機器學習期末project報告

R04325008　社研二　歐乙平

工作分配

1. 邏輯斯迴歸模型（Logistic Regression Model）

因為問題這次期末project的作業是二元分類的問題，所以第一個選擇的模型是可以處理分類問題、計算比較快、比較簡單的logistic regression model作為基準線。

1. 一開始只有使用第一天的資料，且只使用136使用者特徵。因為此次資料1、0的比懸殊，若沒有使用權重（weight），訓練後所得的邏輯斯迴歸模型會全部預測0，這會使得track2的精確率（precision）等於0，使得f1分數為0。處理這個問題的方式包括在既有的樣本中重新抽樣（oversampling or downsampling）或是在運算時考慮權重。因為原本的樣本數已經很大了，所以再抽樣更多1的樣本會讓計算更花費時間，而減少0又會失去某些資訊，因此我們選擇在最佳化時給予樣本不同的權重（給1更多的權重）強迫模型更準確的預測1。這個最基本的模型public score為0.097694。
2. 接下來我們試著嘗試更多的特徵（feature）以及使用所有的資料。在原本的資料當中有時間戳（timestamp）和ID這兩個資訊尚未被使用。我們將時間戳轉換為周日（weekday）和小時（hour），均以one hot的方式編碼，前者分為以7個特徵代表，後者以24個特徵代表。另一個原始資料中存在但未被加以利用的資訊是ID。我們嘗試了兩種在模型中考量ID的方法。第一種方法是將所有的資料依照ID重新整理，也就是將所有的資料依據分成623份[[1]](#footnote-1)，每一份裡的觀察值來自所有的檔案，且ID均相同。接著我們針對每一個ID訓練一個模型，訓練出623個模型。在預測測試資料的時候，每一筆資料都已與其ID相對應的模型預測。這種方式的好處是讓每一個ID的模型均不同，讓不同特徵對不同ID有不同的係數，也就是影響，增加機模型的複雜度。缺點則是每一個模型所使用的資料量較少。在這種方式下，使用第一天的資料得到track2的public score為0.101685。
3. 第二種加入ID的方式是將ID依照one hot的方式重新編碼，增加623個特徵，在將所有的資料整合再一起，訓練模型。以one hot的方式處理ID的好處是訓練模型是充分的利用大部分的資料。所得的public score為0.107756。
4. 最後，我們嘗試調整C，簡單的看看模型是否有過適（overfitting）的問題[[2]](#footnote-2)。我們使用C = 1、0.1、0.01，發現三者的差異並不大（track 2 的public score分別為0.107756、0.107764、0.107709），模型的過適問題應該不嚴重。

1. 在15天的資料中總共有623個不重複的ID [↑](#footnote-ref-1)
2. 我們使用sklearn的LogisticRegression訓練模型，其中C用於regularization，C越小regularization越強。我們使用邏輯斯迴歸模型幫助我們收息資料和練習訓練模型，因此沒有嚴謹的使用validation，在此僅用public score看模型的過適問題。接下來的模型訓練將會更仔細的使用validation挑選模型。 [↑](#footnote-ref-2)