以 Python 進行 SVM 與 Naive Bayes 進行文件分類預測及分類績效評估

B10423004 蔡宗穎，國立雲林科技大學

B10423013 曾鈺雯，國立雲林科技大學

B10423044 王彥淇，國立雲林科技大學

摘要

本團隊使用支援向量機（Support Vector Machines）及貝氏分類器（Naive Bayes）針對文件進行分類，並比較兩種分類器預測結果及分類績效評估，針對新聞的訊息去預測文章分類，並進行比較，SVM 挑選最適當的懲罰值來建構模型，Naive Bayes 挑選最適當的平滑值來建構模型。

演算法：支援向量機（Support Vector Machines）、貝氏分類器（Naive Bayes）

一、緒論

1.1 動機

看到每天人工進行新聞分類是很困難的，所以我們打算讓新聞分類變成自動化的，所以利用 mini_newsgroups 資料集，來進行文本分類預測及評估績效，藉此來模擬股票短期投資的模型。

1.2 目的

利用目前所學的兩種演算法，針對文本進行分類預測及評估績效，並了解該資料集文章與文章之間的關聯性，達成自動分類新聞文章。

二、方法

2.1 程式架構

1. 將檔案解壓縮，並轉碼，方便讀取資料集
2. 打亂原本資料集的順序，藉此來達到每次執行所抓取的資料集不一樣。
3. 對文字出現頻率進行預先處理
4. 經過 k fold 交叉驗證，k 等於 5，來驗證訓練及測試正確率
5. 使用 SVM 套件去建立模型，並找出最佳的 C 值(懲罰值)
6. 使用 Naive Bayes 套件去建立模型，並找出最佳的 alpha 值(平滑值)
7. 比較 SVM 訓練及測試模型正確率並輸出成圖檔
8. 比較 Naive Bayes 訓練及測試模型正確率並輸出成圖檔

2.2 執行程式的方式

透過 Spyder 執行 main.py 檔，使用 python3.6 版本，注意 matplotlib(畫折線圖套件)須為 2.2.3 版，其他需安裝的套件為 sklearn(演算法套件)、nltk(語言分析套件)、numpy(科學計算套件)。

三、實驗

3.1 資料集

使用 mini_newsgroups 資料集是用於文本分類、文本探勘和信息檢索研究的國際標準數據集之一，此資料集被分為 6 個類別，分別為 ：

名稱	筆數	型態
alt.atheism	100	字串
comp.graphics	100	字串
misc.forsale	100	字串
rec.autos	100	字串
sci.crypt	100	字串
talk.politics.guns	100	字串

comp.＊ - 關於電腦的相關話題

sci.＊ - 關於科學方面的討論

rec.＊ - 關於娛樂活動的討論（遊戲、愛好等等）

talk.＊ - 關於社會及宗教熱點話題的討論

misc.＊ - 其他無法歸入現有層級的討論

alt.＊ - 無法歸於其他類別或不願歸入其他類別的話題

3.2 前置處理

利用亂數來打亂原本資料集的順序，藉此來達到每次執行所抓取的資料集不一樣。

TFIDF ： 用來衡量文字字詞出現的頻率，並降低停用詞的權重

$$w_t = tf_t \times \boldsymbol{idf_t} = tf_t \times \log \frac{\boldsymbol{N}}{\boldsymbol{df_t}}$$

名稱	型態	次數
test	int32	45731
zoo	int32	50608
ask	int32	13047
seem	int32	42228
respect	int32	40574

3.3 實驗設計

支援向量機（Support Vector Machines）：是在分類與迴歸分析中分析資料的監督式學習模型與相關的學習演算法。給定一組訓練實體，每個訓練實體被標記為屬於兩個類別中的一個或另一個，SVM 訓練演算法建立一個將新的實體分配給兩個類別之一的模型，使其成為非機率二元線性分類器。

$$f(\mathrm{z}) = sign\left(\sum_{i=1}^{N} \lambda_i y_i \mathrm{x}_\mathrm{i} \cdot \mathrm{z} + b\right)$$

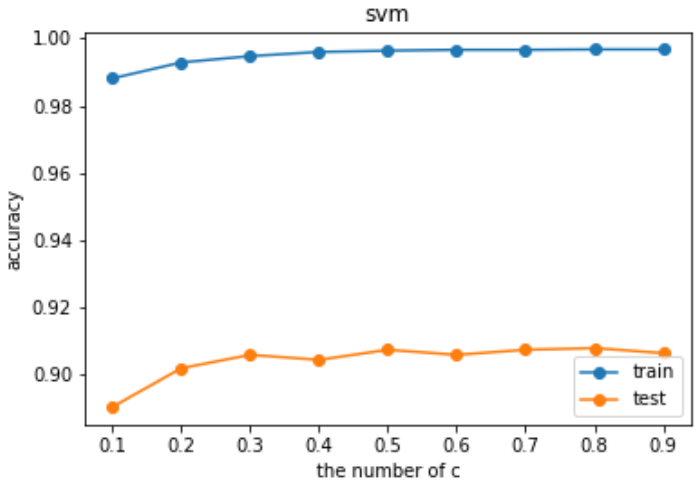
貝氏分類器（Naive Bayes）：是一種構建分類器的簡單方法。該分類器模型會給問題實體分配用特徵值表示的類標籤，類標籤取自有限集合。它不是訓練這種分類器的單一演算法，而是一系列基於相同原理的演算法：所有單純貝氏分類器都假定樣本每個特徵與其他特徵都不相關。

$$C = \arg\max_{C_j} P(C_j) \prod_{i=1\ldots n} P(A_i \mid C_j)$$

3.4 實驗結果

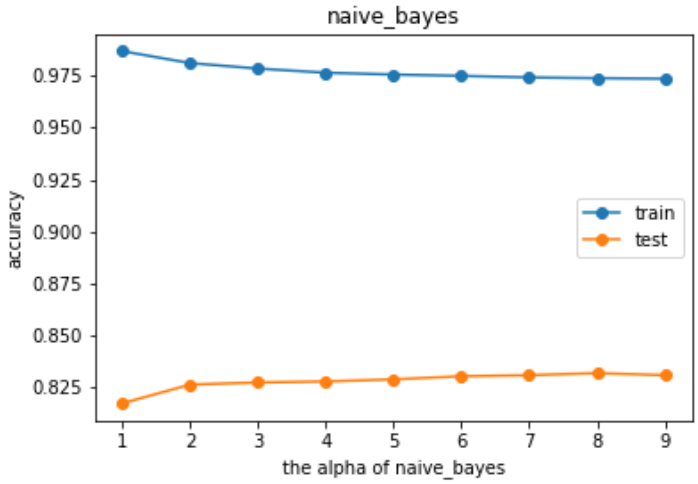
支援向量機（Support Vector Machines）

- 圖表 1.使用支援向量機，比較訓練及測試正確率



貝氏分類器（Naive Bayes）

- 圖表 2.使用貝氏分類器，比較訓練及測試正確率



四、結論

1. 不管是訓練還是測試資料集皆是 SVM 的正確率大於 Naive Bayes。
2. SVM 的最高 C 值(懲罰值)等於 8
3. Naive Bayes 的最高 alpha 值(平滑值)介於 7~8 之間
4. 訓練資料集正確率 SVM = 0.9967、Naive Bayes = 0.9866
5. 測試資料集正確率 SVM = 0.907、Naive Bayes = 0.832