HW2

Neural network、Decision Tree 及 K-means 三種方法應用於紅酒品質之比較

1. 資料前處理

在資料科學中有非常多方法對分布不均勻的資料進行修正，例如：用於修正偏態資料的平方根轉換、僅能用在偏左資料上的次方轉換或是本次作業用到的方法-離群值剔除。

1.1 剔除離群值

一般而言， 資料若位於四分位距外就會被合理認為是離群值，但鑒於不同資料間的特性，實作上將離群值的判斷修正為四分位距乘上權重係數 n，以利離群值數量的控制。

1.1.1 刪除距離中位數較遠的值

高斯分布是資料常見的分布形式，因此基於這個原因，通常在資料前處理時會刪除距離中位數較遠的值，因為這些離群的資料會使模型的訓練往特定一邊傾斜，不利於將來模型的預測。

1.1.2 刪除距離中位數較遠的值

然而有些特徵彼此數值之間離散程度較大，此時就不適合刪除距離中位數較遠的值，因為對於該筆特徵而言，距離中位數較近的資料反而才是離群值，雖然這種資料不會使模型訓練後權重偏向特定一邊，但這種資料仍舊會因為特徵仍過於纏結，使訓練不容易進行。

1.1.3 修正方案

鑒於不同特徵之間分布的形式皆不盡相同，因此應該在進行離群值處理前，先利用資料視覺化或是資料特徵的標準差、四分位距或是中位數，判斷該特徵的分布形式，再根據判斷的結果妥善分配離群值的修正方式，而不是每個特徵皆採用刪除距離中位數較近的值或是刪除距離中位數較遠的值。

2. 三種方法之間的比較

在訓練模型時，Neural network 具有相較後兩者更多的彈性，Neural network 有更多的參數可以調控，不僅可以調整層數和通道數量，也可以調整不同激活函數選擇最佳的非線性方式，這些特性使得 Neural network 有著相較於後兩者更好的解纏結特性，同時 Neural network 也更大程度地使用到特徵之間的關係，因此相較於模型的調整，我更認為決定 Neural network 預測準確度的是資料優劣，因此我較多著墨於資料的前處理。然而許多課程都認為學生國慶日都沒事，指派分量過多的作業，使得學生必須妥善分配不同作業之間的時間，無法在同一個實驗上花費較多的時間做研究，沒有足夠時間分析離群值剔除對於同個模型的影響，實屬可惜。然而 Neural network 的缺點也非常明顯，有著三者中最長的訓練時間，

不過今天紅酒品質檢測是較小的資料集，因此影響不大。

而 Decision Tree 及 K-means 能調整的參數較少，例如 K-means 最大的調整方向就是索引的數量，訓練時也較多著墨在索引大小和精確度的影響，而不是訓練集輸入的方式。同時 Decision Tree 和 K-mean 都屬於縣性的分類器，因此解纏結程度較不如 Neural network。

3. 第四種方法

我本來是想標新立異，透過輸入雜訊獲得準確率當作標籤，設計一近似 GAN 的模型架構，但由於雜訊輸入後得到的標籤都在 50%上下，因此模型非常難收斂因此作罷，不過我仍在附件上附上程式以供參考。