MACHINE LEARNING-Assignment2

尹健璋

*Department of Statistics*

*National Cheng Kung University*

Tainan, Taiwan

R26111052@gs.ncku.edu.tw

# Introduction

本次作業包含應用kernel方法將資料投射至高維空間後，使用不同模型去建構集成式學習之演算法以及建構以2-layers MLP模型取代決策樹模型之隨機森林演算法並與原始隨機森林演算法的性能做比較。

# Data

本次作業的data來自kaggle公開資料集https://www.kaggle.com/datasets/ifteshanajnin/carinsuranceclaimprediction-classification?select=train.csv : 此為汽車保險索賠的資料集其中包含 58592 筆資料、44 個變項，其中”is\_claim”這個變項將作為預測目標，它包含”0”及”1”兩種類別，由於本次資料數很多要執行所有分類器演算法以及其他演算法需要花費很多時間，因此本次作業僅從資料集中抽取25000筆資料來使用，抽取方式為了防止隨機抽取會使抽取到的資料的預測目標類別比例不同，因此隨機抽取時，已取出不放回的方式依照原本預測目標各類別的數量乘以25000/59582並四捨五入至整數的數量去抽取。

預處理部分先將”police\_id”刪除，在Linear classifier 中，有兩種類別的類別變型特徵會使用one-hot encoding 轉換成連續型，而兩個類別以上的類別型特徵則使用frequential encoding做轉換並與連續型特徵一同做標準化；而在手刻的Decision Tree中，則會把所有連續變數轉換為兩類別的類別變數，方法為觀察切在不同值的cross entropy並選出cross entropy最小的切點來對連續型特徵進行分類，而在決策樹的連續型特徵轉換為類別型特徵會在訓練模型時進行。

# Application of Kernel Methods in Ensemble Learning Across Various Learners

本題使用 Radial Basis Function Kernel 針對所選的四個特徵來進行計算將特徵空間擴展至高維度空間來進行後續的建模，本次使用 Neural Network Model 以及五個 Decision Tree Model 來進行整合預測整合，而整合方式依照各自於驗證集的預測準確率去計算個模型權重，再讓總計六個模型針對測試集去預測並透過權重來計算最後的預測結果並輸出整合後的預測準確率。

選取特徵來做 Radial Basis Function Kernel 部分，針對連續變數去計算相關性並進行排序後挑選最相關的四個變數來挑選，相關性表格如下：

1. 連續變數相關性排序表

|  |  |  |
| --- | --- | --- |
| colname1 | colname2 | abs(correlation) |
| displacement | length | 0.9617 |
| turning\_radius | length | 0.9449 |
| length | width | 0.9159 |
| displacement | width | 0.8993 |

透過上表可以發現這四種特徵之間互相都是高度相關性，因此使用這四種特徵去做kernel 方法來擴展特徵空間建立新的特徵集。

另外，關於神經網絡模型設定訓練的迭代次數並觀察其 loss 是否有下降趨勢，如下圖：

1. 神經網絡 loss 圖

一張含有 文字, 螢幕擷取畫面, 圖表, 行 的圖片

自動產生的描述

依照此圖可以確定我們的NN模型是有訓練至收斂的。

再來將新的特徵集切分成訓練集、驗證集與測試集並將訓練集放入一個神經網絡模型與五個決策樹模型去做訓練並使用驗證集去預測且計算準確率，再依照各模型對驗證集的準確率去計算權重來確定哪個模型結果比較可信，如下表：

1. 各模型權重

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | NN | Tree0 | Tree1 | Tree2 | Tree3 | Tree4 |
| 權重 | 0.468 | 0.1365 | 0.065 | 0.206 | 0.0185 | 0.119 |

最後依照權重去整合各模型的預測結果做投票作為最終的結果，而依照此方法對測試集的預測準確率為0.9378。

# Ensemble of Deep Learning-Based Non-Tree Weak Learners

此題為比較隨機森林模型與將隨機森林中的決策樹改成2-layers MLP模型性能是否有提升，首先兩種模型皆依照相同的特徵個數和模型數來進行訓練與預測，設定的特徵數為40、模型數為5，而決策樹的部分設定深度為10、MLP的隱藏層數為10，而在MLP部分迭代次數為500、更新率設定為0.00001，隨機森林對測試集的預測結果為準確率0.2879，而替換成MLP對測試集的預測結果為準確率0.9378，從結果來看替換後的模型性能有顯著的提升，但準確率差異過大，所以初步判斷有這樣的原因可能為單一決策樹模型性能不如MLP模型來的好，在相同子模型個數下預測結果會差異很大，因此標準的隨機森林模型可能需要建立更多的子決策樹模型來提升預測性能。