

# 圖形識別

作業三:SVM

曾宏鈞 06160485

資料:藉由信件內容, 判斷是否為垃圾郵件

## 步驟

1. 原始email資料:txt檔
2. 刪除標題等字詞清除數據
3. 建立詞頻查詢表 (考慮最常用3000個)
4. 將資料轉換成詞頻向量

訓練集 (702筆)

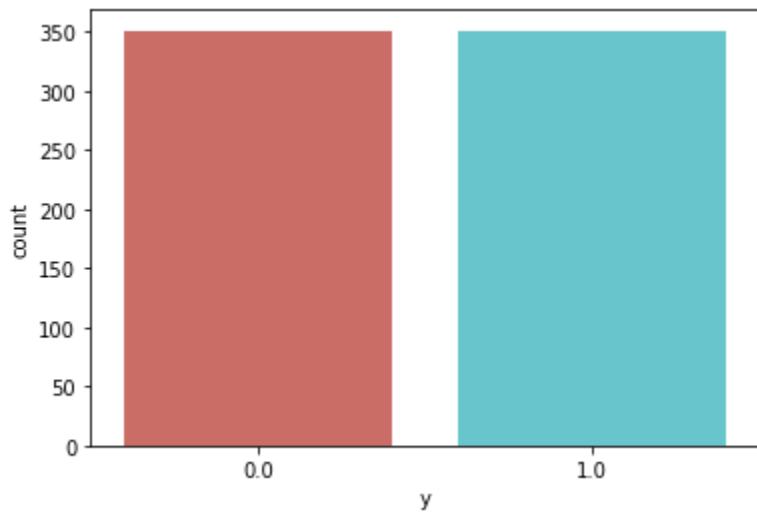
- $X(\text{feature})$  : 3000維
- $Y(1|0)$

測試集 (260筆)

- $X(\text{feature})$  : 3000維
- $Y(1|0)$

# 作業

- 練習1 (kernel="linear", C = 1, gamma = 1) : 0.95
- 練習2, 與貝式時間相比, 貝氏訓練時間為0.037、SVM為0.138, 訓練時間為model.fit的執行時間, 貝氏只需求先驗機率、SVM需要找決策邊界
- 練習3, 準確度: 貝氏的準確度為0.96, 比SVM高0.1
- 練習4 只用原本1/10資料集做訓練, 訓練時間從原本0.138下降到0.019相當於差了7倍
- 練習5, 練習6:



- 由上圖可知, 在訓練集的兩類資料回平衡的資料, 不需做資料平衡。

## 微調模型

我們使用以下參數對模型進行微調：

- Kernel: rbf, linear, sigmoid
- C: 10, 100, 1000, 10000
- Gamma: 0.1, 0.01, 0.001, 1, 10, 100

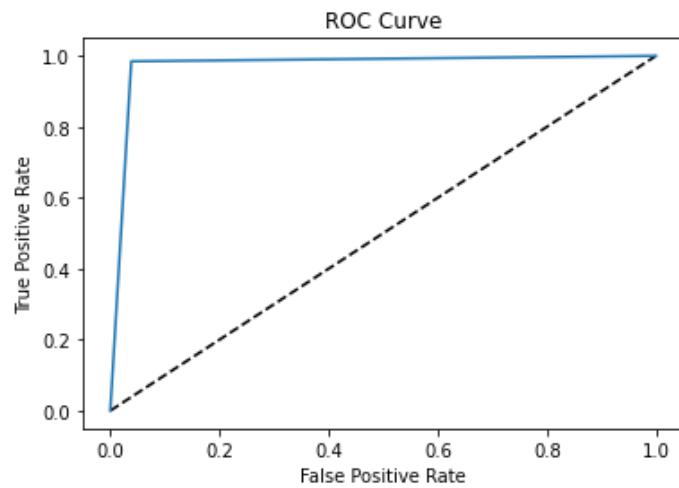
得出最佳準確度為0.973, 而他的最佳參數為 C=100, gamma=0.001, 其中kernel為 使用kernel trick的核函數、C為正規化的參數, 值愈大正規化程度越小。gamma為kernel的核心係數, 下圖為最佳模型的 report及ROC, 由於他的AUC score同時也為0.813, 因此此模型的效能非常好。

precision recall f1-score support

0.0	0.98	0.96	0.97	130
1.0	0.96	0.98	0.97	130

accuracy		0.97		260
macro avg	0.97	0.97	0.97	260

weighted avg    0.97    0.97    0.97    260



根據模型的核心參數gamma與正規化參數去做探討，模型在C=100時準確度最高為0.976，值愈高準確度越低(練習5)，而gamma為0.001時準確度最高為0.93，離0.001越遠準確度越低。(練習6)

