系級：電機系 2E
姓名：張峻瑋
學號：110511194

機器學習導論 作業 2：Classification

Part I.

Consider that there are a group of college basketball players, who come from four different high schools. Now we analyze the Performance-Rating of “skill” and the athleticism” for every player, as shown in Figure 1, where the blue, green, red, and black dots represent the players from High school 1, High school 2, High school 3, and High school 4, respectively. Noted that there are 400, 250, 200, and 150 players from High school 1, High school 2, High school 3, and High school 4, respectively.

You are given a HW2.xlsx file. Please implement the algorithms of the generative model and the discriminative model to classify the data and plot the corresponding decision boundaries, like Figure 2.

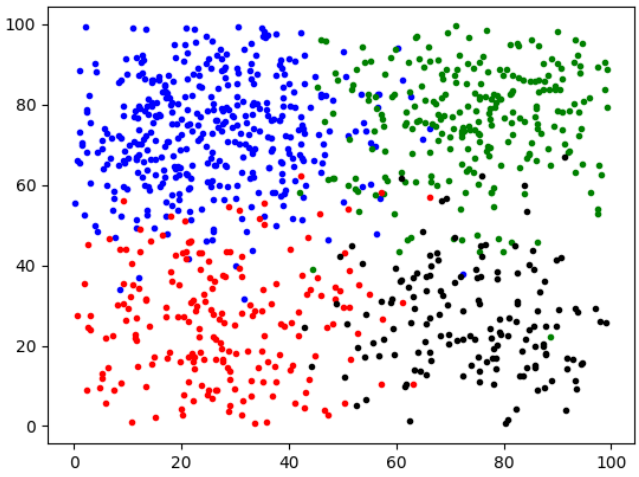

圖 1：4 分類下的資料點分佈

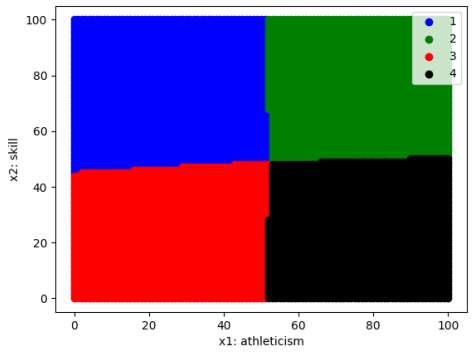

圖 2：generative model 之結果

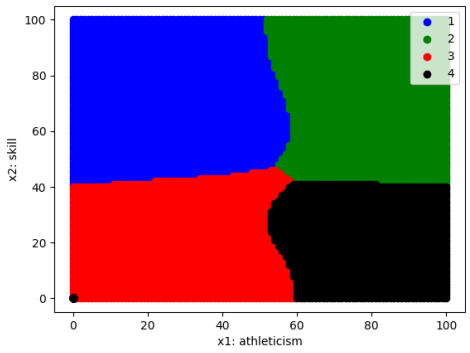

圖 3：discriminative model 之 IRLS 5 次結果

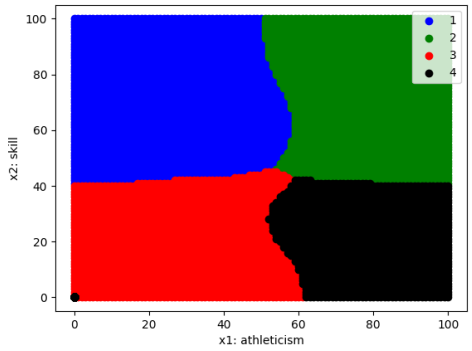

圖 4：discriminative model 之 IRLS 10 次結果

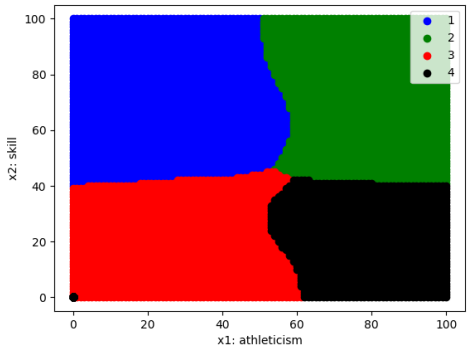

圖 5：discriminative model 之 IRLS 15 次結果

結果分析：

此部分分成 2 個小題，第 1 小題使用 generative model 進行分類，而第 2 小題使用 discriminative model 進行分類。

使用 generative model 時，這裡假設他是 4 個高斯分佈所產生的結果。首先先計算 4 個組別的各別平均值，接著假定這 4 個高斯分佈的共變異數矩陣是一樣的，去計算出該矩陣。然後，藉由共變異數矩陣與平均值分別計算出 $\mathbf{w_k}$ 與 w_{k0}。$\mathbf{w_k}$ 的轉置矩陣乘上各資料點加上 w_{k0} 即可得各點之 a_k，a_k 有點類似某點在第 k 類的機率，是故我們可以依據 a_k 得到每一點最可能是第幾類。依序著色即可得圖 2。

Discriminative model 則不假定資料分佈的可能性，屆由調整權重的方式得到最要的權重，依此進行分類。首先我們先設定一個 basis function，這裡採用 logistic sigmoid function，並令 M = 5。將已知資料點丟進 basis function 後得到一設計矩陣。由於權重 w 是可以不斷調整的，這裡我們先設全部皆為 1。

在調整參數的過程中，我們先用同 generative model 的方式求得 a_k，進而求得 y_k。有了 y_k 後，用(y-t)乘上Φ求得 error function 對 w 的梯度，我們的目標是使此梯度趨近於 0。調整 w 的過程中，需求得黑塞矩陣。新的 w 即為舊的 w 減去黑塞矩陣的反矩陣乘上誤差函式的梯度。而黑塞矩陣可透過 $\Phi^T R \Phi$ 求得，R

為 y(1-y)所構成的對角矩陣。

當調整 w 直到梯度很小時，即可將該 w 運用於資料點上，求得每一點為特定組別的機率，而畫出所求圖形。

Part II.

According to the overall Performance-Rating of “skill” and the “athleticism” for each high school in Figure 1, we classify High school 2 to be class A, High school 1 and High school 4 to be class B, and High school 3 to be class C, as illustrated in green, blue, and red dots, respectively in Figure 3. Please classify the data in Figure 3 for the same requirement in Part I.

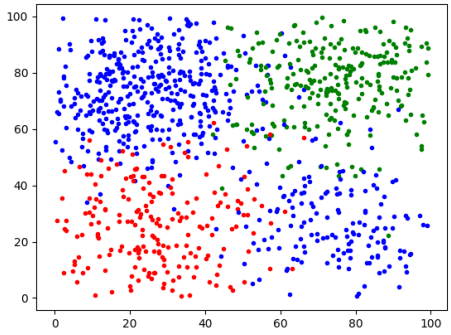

圖 6：3 分類下的資料點分佈

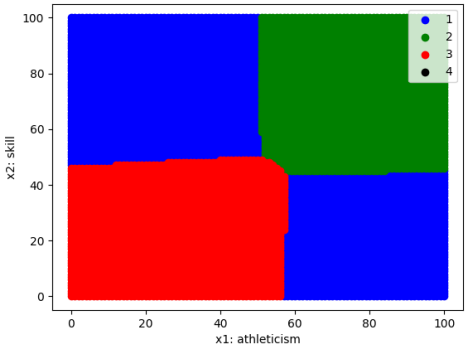

圖 7：generative model 之結果

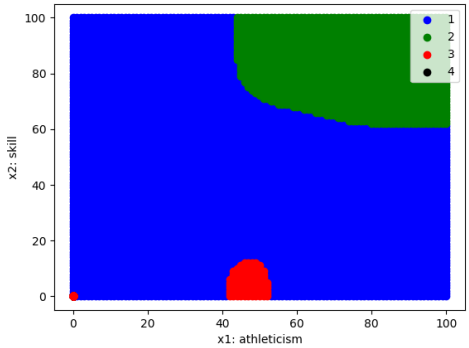

圖 8：discriminative model 之 IRLS 5 次結果

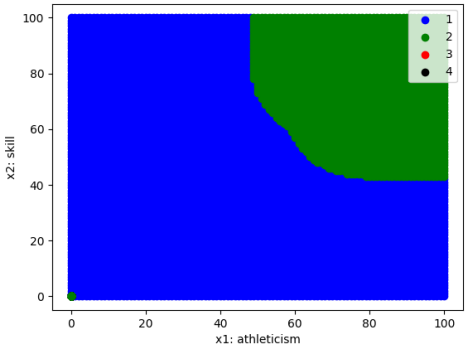

圖 9：discriminative model 之 IRLS 10 次結果

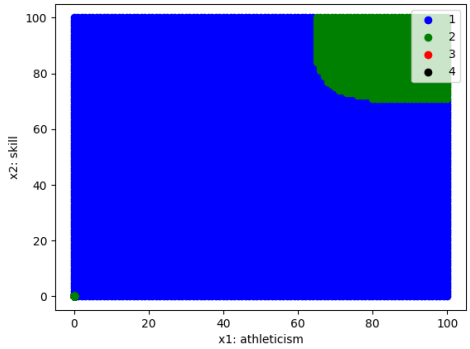

圖 10：discriminative model 之 IRLS 15 次結果

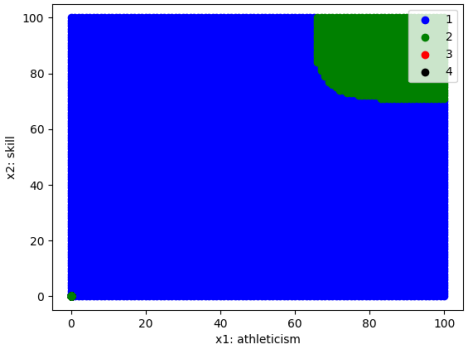

圖 11：discriminative model 之 IRLS 20 次結果

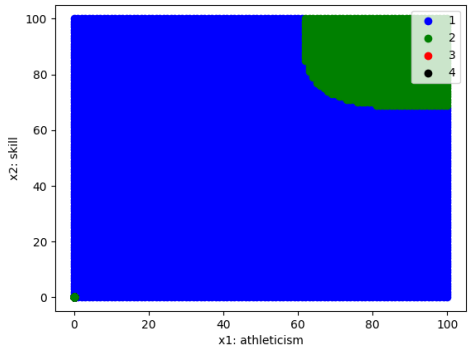

圖 12：discriminative model 之 IRLS 25 次結果

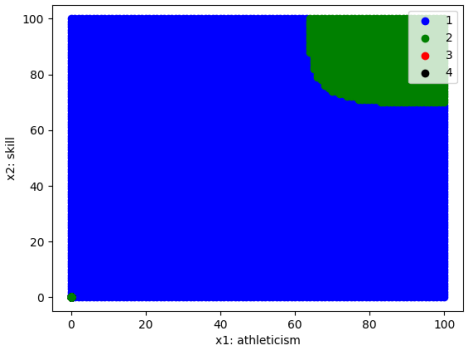

圖 13：discriminative model 之 IRLS 30 次結果

結果分析：

使用 generative model 時，同第一部分，只是將第一組與第四組做 mixture Gaussian，所得出的結果。

使用 discriminative model 時，由於沒有假定其分佈，故直接將資料點丟進去跑同第一部分的函式。結果如圖 8 到圖 13。