Machine Learning - Assignment II

All about supervised learning

朱信榮  
*Department of Statistics,  
 National Cheng Kung University  
Tainan, Taiwan  
morris70121@gmail.com*

# Introduction

本次作業內容包含第一部分將kernel method應用在ensemble learning，根據選擇使用的集成式學習方式進行訓練與預測後做出結論，第二部分則是將隨機森林中的決策樹部分替換成神經網路，也就是架構為隨機森林但是內部為使用神經網路進行訓練，並且比較與原本使用決策樹的隨機森林結果與作出結論。

# DATA & preprocessing

本次作業使用的資料集為來自kaggle的公開資料集<https://www.kaggle.com/datasets/ifteshanajnin/carinsuranceclaimprediction-classification/data>

該資料及總共具有58592筆資料、44個變數，並且以其中的“is\_claim”變數作為最後準備預測的目標，該變數為包含”0”及”1”兩種類別。

在進行訓練前，首先針對數值型部分資料並且數值範圍太大的變數進行標準化（包括變數如下："population\_density","displacement","length","width","height","gross\_weight"），之後再針對類別行資料做處理，針對類別行資料的處理為使用one-hot encoding 轉換成連續型資料。

# Application of Kernel Methods in Ensemble Learning Across Various **Learners**

根據題目給定的Hint 1，這邊使用的集成式學習中包含了兩層深度的神經網路（2-layer NN）與使用隨機部分特徵的決策樹，並且將訓練資料先進行kernel method選取兩個特徵進行運算增加新特徵加入到原本資料中（使用RBF kernel將兩特徵進行運算合成一個新特徵，即投射到高維度）

* RBF kernel :

為兩特徵。

設定將一次神經網路模型與五次決策樹模型訓練的結果進行validation set資料判斷各自六個模型的validation accuracy當做最後集成式學習輸出結果的權重，下圖為兩層NN的訓練loss結果：

一張含有 行, 繪圖, 圖表, 螢幕擷取畫面 的圖片

自動產生的描述

最後根據各自模型輸出結果與權重，例如第一筆測驗資料各自模型輸出結果為1,1,0,0,1,1，先將0,1分類後，利用各自的模型權重計算判斷最後輸出結果為0或是1。最後計算出來的結果，測試集的accuracy為0.9377880184331797。

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

利用第一部分的隨機森林架構，將裡面的決策樹部分替換成兩層神經網路（2-layer NN），並且比較輸出結果差異做出結論，我們可以得到兩者分別的accuracy為(Original random tree) 0.28793309438470727 (Random tree with 2-layer NN) 0.9377880184331797

##### Conclusion

首先由於資料label的不平衡（大部分是0），因此模型在進行預測的時候會很容易預測成0導致最後預測準確率很高，但其實模型並沒有學到東西。另外第二部分的將隨機森林中決策樹部分替換成NN的結果相較於原本隨機森林的結果為家，這部分有可能是因為原本隨機森林為使用部分特徵的資料而已，那可能因為這部分的特徵並沒辦法很好的解釋與預測結果，以及原本資料的不平衡性存在，因此導致結果會不好，但是替換成NN 的及果其實也只是都預測成0而以，並非有真正學習到東西，因此可能無法很好的進行模型之間的比較與分析，這個部分可能會需要後續進行改進。

##### References

1. https://www.kaggle.com/code/dapy15/mlp-from-scratch

**Github : https://github.com/MorrissChu/ML--Assignment-II-2023-12.gi**