

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

LDA classifier of Versicolor and Virginica

	weight_vector	bias	CR(%)
fold1	[[-2.09 -10.46]]	[[28.1]]	94.00
fold2	[[-3.73 -7.85]]	[[31.08]]	94.00
average CR	nan	nan	94.00

表 1

2. ROC and AUC:

- I. 4 features:

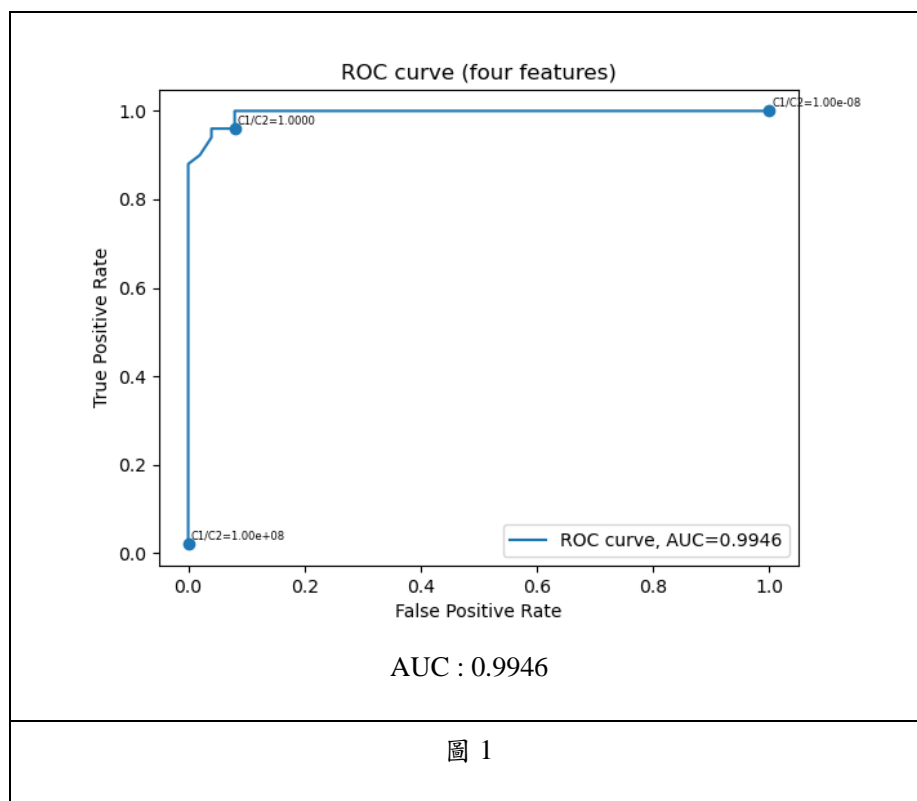
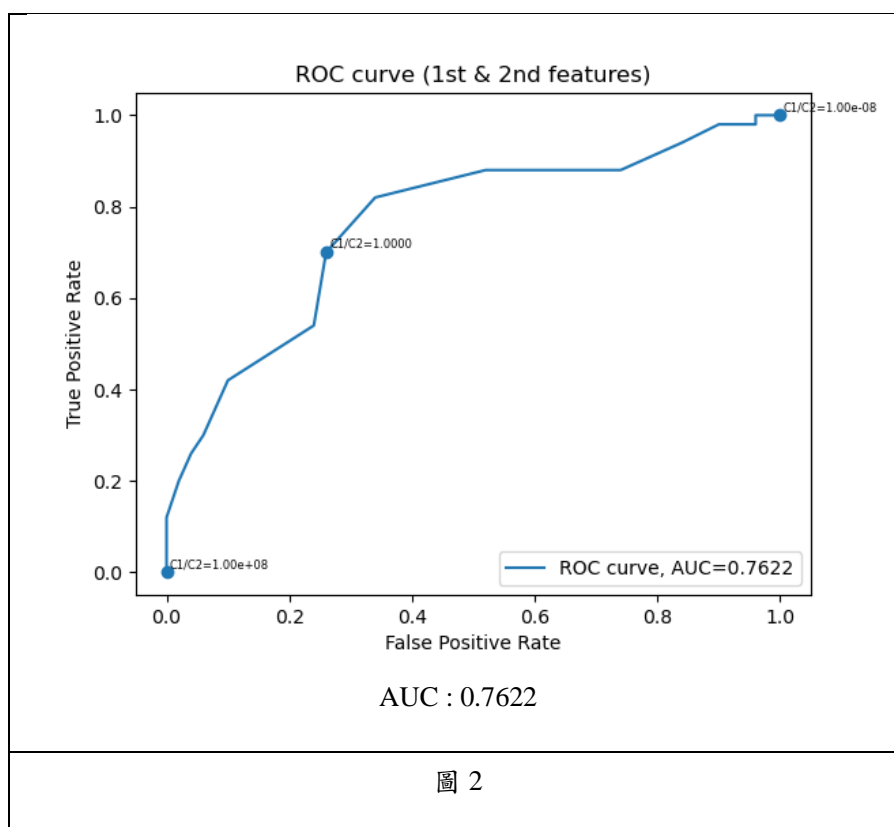
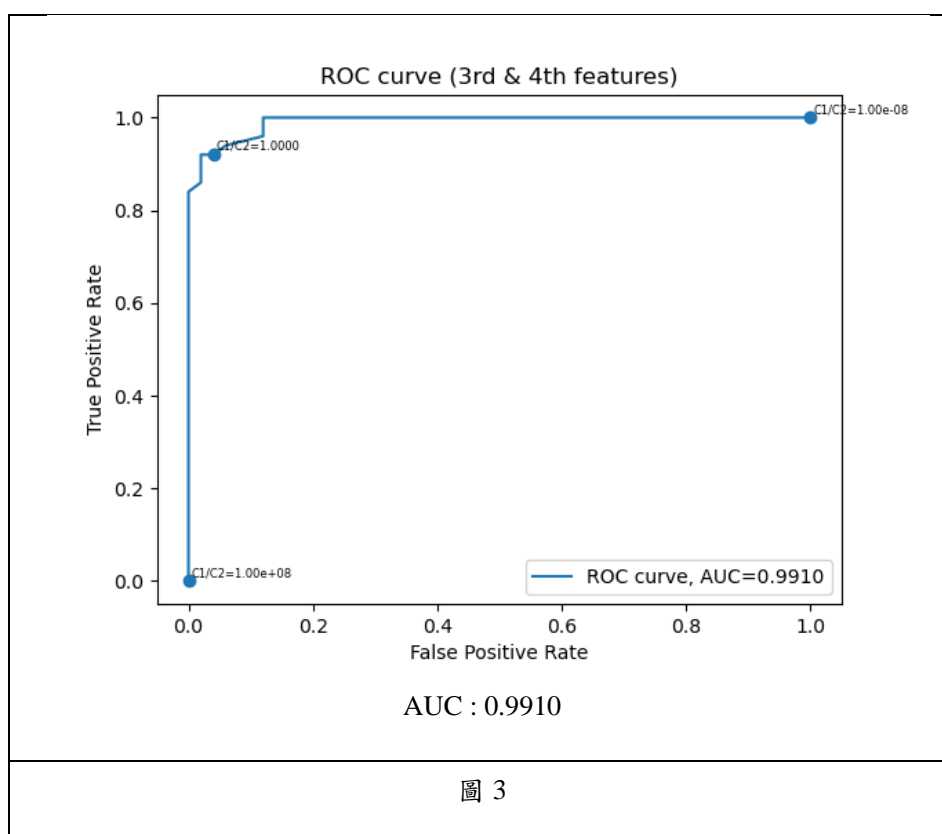


圖 1

II. 1st & 2nd features :



III. 3rd & 4th features :



3. Multiclassification -- One against one strategy:

multi-class LDA classifier

	CR(%)
fold1	96.00
fold2	96.00
average CR	96.00

圖表 2

4. 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 $10^{-8} \sim 10^8$ 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:

```
np.logspace(-8, 8, num=97)
```

```
17 0.0000068129
18 0.00001
19 0.000014678
20 0.0000215443
21 0.0000316228
22 0.0000464159
23 0.0000681292
24 0.0001
25 0.0001467799
26 0.0002154435
27 0.0003162278
28 0.0004641589
29 0.0006812921
30 0.001
31 0.0014677993
32 0.0021544347
33 0.0031622777
34 0.0046415888
35 0.0068129207
36 0.01
```

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。