F74076027 林政傑 FDA_HW3-1

i. 資料前處理:

1. 將開盤價和收盤價相加除以二得到中間值，放在'Mid Price'欄位，然而後來跟同學討論才知道只需要關注收盤價即可，所以這個 feature 作廢。
2. 將原本日期的格式修改，分為 'Day'、'Month'、'Year' 三個欄位，並且把年份從 2009..2017 改為 1..9，然而股票漲跌似乎跟月份日期沒有直接關係，尤其 train data 中的年底收盤價全都是拉升，然而 2018 年的資料發現年末幾乎都是重挫，所以便沒有將月份日期放入 trainning feature。
3. 計算每日跟前一日收盤價的價差，在 'Delta Close Price' 欄位，單獨使用這個 feature 可以直接讓模型正確率變成 1.0。然而這是個不切實際的 feature，因為在現實中，不可能在股票收盤之前就知道收盤價進而算出收盤價差，如果可以就等於是未卜先知了。所以不能使用這個 feature 訓練模型。
4. 承第三個前處理，收盤價差可以進而判斷當日是漲還是跌，將漲或維持不變標示為 1，跌標示為-1，儲存在 'Ups and Downs' 欄位，此欄位即是模型要預測的東西，也就是 Y。
5. 計算 n 日均線，均線是股票經常觀察的指標，我一開始算了 5 日均線和 20 日均線，後來獲得同學的建議，又算了 2 日均線，並把 20 日均線刪除，因為 20 日均線在這個題目已經算中長期指標，不太適合拿來判斷每日漲跌。n 日均線會放在 str(n) + ' day mean' 欄位，計算均線後，額外增加兩個欄位，第一個欄位為'Delta ' + str(n) + 'day mean' ，以 1 代表均線為上升或水平，-1 代表均線下降；第二個欄位為 'Delta ' + str(n) + 'day mean 2'，為均線升降的原始值。有些模型適合放入二元值訓練，有些適合原始值。

ii. 1. 使用 Logistic Regression 作為梯度公式的 SGDClassifier 準確度最高，原因也許是我放入的 feature 比較少，只有 5 日均線跟 2 日均線升降的二元值，較不會對梯度下降法的能力有抑制性，而 neural network、random forest、decision tree 等需要比較充足且有用的 feature 才會大幅提升其準確度。在選擇 feature 時，每當選入絕對的值，如 'Close Price' 就會導致準確度下降，而選 Delta 類的欄位才有機會提升準確度。

2. 個人推測用不同 data set 結果會有很小的差異，但準確度應該差不多，因為 5 日和 2 日均線的差值非常接近收盤價差值，而且跟時間和收盤價格完全沒有關係，只和前後 n 日的升降有關係，所以應該要可以套用到所有股票。
個人推測用不同 data set 結果會有很小的差異，但準確度應該差不多，因為 5 日和 2 日均線的差值非常接近收盤價差值，而且跟時間和收盤價格完全沒有關係，只和前後 n 日的升降有關係，所以應該可以套用到所有股票。

iii. 嘗試使用不同的超參數、嘗試套用及抽出不一樣的 feature、試著對 feature 二值化或標準化或其他進階的處理、和同學討論互相交流 feature 跟分析方法。