函數的繪製

林哲兆

January 10, 2024

前面的章節介紹了多種學習器，雖然它們的原理各不相同，但其共同目標是將樣本進行正確分類。因此，在這一章節中，我們將運用相同的資料集來訓練不同的學習器，並比較它們在相同資料下的訓練成果，找出優劣之處。為了確保公平的抽樣，本文將採用蒙地卡羅模擬進行重複抽樣，以最小化抽樣所產生的誤差。

1 蒙地卡羅模擬

蒙地卡羅模擬是一種數學的計算方法，主要用於模擬和分析不確定性事件，這種方法依靠隨機抽樣和統計推斷來解決問題，特別適用於沒辦法準確預測結果的複雜系統。蒙地卡羅方法通常分為三個步驟：

- 第一步　通過隨機抽樣生成大量可能的輸入值
- 第二步　將這些輸入值放入模型進行運算與模擬
- 第三步　通過統計分析處理這些結果，以做出統計推論

蒙地卡羅模擬在金融、工程、物理學、生物學等領域都有其應用，因為它能夠處理高度複雜和不確定的問題，所以受到很大青睞。

2 模型比較

本小節將使用兩筆資料進行模型訓練，而每次訓練都會重新分割訓練資料與測試資料，達到重複抽樣的目的。最後，將資料放進六個選定的模型進行比較。

2.1 亂數產生資料

在樣本設定為 200 下，分別生成兩個連續型變數的特徵與一個類別變數，以下是常態亂數的參數設置與資料散佈圖：

$$\mu_0 = \begin{bmatrix} -2 \\ 3 \end{bmatrix} \mu_1 = \begin{bmatrix} -1 \\ -2 \end{bmatrix} \Sigma_0 = \begin{bmatrix} 15 & 1.2 \\ 1.2 & 20 \end{bmatrix} \Sigma_1 = \begin{bmatrix} 5 & 2 \\ 2 & 18 \end{bmatrix}$$

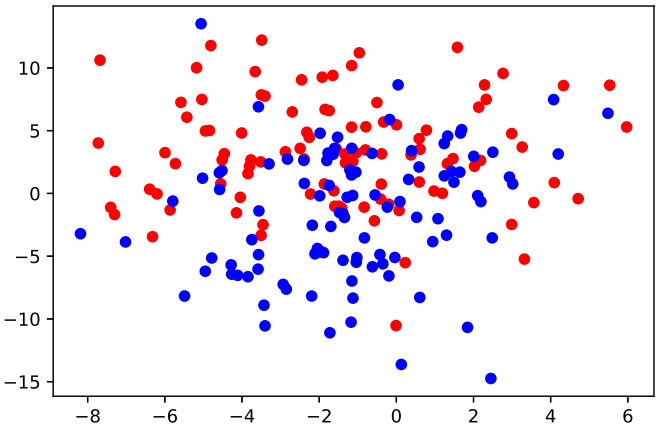

圖 1: 常態亂數資料

在進行重複抽取訓練資料 (70%)100 次後，將資料放入分別的七個模型：LDA、QDA、Logistic Regression、KNN(K=5)、KNN(K=15)、ANN(神經元數 10)、ANN(神經元數 20)，進行模型配適，最後得到的結果如下：

表 1: 七種模型在亂數資料的誤判率

模型	**LDA**	**QDA**	**Logistic**	**KNN(K=5)**	**KNN(K=15)**
誤判率	0.2173	0.218	0.2052	0.2175	0.213
模型	**ANN**(神經元數 10)		**ANN**(神經元數 20)		
誤判率	0.2082		0.2502		

在這筆資料中，七種模型的差異不大，除了 ANN(神經元數 20) 模型的誤判率比較大之外，其他誤判率都介於 0.2 ∼ 0.22 之間。

2.2 消費者資料

第二筆資料是來自 kaggle 的一份消費者購買行為的分類資料，其特徵包括兩個：觀測者的年齡與薪資，類別則是其購買與否 (0、1)，樣本數為 400，以下為資料散佈圖：

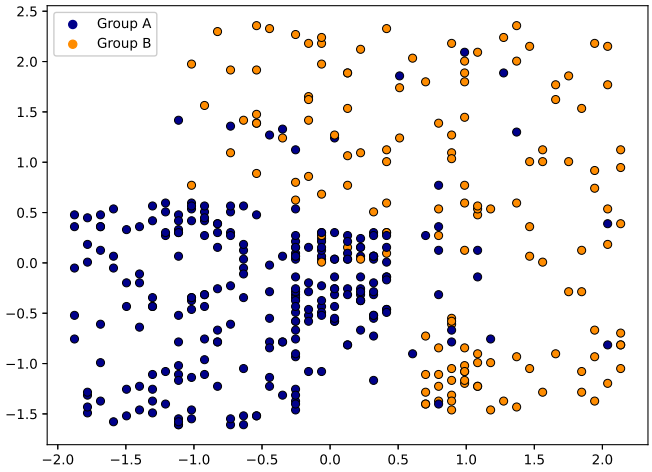

圖 2: 消費者購買行為資料

同樣重複抽取訓練資料 (70%) 100 次後，將資料放入分別的七個模型，最後得到的結果如下：

表 2: 七種模型在消費者資料的誤判率

模型	**LDA**	**QDA**	**Logistic**	**KNN(K=5)**	**KNN(K=15)**
誤判率	0.1528	0.1517	0.0988	0.0789	0.0884
模型	**ANN**(神經元數 10)		**ANN**(神經元數 20)		
誤判率	0.1232		0.1116		

從表 2 可以看出在這筆資料中誤判率有些差異，KNN(K=5) 的誤判率是最低的，LDA 則是最高，特別的是這筆資料的執行時間從上一筆的 5 分鐘變成 20 分鐘，可知樣本數對運算時間有很巨大的影響，在樣本增加 2 倍時，時間增加了 4 倍。