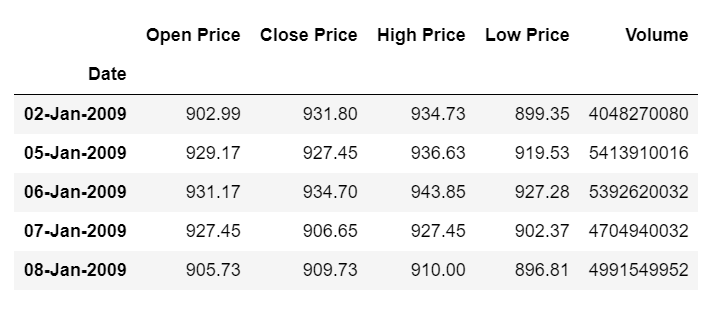
FDA hw3 stock movement predict

F74046030 李昕

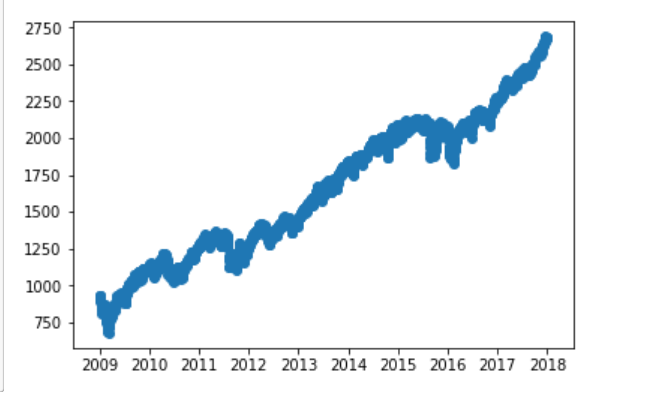
1. 目標

利用logistic regression、svm、neural network三種分類方法，來分類預測股票的上升下降趨勢。

1. Data preprocessing
2. 原本的data長相 :



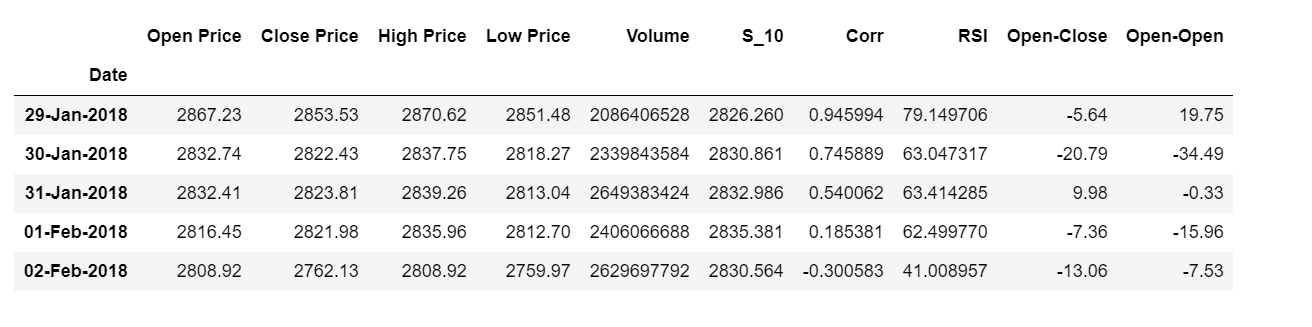
包含了open、close、high、low、volume五種資訊。



訓練資料的收盤價趨勢圖，可以看到是上升的趨勢。

但就算是這樣，資料中上升和下降的比例還是大約在1:1，不一定有辦法用整體的趨勢判斷。

1. 加入新的特徵 :



S\_10 :　每10天股票收盤價的平均值

Corr : 相關係數

RSI : 相對強弱指數

Open-Close : 用今天的開盤價減掉昨天的收盤價，找到中間的差距

Open-Open : 用今天的開盤價減掉昨天的收盤價，找到中間的差距

1. 把含有NaN的資料drop掉
2. 建模預測結果 :
3. 用原本的資料訓練 :

‧logistic regression



‧svm



‧neural network



雖然三個方法裡頭算出來每一筆資料的機率還是不太一樣，但結果都是1的機率比-1還要大，所以最後大家都猜1，在同樣的測資情形下，準確率也就都一樣了。因為覺得可能是資料的特徵不夠強，害模型沒有辦法學到東西，所以決定加入新的特徵訓練。

1. 加入新的資料訓練 :

‧logistic regression



‧svm



‧neural network



加入了新的特徵之後狀況還是一樣。但不同的是這次準確率卻只有0.50多。推測應該是在建立新的特徵的時候有把一些本來沒有出現NaN但後來出現NaN的資料drop掉了，才導致最後準確率的改變。但關於模型的預測，除了機率有一點點小小的不同之外，全部都猜1的情形還是沒有改變。

1. 結論

這三種分類器在這次的試驗當中都沒有辦法很有效的做到我們想要的結果。 一開始只用原始的資料來訓練的時候以為是資料不足、特徵不足的問題，所以加入了新的特徵，但結果也還是不優。這樣的話，如果不是我的資料設的有問題，那就是這個資料本身對於預測就有一定的難度。

不過想想也有道理啦。如果股票有這麼好預測的話，高雄就真的可以發大財了。征服宇宙!

1. 討論
2. preprocess :

‧把不需要的資料拿掉(ex : date)

‧把空的資料處理掉(NaN)

‧根據現有的資料產生新的資料作為用來訓練的特徵(ex : 利用兩天的收盤開盤價格相減就可以得到差距和上升或下降的趨勢)

1. 三個分類器最後的結果基本上都一樣，目前還沒有辦法判斷哪一個分類器最適合這一筆資料。但因為這樣的情況可能是資料的性質導致的問題，所以如果換了另外一筆資料，說不定就會有不一樣的結果。
2. Improve :
   * 資料方面 :

這次雖然有增加一些新的資料當作特徵，但我並沒有想辦法把一些比較不重要的資料篩選掉，就這樣放著給他們訓練，最後造成干擾，讓分類器沒辦法學到真正的特徵。或許可以試試看feature selection的方法。

另外也可以再想想看有沒有其他可能可以嘗試的特徵。例如10天之內的成長趨勢之類的。這次沒有把一些有想到的特徵加進去，下次可以試試看。

* + 分類器方面 :

除了資料的處理之外，分類器的選擇跟參數的設定也有一定程度的影響。每一個分類器適合不同的情況，也可能會需要一些個別的預處理。例如在使用svm時，可能就會需要對資料做scaling，把資料都縮進同一個範圍。這次在用svm的時候也有嘗試對資料做scaling，但因為發現效果跟沒做差不多，而且想要和另外兩個資料比較，就維持了資料的一致，沒有在用svm的時候做scaling。

可能的話，透過分類器的選擇和參數的調整，說不定可以讓結果再更好看一些。