Report

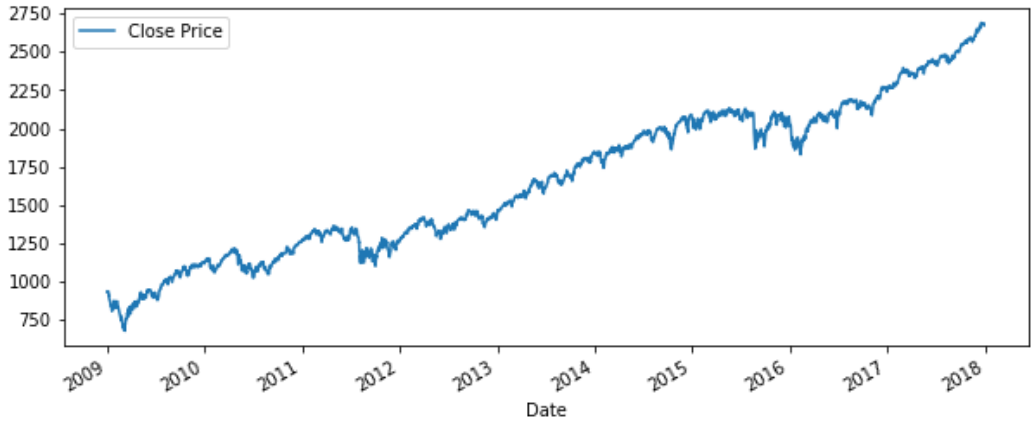
F74052201 陳鈺潔

- **How did you preprocess this dataset ?**

1. 定義問題：預測每天的收盤價漲幅變化。在資料集中建立 Up Down 作為 ground truth，1 代表相較前一日收盤價上漲、0 代表相較前一日收盤價下跌。

	Date	Open Price	Close Price	High Price	Low Price	Volume	Up Down
0	2009-01-02	902.99	931.80	934.73	899.35	4048270080	1
1	2009-01-05	929.17	927.45	936.63	919.53	5413910016	0
2	2009-01-06	931.17	934.70	943.85	927.28	5392620032	1
3	2009-01-07	927.45	906.65	927.45	902.37	4704940032	0
4	2009-01-08	905.73	909.73	910.00	896.81	4991549952	1

2. 把日期資料轉化成純數字格式，作股價變化圖，發現收盤價逐年上升，但是每天還是漲跌不定。



3. 把 High Price 和 Low Price 取出，求出其中間值 Mid Price，只使用 Mid Price 作為 feature，捨棄 High Price 和 Low Price，因為 Mid Price 的值較接近 Close Price 和 Open Price，較有參考價值。
4. 把 Volume 劃分成五個區間，發現最大的區間的上漲機率最低，因此分成在最大區間(High Volume=1)和其他(High Volume=0)兩種，並命名為 High Volume 特徵。使用此特徵並放棄 Volume。

	VolumeBand	Up Down
0	(509556458.048, 2238546790.4]	0.561529
1	(2238546790.4, 3958935180.8]	0.540488
2	(3958935180.8, 5679323571.2]	0.545802
3	(5679323571.2, 7399711961.6]	0.563636
4	(7399711961.6, 9120100352.0]	0.366667

使用 Open Price、Close Price、Mid Price、High Volume 四個特徵訓練。

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

  Accuracy：Adaboost(83.7%)最大，Logistic Regression(82.1%)次之，Neural Networks(80.1%)最低。

  - **Why ?**

    當直接使用 Volume 特徵時，每日 Volume 變化極大，無法看出與收盤價的相關性，Logistic Regression 準確度只有五成，但轉換成 High Volume，特徵更加明顯，準確度上升。

    Adaboost 使用 Logistic Regression 作為基底 model，更著重在分錯的資料上，利用 100 個 Logistic Regression 的模型加強訓練，因此準確度較 Logistic Regression 高。

    Neural Network 只使用四層 layer，activation function 的部分前三層使用 Relu、最後一層用 sigmoid，並用隨機梯度下降以 learning rate=0.001 做最佳化，在 20 個 epoch 內的 loss 下降到特定值後浮動、acurracy 也是，故在特徵少(4 個特徵)的情況下，此模型較難 fit 資料集，準確率為三者最低。

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

    使用 Google Stock dataset，並取用與 S&P 500 相同的特徵、同樣預測每日收盤價的漲跌。在使用與 S&P 500 相同的模型超參數的情況下，Logistic Regression 準確度最高(75%)，Adaboost(65%)次之，Neural Networks(60%)依然最低。三種分類器的準確度都下降，可能與資料集本身的特徵強度、沒有使用最合適的超參數有關。

    Adaboost 準確度低於 Logistic Regression 的原因為 learning rate 太大，應該修改 learning rate 或是調整重複訓練次數，方能使分類準確度上升。Nerual Network 的 loss 還在逐漸下降中，由此可知還沒找到最佳解就結束訓練了，因此準確度偏低，應重新調整 learning rate 與 epoch。

- **How did you improve your classifiers ?**

  1. 移除 High Price、Low Price 特徵，因為此兩項為極值，變異性較大，

沒有明顯的特徵會影響收盤價。

2. 新增特徵 Mid Price，把 High Price 和 Low Price 取中間值，此值與 Open Price 和 Close Price 較相近，較值得參考。
3. 觀察特徵 Volume 並作數值轉換。由於 Volume 為極大值且數值不易觀察與收盤價關聯，切成區間較能觀察特定範圍內的 Volume 是否與漲跌有關，發現 Volume 值最大的區間上漲機率不到四成，因此新增一個特徵用來判斷是否在該區間，使用此特徵後，Logistic Regression 的準確度就大幅上升。
4. 調整超參數
   - Logistic Regression 使用 L-BFGS 最佳化，收斂速度快；使用 multinomial 作為分類方式，準確度較 ovr(one-vs-rest)高。
   - Adaboost Classifier 的準確度受基底模型的分類準確度、重複訓練次數、學習速率影響。使用最大深度為 1 的決策樹分類時，僅有五成準確度，但使用八成準確度的 Logistic Regression 後，準確度表現突出。重複訓練次數與學習速率之間則是互相制約，提高重複訓練次數同時也提高學習速率時，反而會使準確度下降，最後取用 100 次、學習速率為 2.0，可以達到不錯的準確度。
   - 使用不同的 activation function、optimizer 會使 Neural Networks 的準確度提升，輸出層使用 sigmoid 激活、隨機梯度下降法能夠使 loss 在 learning rate=0.001 時快速下降，但若使用 mean squared error 而不是 binary crossentropy 作為 loss function，每次 epoch 的 loss 與 accuracy 值並不會改變。