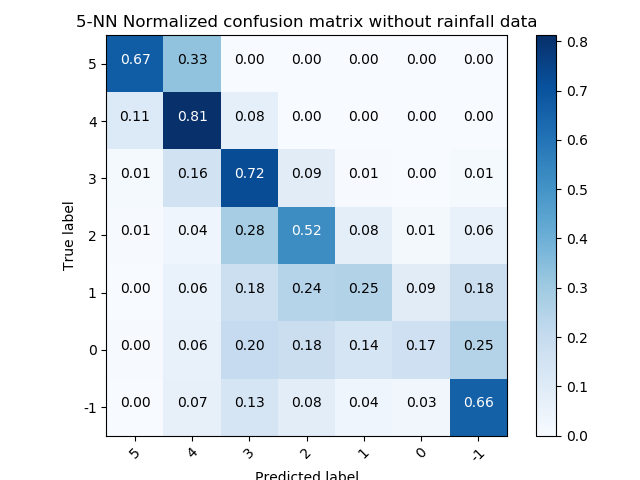
1. K-nn

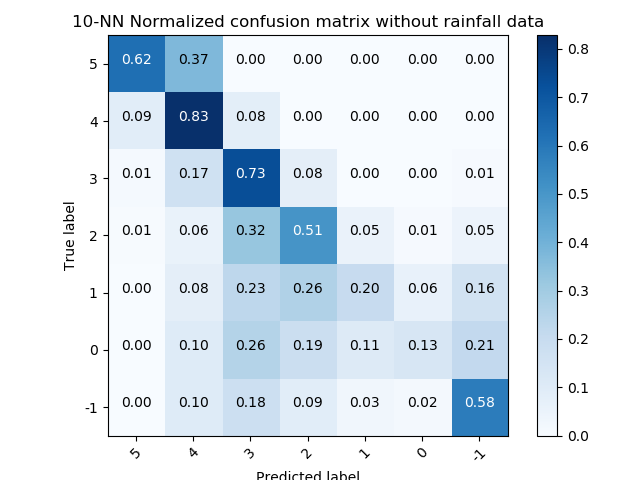
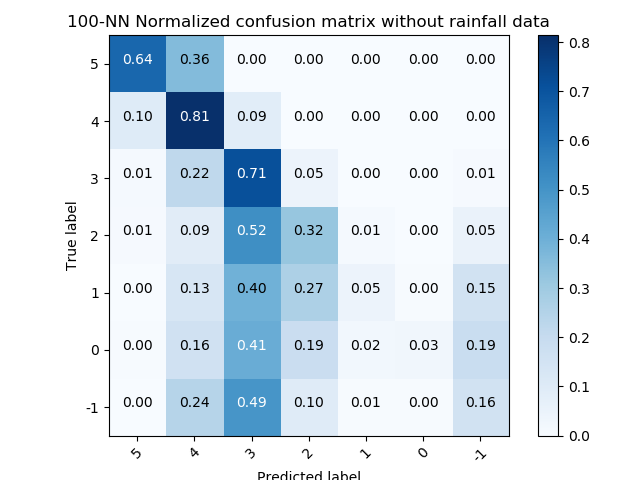
accuracy with rainfall without rainfall

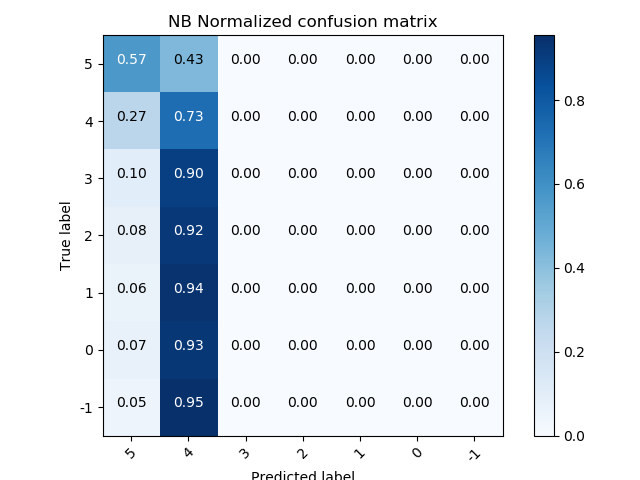
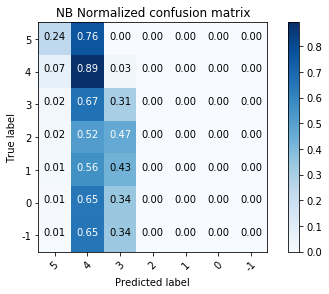
1-nn 0.69274392165 0.68404773923

5-nn 0.730028761429 0.726163415056

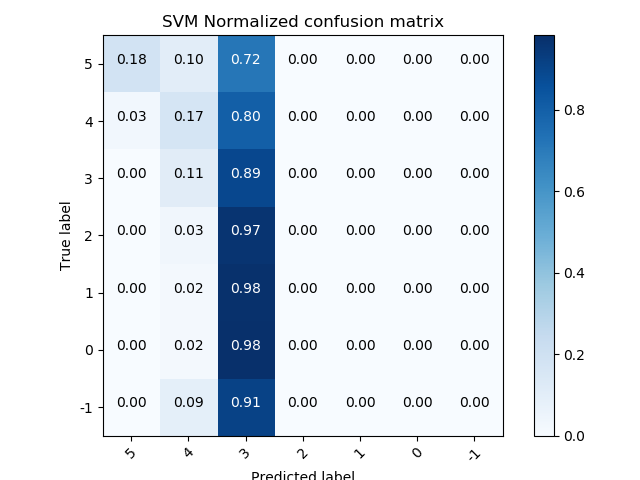
10-nn 0.724330678519 0.722433576004

100-nn 0.682367448431 0.681910111218

1. Naive bayes

**不含天氣accuracy : 0.529372905988 含天氣Accuracy : 0.446025739616**

1. SVM

**含天氣 accuracy 0.418033314475**

三.結果分析

根據training data target 的分佈可以看出，大部分的車速都維持在90~100km/hr，資料分布不均導致在預測的target為資料較少時，準確度有所降低。雖然資料看起來很多，但因為取樣時間只有三周，導致每一時間點實際上只有三筆不同的資料，扣除一些有問題的資料，使得可以用來training 的資料更少。

在RandomForest方面，有雨量的accuracy 與無雨量的十分接近，甚至還略差一點，我們推測這可是因為我們的雨量來源太少，導致資料不夠精確得到當時的真實雨量。

另一方面，雨量資料極大多數為0，這也導致Naïve Bayes 的準確率進一步降低，因為這明顯不屬於Gaussian distribution。再者我們的feature也並沒有符合NaiveBayes的假設，例如:速限跟里程數是一一對應關係，並不是在條件機率下independet.

k-nn在k值較低時較不受資料分布不均的影響，命中率較高。但k = 100 時可以看到速度區間為-1的recall比其他k值差了許多，更容易判斷為3或4。

因為sample的數量太大，我們嘗試用除了linear的kernel 丟給SVM跑但是跑不完，在經過討論後，我們使用sklearn 的 LinearSVC 才跑出結果，但是結果顯然是training instance 的分布不屬於linear 導致accuracy其差無比。

四、問題討論

* 氣象局提供之天氣資料之測站數不夠密集，導致無法取得精確雨量資料，且無天氣先關資訊，欲取得更詳細資料須付費購買。
* 近日缺乏降雨，導致dataset中降雨量資料少有變化。
* 根據高度修正氣壓資料，始可使用。
* 缺乏跟能見度有關的氣象資料
* 在接近最低時速區間資料不足，資料皆集中 > 90 or < 60
* 部分路段資料問題(ex. 休息站)，希望可以找到清理這些資料的方式
* 增加是否為連續假期作為feature
* 取得其他來源平均時速資訊作為驗證，例如國道VD
* 加入附近流動人口數量，目前政府公布的數據只有以直轄市等大地區設籍人口為單位，難以配合高速公路的分布
* 未來可以加入國道三號資料，或針對一路段做長時間的分析

五、參考資料

[http://eservice.cwb.gov.tw/wdps/obs/state.htm#%E7%8F%BE%E5%AD%98%E6%B8%AC%E7%AB%99](http://eservice.cwb.gov.tw/wdps/obs/state.htm)(中央氣象局測站資料)

<http://cit-etc.net1.tw/download.html> (交通部高速公路局)

六、工作分配

|  |  |  |
| --- | --- | --- |
| 組員姓名 | 學號 | 工作貢獻 |
| 黃紹瑜 | 0416211 | 資料蒐集、資料前處理、Naïve Bayes model分析 |
| 黃兆宇 | 0416094 | Random Forest model 分析、ptt製作 |
| 黃品任 | 0416033 | K-nn Kd-tree model 分析、書面報告撰寫 |
| 李明峻 | 0416206 | SVM model 分析 |