Report

F74066527 洪邵澤

資料觀察

- 首先將 close price 畫出來觀察看看走勢
- 再來觀察 Volume 是否與 close price 有什麼關聯性
- 發現 4 種 price 在 scale 過後會幾乎長得一樣
- 發現 漲 的比例約莫為 54%，資料分布還算平均

How did you preprocess this dataset ?

根據不同的方法我做了不同的處理

- 如果用到 Volume 時我就會把所有的 data 做 scale，讓彼此之間的差距不會太大
- 需要用到前 N 天的資料時，我就會把 N 天的資料放在一起，並且做 reshape 符合 model 的 input
- 其中一種方式我的資料是直接放漲跌的情形下去 train， 所以我要先將每一天的 Close Price 做 diff 才能拿到漲疊情況
- 當我需要 y_label 時，我會自己去 diff Close Price ，並且把漲跌轉換為 1 跟 0

Which classifier reaches the highest classification accuracy in this dataset ?

- 使用前面數天的漲疊結果下去當訓練資料的 RandomForestClassifier 是最高的
- trainset: 0.9315
- testset: 0.5714

Why ?

1. 我認為直接拿漲疊情況(1 或 0)下去 train 的好處在於他不會受到漲幅或跌幅的影響，重點應該就會放在連續數天的漲或跌
2. 我認為 RandomForestClassifier 是經過 ensemble 的方法，理應可以應對更多狀況，表現會比 logistic regression 好是可以理解的

3. 我認為直接拿股價下去 train 會遇到漲幅跌幅影響，而且漲幅與跌幅不一定會直接與之後幾天的漲跌有直接關係，所以直接拿漲或跌下去的結果會比較好。

Can this result remain if the dataset is different ?

- 雖然在 RandomForestClassifier 看出明顯的 overfitting，但是他在的 testset 的 accuracy 還是最高的 0.57，所以我保留了此結果。
- 再拿 logistic regression 的來看，會發現在 train 跟 test 上都是 0.55 左右，看似效果還不錯，在兩種 dataset 上有差不多的準確度。
- 但 logistic regression 的 confusion matrix
  tn, fp, fn, tp: 28, 91, 20, 113
  會發現 logistic regression 的 false/true positive 都滿高的，可見 model 猜了一堆 1，我認為這個 model 有一點學壞了。
- 而在 RandomForest 的 confusion matrix
  tn, fp, fn, tp: 63 56 52 81
  雖然 overfitting，但是猜測分布還算平均，所以我認為還是個不錯的結果。

How did you improve your classifiers ?

- 嘗試 3 種不同的方式去處理 data
  - 拿前一天的資料
  - 拿前 N 天的資料
  - 拿前 N 天的漲跌情況
- 利用這 3 種不同的方式當 train data 分別去訓練了不同的 model，我認為嘗試不同的 train data 可以方便我找出比較適合用來訓練這個 dataset 的資料處理方式

- 調整參數
  - LogisticRegression
    - solver: 在預測股價時選用 liblinear, 在二分類時選用 saga
  - RandomForest
    - n_estimators: 產生更多種可能
    - max_depth: 降低 overfitting
    - min_samples_split: 避免將資料劃分太小
    - oob_score: 設成 True 讓他可以自行驗證

- Neural Network
  - 選用 LSTM，也用了雙向的 LSTM 但效果差不多
  - 選用 Conv1D 但效果欠佳
  - 調整 learning rate 改變 gradient descent 效果
  - Loss 選用適合用在數值的 mse
  - activation 用較不容易 vanishing gradient 的 relu
  - 使用 early stopping 降低 overfitting
  - 調整 epoch，後面發現 epoch 小一點的效果會比較好，我認為是跟 overfitting 有關係

Summary

- NN 的效果比我用 randomforest 還差，我想是我不太會疊 model
- 在 testset 上的作圖，我發現 scale 過後其實滿相似的，但 accuracy 還是很差
- 有些時候 confusion matrix 分布不平均但 accuracy 變高，因為 model 幾乎都猜 1，就可以有大概 54% 左有的準確率