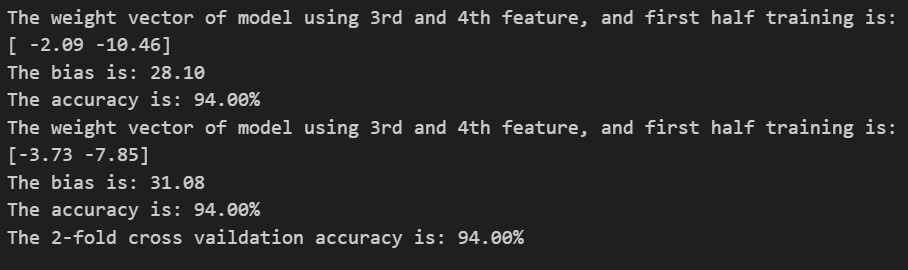
Question No. 2



Step 2: 將變色鳶尾設為正類別(Positive class)，維吉尼亞鳶尾設為負類別 (Negative class)，採用第三種和第四種特徵，並且以兩類別前25筆數據做為訓練，後25筆共50組數據做為training set

Weight Vector: W = [ -2.09 -10.46]^T

Bias: b=28.10

Step 3: 模型分類率94.00%

Step 4: 將訓練和測試資料集對調，進行交叉驗證(Cross validation)，對調後結果如下：

Weight Vector : W = [ -3.73 -7.85]^T

Bias: b=31.08

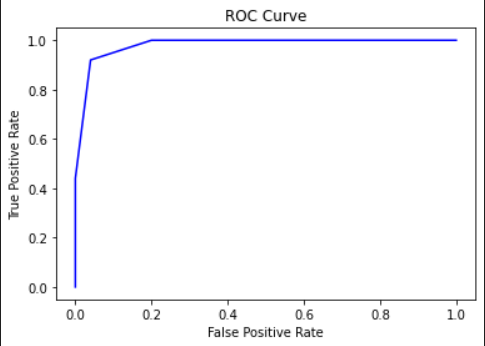
分類率94.00%

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 資料組合 | Class | Weight vector | Bias | Accuracy |
| 2,3 | 2,3 | [ -2.09 10.46] | 28.10 | 94.0% |
| 2,3 | 2,3 | [ -3.73 -7.85] | 31.08 | 94.0% |
| Avg.accuracy | | | 94.00% | |

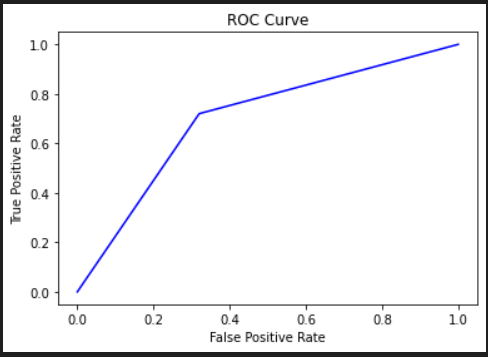
註：第一次為使用前25筆資料做訓練，後25筆做為測試集；第二次則相反。

Question No.3

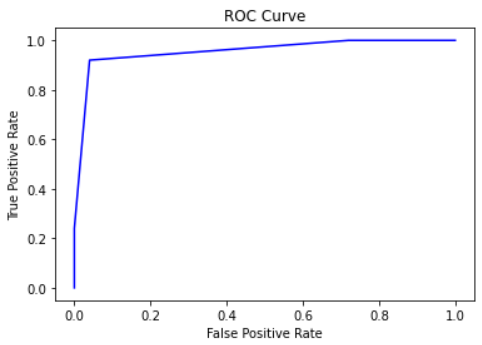
Step1: 使用四種特徵的ROC curve



Step2. 使用前兩種特徵的ROC curve



Step3. 使用後兩個特徵的ROC curve 以及AUC value



Question 3 AUC value form:

|  |  |
| --- | --- |
| 使用特徵 | AUC value |
| 全部 | 0.9808 |
| 1,2 | 0.7000 |
| 3,4 | 0.9560 |

Step 4討論：

若整體ROC曲線越靠近左上角(1,0)這個點，表示AUC值越靠近1，意味著分類器的性能越佳(理想值為1.0但幾乎不可能達到)，在三個不同的模型內，使用全部4個特徵的有著最高的AUC值，符合預期(有越多線索我們越能清楚分類兩種不同的植物)，但只使用前兩個特徵的AUC值明顯和使用後兩個特徵、使用全部特徵的AUC有一段差距，這意味著如果要簡化計算，僅使用兩個特徵來分類時，使用第三和第四個可能會是能夠接受的，畢竟分類精確度和計算的繁雜程度會呈正相關。

Question4.one-against-one strategy :

|  |  |
| --- | --- |
| 組合 | 分類率 |
| 每種前25筆當測試；  後25筆作為訓練資料集 | 96.0% |
| 每種後25筆當測試資料；  前25筆作為訓練資料集 | 96.0% |
| 平均 | 96.0% |

訓練資料size:

|  |  |
| --- | --- |
| 資料集類別 | Size |
| 正類別資料 | 25, 2(Only feature) |
| 負類別資料 | 25, 2(Only feature) |
| 測試資料集 | 75, 2+75,1(Label) |

其他討論：

由近兩次的作業發現，只要決策過程中出現「投票」決策的演算法，對於irisdata的分類率都相當高，但由於只有分類單一個數據組，並沒有嘗試進行其他數據的測試，可能暫時無法得到究竟是「凡有投票的正確率都會高」還是單純「這組數據在投票表決的演算法中可以被精確地分類」，可能還要以不同數據組測試，才能得到較明確的答案。(也有可能這是一組非常適合二分法的數據資料)