Machine Learning HW2

312512005 黃名諄

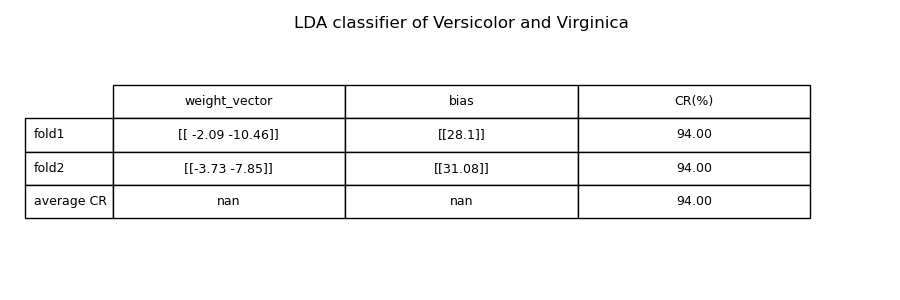
1. 2-fold cross validation : Iris data set的變色鳶尾（Versicolor）以及維吉尼亞鳶尾（Virginica）分別設為positive class與negative class，每個class皆採用第三種與第四種特徵

表 1

1. ROC and AUC:
2. 4 features:

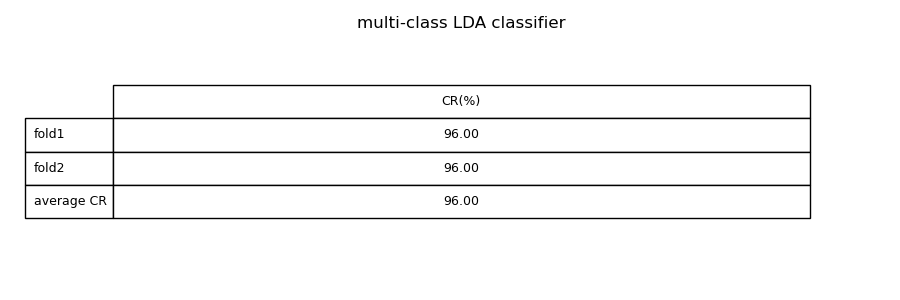
|  |
| --- |
| AUC : 0.9946 |
| 圖1 |

1. 1st & 2nd features :

|  |
| --- |
| AUC : 0.7622 |
| 圖2 |

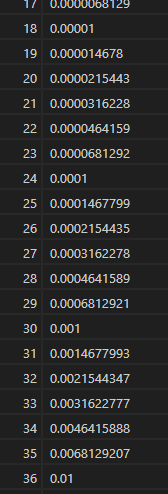
1. 3rd & 4th features :

|  |
| --- |
| AUC : 0.9910 |
| 圖3 |

1. Multiclassification -- One against one strategy:

圖表 2

1. 結果討論:

* 在part1的實作LDA兩類別分類中，模型原理部分和上課所學相同，而對於在decision boundary(D=0)上的資料則當作預測錯誤不歸為任何一種class來處理，此模型表現不錯，有94%分類率，而對照HW1的散佈圖，Versicolor 和Virginica的分佈交界也有些混雜，因此線性分析難以達到100%分類率。
* ROC部分，對於每種情況，調整比值從10-8 ~ 108 取對數分布共97個值(每10倍間有4個數)，如下取其中一小段所示:

從ROC及AUC結果可看出，圖一及圖三AUC較大，圖二AUC較小，而在ROC中最左上角，(0,1)的點FPR=0，代表所有negative class沒有被錯誤判斷為positive的，也就是沒有偽陽性；TPR=1代表所有positive class皆被正確判為positive，也就是沒有偽陰性，兩類別都被正確預測，因此(0,1)的點代表了一個最完美的預測結果，也因此AUC最大值為1，當AUC越大，表示模型整體ROC越靠左上角，可推論在當前使用的特徵情況下，變動C1/C2比值所得到的模型整體上能較靠近完美預測而有效分類，像是圖一使用4個特徵時及圖三使用第三和第四個特徵時，AUC都接近1，表示在選用這兩種特徵組合下，模型的預測能力會較接近完美預測，而另一方面圖二使用第一和第二個特徵時，ROC曲線明顯較沒有靠近左上方，AUC也因此較低，表使用這兩個特徵時整體上離完美預測較遠，分類效果可能較差。將這些結果對照上次HW1的結果，使用第一和第二個特徵時，兩個類別的分界非常模糊，分類率也因此很差，和AUC的推測呼應，而使用到第三和第四個特徵時，兩類別的分界較為清楚，分類率較高，也和AUC有相同的推論結果。因此可得到結論，AUC大小可對應到使用不同特徵下模型的分類效果，AUC越大其整體可能越接近完美分類，AUC越小則表示分類可能較差，因此由前所述，AUC是可以用來比較不同特徵組合下模型好壞的量化工具來做使用。

* ROC中，三種情況下都在比值接近1時較靠近左上角完美預測點，可推論在各特徵組合中，總體上比值在1附近會得到最好的預測效果，而如老師上課說的一樣，有特別需求想判對哪一類時也可參考此曲線去做選擇。
* 多類別分類主要架構還是建立在part1兩類別分類器上，將part1的分類器分別建立one against one strategy兩兩一組的model共3個，比起兩類別分類，此時test需經由3個模型預測結果最後以多數投票的方式得到最終預測，同票則設為分類錯誤，在這樣的條件下，test data是經由3個模型預測共同決定結果，更具泛用性，較不會因單一模型的缺陷而預測錯誤，例如某一模型overfitting等，因此可推論預測正確的機率比起單兩類別應該更高，結果也表明其分類率高達96%，比part1單兩類別有更好的表現。