KDD CUP

組員：陳易煒
學號：M10715053

資料描述

- Plans
  - Sid : 每筆資料的ID
  - Plan_time： APP產生路線規劃資料的時間
  - Plans : 各種路線規劃包含其距離價錢及花費時間(最多11種)
- Queries
  - Pid : 用戶的屬性
  - Req_time : 用戶提出需球的時間
  - O, D : 起點及終點
- Profile
  - P0 ~P65 : 用戶的的特徵

實驗環境及做法

- 平台：Windows10, Linux
- 工具：Tensor-flow, Keras, Scikit-Learn
- 使用的model：MLP, KNN, SVM, Decision Tree, Random Forest
- 使用feature：Pid, req_time, Plans, o, d

選用的feature及其處理

- req_time -> day_of_week(星期一～五),quarter_hour(每十五分鐘分成一段)
- Pid -> 透過有無pid將資料拆成分兩部分，並分開train
- Plans -> 資料數量不一致將資料不足部分補0（共取七筆資料），並增加一個mode0
- o, d -> 將起點及終點的經緯度拆開分
- Weather -> 包含溫度濕度氣候
- Distance -> o, d換算成距離

期中後的改進

- 將經緯度換算成距離
- 將天氣分得更細，加入溫度濕度等因素
- 使用Scikit-learn PCA對profile降維
- 將各個model train出來的結果(knn,svm,cnn,decision tree)加進來一起train

結果及遇到的困難

- 分數：0.68903657；排名：514
- 始終找不到合適的方法填補缺失的資料。
- 加入溫度濕度只有些微的提高。
- 使用的model準確率始終無法突破0.69。

比賽心得

- 個人覺得這次比賽的資料有點奇怪，每個人的分數幾乎都差不多，不管用什麼方法改善的的程度都非常有限，這讓比賽的體驗變得非常差。雖然如此但是也因為這樣，我在這次比賽中為了解決問題嘗試了非常多以前不曾接觸的東西，像是了解了很多不同的機器學習模型，以及許多不同的正規化方法等等，雖然比賽沒能打出好成績但學到了這麼多也是值了。

組員分工

- 外部資料收集：他校同學
- 資料處理：陳易煒
- 模型建置/訓練：陳易煒＋他校同學
- 報告：陳易煒

Thanks for listening