11120CS 554000 Pattern Recognition

台灣股市趨勢分析

Team Members : 109062206 張雅涵、109062211 張惇媛

Abstract—此報告的主題為台灣股市趨勢分析，我們希望從特定股市時間區段內的相關資訊，來預測下一交易日的漲跌趨勢。在實作方法中，分別使用到機器學習之迴歸與分類，但獲得不好的表現，最後進行分析並改善。

I. INTRODUCTION

由於剛開始接觸股市的新手沒辦法根據經驗判斷股票走勢，加上目前股票市場上的技術指標皆複雜難懂，因此我們的目標是從特定股市區間段內的相關資訊，來預測下一交易日的漲跌趨勢，幫助新手在買賣股票時有判斷的依據。

我們將台灣股市聚焦在台股加權指數上，因為台股加權指數具有衡量整體股市表現的指標性，透過分析大盤，能夠了解台灣的股票市場。

因此我們希望設計出一個分類器，可以在輸入想預測的交易日的股市相關資訊後，輸出該交易日的趨勢為漲或是跌。

在這個漲跌的分類問題上我們嘗試了五種不同實作方式，分別是—用 regression 結果做 classification、k-means classifier 、 Perceptron 、 MLP、LDA。接著我們比較實驗結果，試著找出表現最好的分類方式。

II. METHODS

A. Data Collection

透過 python 中奇摩股市的函式庫 yfinance，來獲得台股加權指數之資訊。

我們將 2019 到 2021 年間所有的交易日作為訓練集(共 729 個交易日)，而 2022 年的交易日作為驗證集(共 246 個交易日)。對於每個交易日來說，擁有的五個特徵分別為「開盤價」、「最低價」、「最高價」、「收盤價」、「交易量」。

其中將開盤價與收盤價做比較，如果收盤價大於開盤價，則為漲勢，標籤為 1；其餘的部分，則為跌勢，標籤為 0，並且利用 datafram.shift，取得前 n 天的所有特徵(n = 1 ~ 10)

［註：為了避免 data missing，收集數據時，我們多收集了前半個月的交易日數據，取得前 n 天的所有特徵後，再丟棄不需要的交易日］

B. Regression

首先我們嘗試對 2019/1/1 ~ 2022/12/31 這段時間內所有交易日的收盤價折線圖做 polynomial regression，得到預測的收盤價後再做分類。我們將 2019/1/1 起的每個交易日依序編號為 1,2,3..，再對{日期編號-收盤價}做 regression。得到 polynomial function 後，我們做分類的方法如下: 把想預測的交易日也做編號後代入 function，得到預測的收盤價，再跟前一日收盤價做比較，判斷預測日的 label 是漲或是跌。我們嘗試了 polynomial order 1~10，並分析 accuracy 的變化來找出表現最好的 polynomial order。

C. Classification

1. Data Preprocessing

在進行 classification 前，由於特徵數量較大的，我們利用 sklearn 函式庫，首先對於 data 進行標準化(StandardScaler)，接著進行 Principal Component Analysis (PCA)，將數據降為二維或三維數據。

2. K-means

特徵取的是前面所提到的，我們有蒐集的所有股市相關資訊。我們想要找到對預測日有影響的時間長度，所以資訊取自預測日的前 n 天，n = 3、5、10 ，分別代表短期、中期、長期。我們先對 data 進行標準化，再做 PCA 降維，並嘗試 component 數 = 2 ~ 3。接著我們用各筆降維後的 data 間的 euclidean distance 做 K-means clustering，並嘗試了 K = 3~9。我們做分類的方式是: 先將預測日的特徵標準化並降維後，找到預測日屬於的 cluster （比較預測日 data point 和各個 means 之間的 Euclidean distance）。再用多數決判斷該 cluster 的 label，也就是 cluster 中如果 label 為漲的 data 多， cluster 的 label 就是漲，反之為跌。預測日的 label 就是它屬於的 cluster 的 label。

3. Perceptron

利用 sklearn 函式庫中的 Perceptron，進行分類。分析在沒有進行降維與當 PCA 降至二/三維，此三種情況下，取前 n 天交易日特徵，以監督式學習之線性區別分析表現 n = 3、5、10)。

4. MLP

如同 Perceptron 的做法。MLP 使用到 2hidden layers、Relu，來進行非線性區別分析。

5. LDA

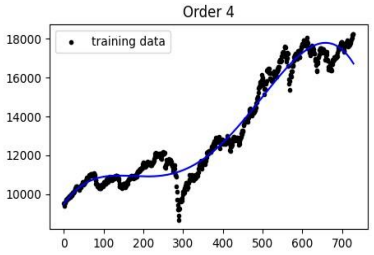
利用 sklearn 函式庫中的 Linear Discriminant Analysis(LDA)，將前 n 天交易日未進行 PCA 降維的特徵，以 LDA 降至一維(預測之類別數 - 1)，直接作為分類器使用。

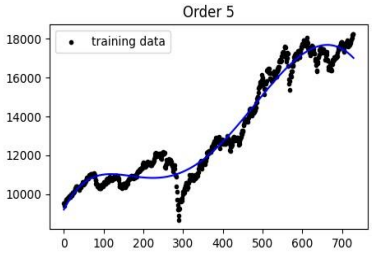
III. RESULT&ANALYSIS

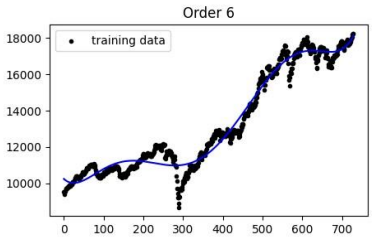
我們判斷實驗結果好壞的指標是 test data 的 Accuracy，公式如下：

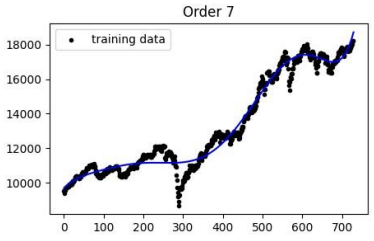
$$Accuracy = \frac{\text{Number of correct prediction}}{\text{Number of testing data}}$$

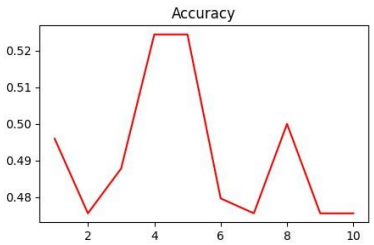
A. Results of Regression









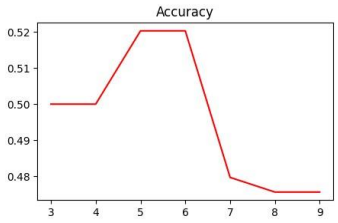


上面四張圖是 polynomial order = 4 ~ 7 的 regression 結果。最後一張圖為 polynomial order = 1 ~ 10 的 accurcy 折線圖，可以看到 order 大於 5 會出

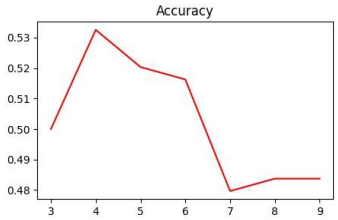
現 over fitting 的情況，order = 4、5 時表現最好，accuracy 為 0.524。

B. Results of K-means

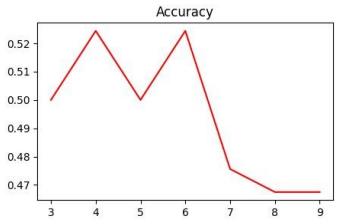
PCA component = 2
look back days = 3



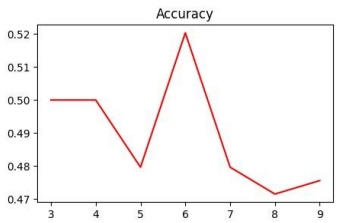
PCA component = 2
look back days = 5



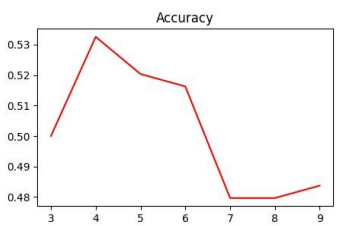
PCA component = 2
look back days = 10



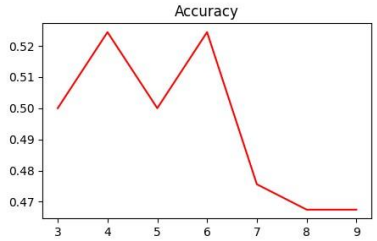
PCA component = 3
look back days = 3



PCA component = 3
look back days = 5

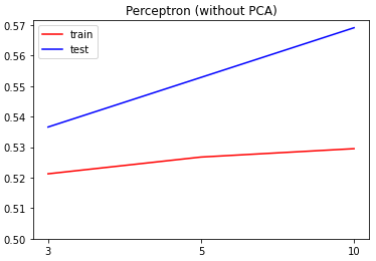


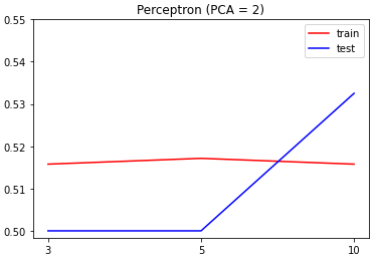
PCA component = 3
look back days = 10

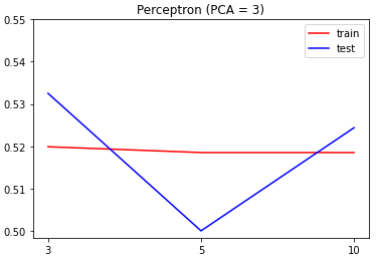


上面六張圖分別為 PCA component 數 = 2、3，特徵取自前 3、5、10 天的股市資訊時，K = 3 ~ 9 的 accuracy 折線圖。可以看到最好的表現出現在 PCA component = 2、3，look back days = 5，K = 4 的時候，最好的 accuracy 為 0.5333。

C. Results of Perceptron



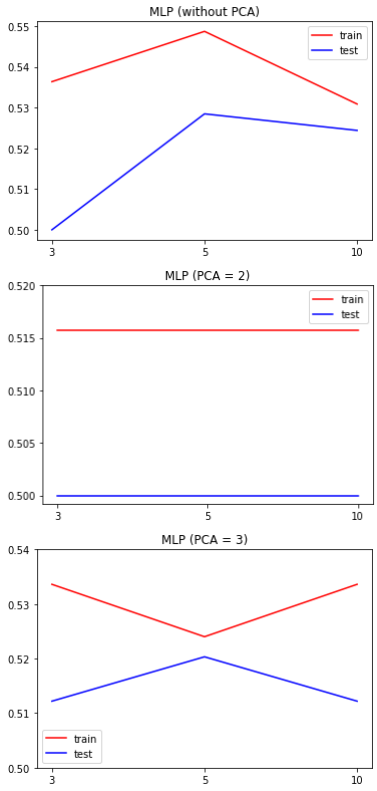




上面三張圖分別代表了沒有降維、降維至二維、

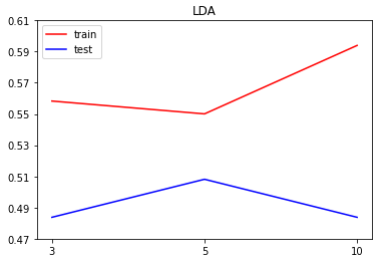
三維，在取 n 天交易日特徵下的各 Perceptron model 的 Accuracy。三者比較之下 without PCA 的表現較好，最好的 accuracy 為 0.569。

D. Results of MLP



上面三張圖分別代表了沒有降維、降維至二維、三維，在取 n 天交易日特徵下的各 MLP model 的 Accuracy。三者比較之下 without PCA 的表現較好，最好的 accuracy 為 0.529。

E. Results of LDA



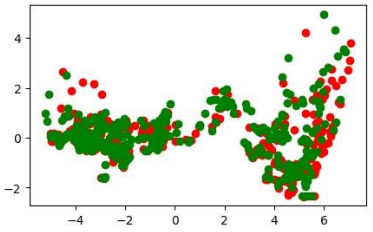
上面的圖片代表了取 n 天交易日特徵下的各 LDA model 的 Accuracy，最好的 accuracy 為 0.508。

F. Results of visualization

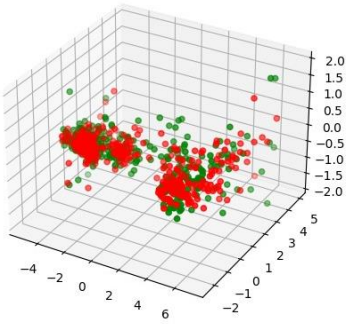
由於上述 model accuracy 皆為 0.5 ~ 0.6，約等於亂猜的準確率(0.5)，模型表現不好，因此，我們決定對 PCA 及 LDA 數據進行視覺化，並分析。

- PCA 視覺化

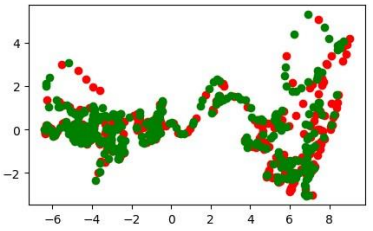
PCA component = 2
look back days = 3



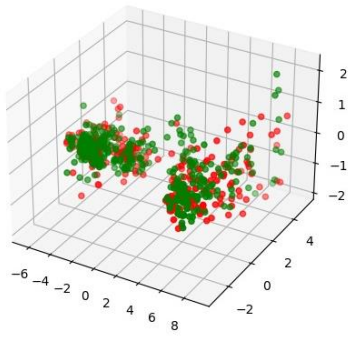
PCA component = 3
look back days = 3



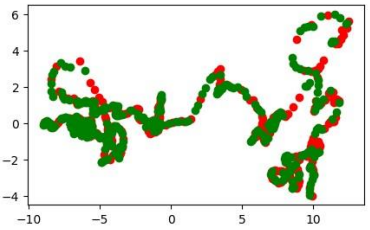
PCA component = 2
look back days = 5



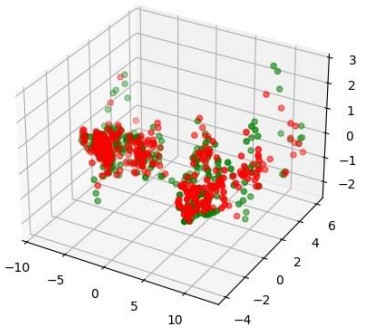
PCA component = 3
look back days = 5



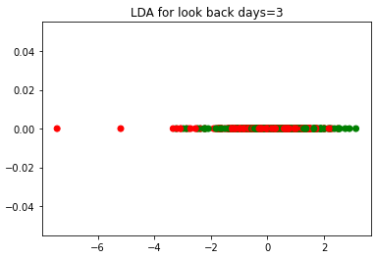
PCA component = 2
look back days = 10

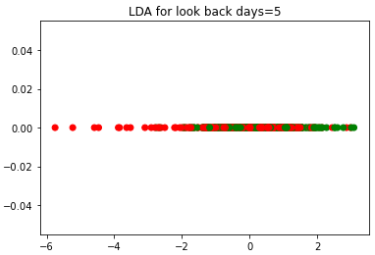


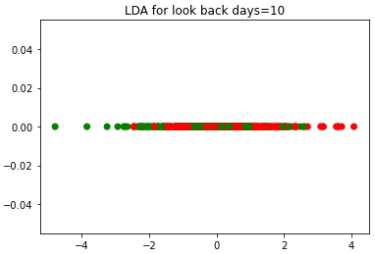
PCA component = 3
look back days = 10



- LDA 視覺化







從視覺化的 PCA 及 LDA 數據可以看出，經過 PCA 或 LDA 的操作之後，對於數據的區分作用不大。

其可能原因是為數據具時間性，而 PCA 及 LDA 不適用於時間序列的數據，且特徵間具有高度關聯性，因此在經過 PCA 或 LDA 降維之後，損失了許多資訊，使得分類結果表現不好。

IV. FUTURE PLANS

我們嘗試根據前面分析出的結論改良分類器，並持續分析改良後的分類器預測錯誤時的情形，找出錯誤的原因並繼續改良，以提升 Accuracy。

A. Long Short-term Memory (LSTM)

分析完前面的分類方式表現不好的可能原因之後，我們為了保持 data 時間序列的特性，嘗試使用 LSTM 模型。模型的架構為一層 LSTM layer 接一層 dense layer，input 為想預測的交易日前 n 天的收盤價(n = 3、5、10)，output 為預測收盤價。

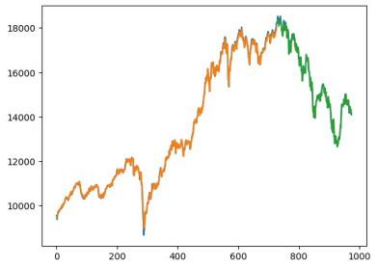
在判斷漲跌方面，我們最初的做法為將預測收盤價與前一交易日收盤價做比較，但卻出現 regression 的 accurcy 高，classifaction 的 accuracy 卻不超過 0.6 的情況。分析 regression 的結果後，我們發現當連續幾個交易

日的收盤價之間差異極小時，預測結果誤差雖然小，但是正誤差或負誤差會嚴重影響 classifacation 結果。

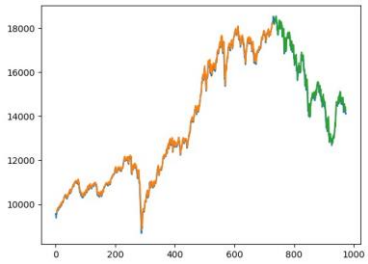
因此與前一交易日收盤價做比較的分類方式，在這種情況下反而不能體現整體股市的漲跌趨勢。於是我們將判斷漲跌的方式改為與前十個交易日的平均收盤價比較，並以同樣的方式改寫 ground truth label。

B. Results of LSTM

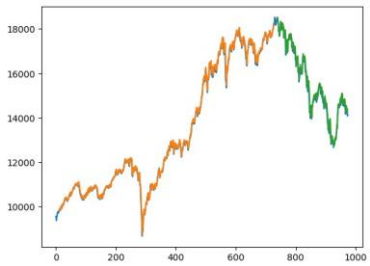
Look Back Day = 3
Accuracy = 0.69



Look Back Day = 5
Accuracy = 0.75



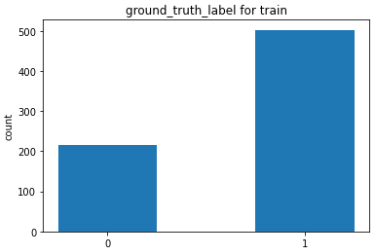
Look Back Day = 10
Accuracy = 0.81

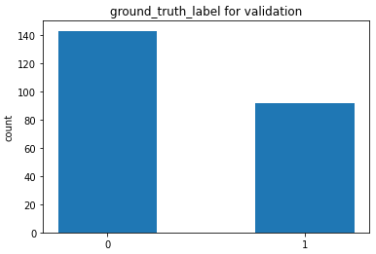


上面三張圖分別是 LSTM 模型在 look back day = 3、5、10 下的 regression 結果，可以看到預測日前十天的數據都會影響到預測結果，因此最高的 accuracy 出現在 look back day = 10 時。最高的 accuracy 為 0.81，明顯高於前面四種分類方式，但是仍有進步空間，因此我們進一步對錯誤的樣本進行分析。

C. 分析錯誤的樣本

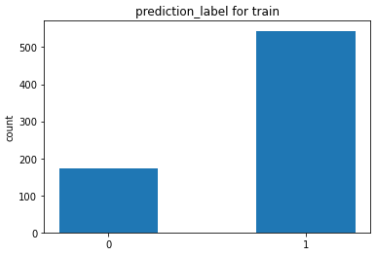
在分析錯誤樣本時，我們首先觀察 ground truth label 的分布情形。

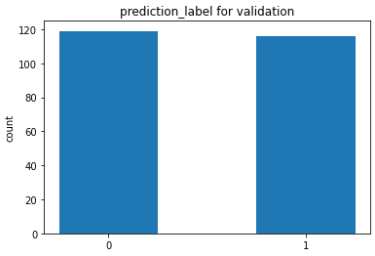




上方兩張圖，分別展示了 train 和 validation 中 ground truth 的分布情形，可以發現兩者間漲跌趨勢分布並不相同。

接著觀察預測之 label 的分布情形。





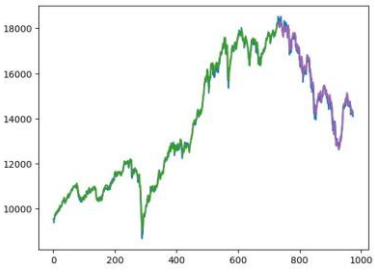
上方兩張圖，分別展示了 train 和 validation 中預測之 label 的分布情形。

與 ground truth 進行比較，可以看出由於 train 的 ground truth 分布中，漲勢較為多數，因此在訓練模型判斷漲跌勢時，模型會較趨向於判斷為漲勢，而在 validation 的 ground truth 中，則是跌勢較為多數，跌勢容易誤判，導致 validation accuracy 較低。

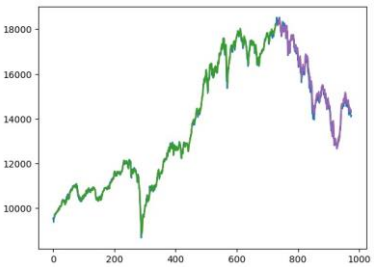
D. 加入月份，進行分析

為了降低資料比例不平衡帶來的影響，讓預測更加精準，我們嘗試將月份以 one-hot encoding 處理後，作為特徵加入模型進行分析，獲得的 accuracy，如下圖所示：

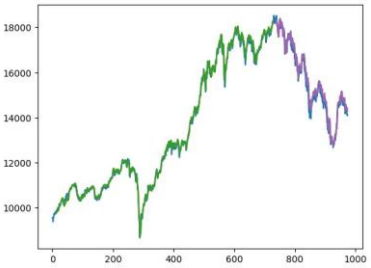
Look Back Day = 3
Accuracy = 0.60



Look Back Day = 5
Accuracy = 0.72



Look Back Day = 10
Accuracy = 0.77



在加入月份之後，模型的表現呈現較差的結果，但是月份應為判斷股市的重要資訊，因此我們認為其原因可能為 one-hot encoding 此編碼方式。

使用 one-hot encoding 進行編碼，可以使月份類別轉換成數值形式，作為特徵進行分析，然而轉換後，特徵維度將會增加 12 維，並且此 12 維的特徵空間具有極高的稀疏度，使得維度增加的同時，有用的資訊又零散地分布在大量數據中，是為 one-hot encoding 編碼導致錯誤分析的可能因素。未來我們將會再嘗試更多 encoding 方式，使月份能作為特徵發揮提升 accuracy 的功用。

V. REFERENCE

A. 銷售量預測 -- LSTM 的另一個應用:
https://ithelp.ithome.com.tw/articles/10195400

B. 股票預測三試 :: 使用小的 Dataset 和 LSTM 做多個測試:
https://ithelp.ithome.com.tw

/articles/10214405

C. [Keras] 利用 Keras 建構 LSTM 模型，以 Stock Prediction 為例:
https://daniel820710.medium.com/%E5%88%A9%E7%94%A8keras%E5%BB%BA%E6%A7%8Blstm%E6%A8%A1%E5%9E%8B-%E4%BB%A5stock-prediction-%E7%82%BA%E4%BE%8B-1