**3 classifiers:**

Logistic Regression

Neural Network(DNN)

SVC

**Discussion:**

* How did you preprocess this dataset ?

計算額外欄位dir(今天漲或跌) dpre(前一天的變化幅度) dh(前一天的最高價變化幅度) dl(前一天的最低價變化幅度)，並用dir(今天漲或跌)來當作預測之目標，在觀察資料的相關係數後決定用dh(前一天的最高價變化幅度) dl(前一天的最低價變化幅度)來當作輸入特徵。

* Which classifier reaches the highest classification accuracy in this dataset ?

SVC在train set中最好(約90%)，但是logistic regression在test set中最好(約73%)

* + Why ?

SVC在train set中表現很好，因為採用的特徵可能剛好都是關鍵或是在train set中產生了overfitting，此外svc本來就適合在維度較少及資料數較少的樣本，但是所有分類器在test set準確度普遍不高，可能是因為股票是時序資料且影響股票漲跌的因素有眾多，不利於傳統分類器的方法(因特徵沒有順序的觀念)，而且訓練與測試集中間有時間差可能市場性質也會發生改變，依目前的方法比較像是在fitting training dataset而已，若要有較大的改進可能需要更多的額外資訊(例如相關新聞)或是改用RNN或是LSTM來做時序預測，不過這樣可能就比較不像是分類器的問題了。而nn的表現約在7成，是因為本次選擇的特徵數和樣本數都沒有很多，導致無法發揮nn的強大fitting能力。

* + Can this result remain if the dataset is different ?

SVC在test set不維持(降至54%)，可能依據dataset改變，因為影

響股票漲跌的因素有很多，不是目前的擁有特徵就能決定的。

反而是logistic regression在test set和train set中表現幾乎相同(約

73%)，以綜合結果來說表現是最好的，可能是因為本來fitting的

程度就沒有到非常高反而留下了一部份的彈性。

* How did you improve your classifiers ?

計算額外欄位和特徵dpre(前一天的變化幅度) dh(最高價變化幅度) dl(最低價變化幅度)

-logistic regression

嘗試不同的solver發現liblinear表現較好

-svc

嘗試不同的kernel發現rbf表現較好

加入max\_iter減少overfitting

-dnn

加入多層hidden layer

調整Activation function

調整Epochs

調整batch\_size

**Future improvement:**

不限制在分類器問題，改用RNN或是LSTM來做時序預測股價的數值。