漲跌預測

問題簡述與方法

在這個機器學習問題中，我們希望在每週一 SPY 收盤後預測下週一 SPY 股價的漲跌（附註：在此問題中，若下週一股價為平盤則分類為漲）。此問題為機器學習中二元分類（Binary Classification）的類型。在這次的試驗中，我們使用了兩種不同的機器學習模型：Random Forest Classifier 以及 Logistic Regression Classifier 來進行預測，並將資料集分為 60%訓練集（Training Set）、27%驗證集（Validation Set）與 13%測試集（Testing Set，2015/1/1~2018/6/29 之間的資料）。我們分別使用訓練集來訓練模型參數，並使用驗證集調整超參數（Hyperparameter），最後使用測試集衡量各種不同模型的表現優劣。

資料預處理（Preprocessing）

在原始資料中，我們僅有 SPY 的歷史收盤價格（Closed Price），而僅有單一的收盤價格不足以建構足夠的資料特徵來讓機器學習模型學習參數，因此，我們必須增加資料特徵（Feature）。首先我們使用 python 中的 yfinance 套件從 Yahoo Finance 的平台下載 SPY 過去的每日價量資訊（每組資訊包括當日開盤價、最高價、最低價等資訊），並使用 talib 套件計算多項技術指標作為資料特徵。產生的資料特徵包括：1）相對強弱指數（RSI），2）KD 隨機指標（包含 Slow 慢速與 Fast 快速兩種），3）動量指標（Momentum），4）指數平滑異同移動平均線類別指標（包含 MACD Line：快速 EMA 與慢速 EMA 之差異、Signal Line：MACD Line 九天期 EMA、MACD Histogram：Signal Line 與 MACD Line 的差異），總計七種資料特徵。

在完成資料特徵的建構之後，下一步是生成各項資料組與其對應的標籤。若下周同日的股價高於或等於當日的股價，將標籤（Label）標為 1，反之，則標為 0。接著，由於本項機器學習的課題是在每週一收盤後預測下週一的漲跌，因此在判段完資料日期是星期幾後，我們僅保留星期一的資料。同時，我們也將缺乏下周同日股價（周一遇休假日）的資料移除，並完成資料的預處理。

模型訓練與表現

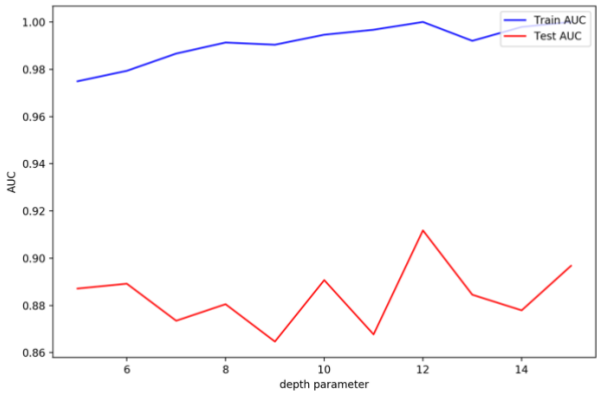
1. 隨機森林（Random Forest）

隨機森林是一種包含多個決策樹（Decision Tree）的分類器（Classifier），其產生的標籤是由個別決策樹輸出的類別的眾數（白話來說，就是個別決策樹投票，票數多者勝出）。個別決策樹生成的原理是使用某些分類指標，將兩種不同類別的資料盡量區分出來（High Purity）。舉個虛構的例子，在 RSI>50 的情況下，下週的股價都是漲，反之則是下跌，這即為一種分類性極佳的指標，而決策樹構築的過程中就要盡量找出這種極具分類效果的指標。回到隨機森林本身，其森林中的每棵樹隨機抽取不同的資料集與特徵，能夠從不同的角度觀察資料集（Bagging），並投票出更好的結果。在這個模型中，主要能教調的超參數為森林中樹的數量，在 sklearn 模組中的預設值為 10，我們測試 5-15 的參數，並比較其在驗證組中的表現，從中選出最佳的超參數作為最終模型的超參數，並測試試驗組的成果。在實驗裡，我們使用 AUC（Area Under Curve）作為衡量依據，當 AUC 越接近 1，則代表預

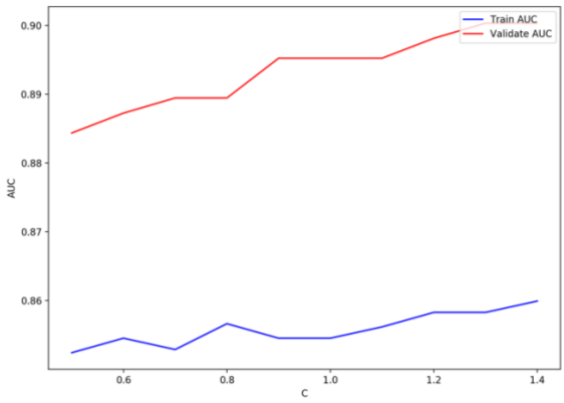
測成果越加。若 AUC 越接近 0.5，則代表預測成果接近隨機猜選。有趣的是，當 AUC 小於 0.5 時，其預測能力是好過 AUC = 0.5 的，因為市場反指標從反向意義來說也是種預測準確的表徵。在使用驗證集測試超參數的優劣後，我們發現當決策樹數量為 12 的時候(圖一)，驗證集的 AUC 達到最高。因此我們將超參數為 12 的模型做為我們的最終模型，並得到測試集的準確率（Accuracy Rate）為 83%以及 AUC 為 0.83。

2. 邏輯回歸（Logistic Regression）

邏輯回歸是一種機器學習分類模型，是利用資料集對分類的邊界建立回歸方程並以此進行分類。在把資料組中的數據放入回歸方程後，回歸方程會輸出一個介於 0 到 1 之間的數，當數字大於 0.5 時，分類器就會把該資料分類為類別 1，反之則分類為類別 0。在這個模型中，主要能教調的超參數為一個懲罰參數 C，主要目的是防止分類器過度擬和 (Overfitting)，在 sklearn 模組中的預設值為 1，我們測試 0.5-1.5 之間間隔為 0.1 的參數，並比較其在驗證組中的表現。我們發現當 C 為 1.3 的時候（圖二)，驗證集的 AUC 達到最高。因此我們將超參數為 1.3 的模型做為我們的最終模型，並得到測試集的準確率（Accuracy Rate）為 85%以及 AUC 為 0.84。



圖一： 隨機森林不同超參數表現



圖二： 邏輯回歸不同超參數表現

模型選擇

兩相比較之下，邏輯回歸在 AUC，Accuracy，Precision 以及 F-score 都相較於隨機森林有較好的表現。唯一的差異是隨機森林在 Recall 擁有較高的表現。Recall 衡量的是在所有漲的數據中，有多少比例的數據被成功辨認出來。在這個本次實驗的情況下，由於隨機森林傾向於把數據都辨認為漲，也因此造就較高的 Recall。但同時，隨機森林的 Precision 也比較低。Precision 代表的是所有辨認為漲的數據中，有多少比例實際是漲的數據。F-1 score 是用來衡量 Precision 與 Recall 兩者重要性的指標，而邏輯回歸的 F1-score 也有較好的表現。多項指標的衡量下，即使隨機森林一般來說被認為是一種較有效果的分類器，但在這次的資料組中，邏輯回歸模型的表現較佳。

	AUC	Accuracy	Precision	Recall	F1-score
Random Forest	0.83	82%	73%	85%	79%
Logistic Regression	0.85	85%	80%	82%	81%

圖三： 最終模型表現比較