FDA HW 3-1 report

陳郁明 F74056255

1. **How did you preprocess this dataset ?**

   Dataset 中有 5 個欄位，其中四個欄位：Open Price、Close Price、High Price、Low Price 分布較相似，大略分布於[670,2700]的區間，因此利用 sklearn.preprocessing.scale 將其標準化至約[-2,2]的區間。剩餘的 Volume 欄位由於數值較大(> 1E8)，因此以 log10 將其轉換到數值較小的空間。

2. Which classifier reaches the highest classification accuracy in this dataset ?

   本次作業中使用的三種分類器：Logistic Regression、Neural Network、KNN，其中以 Neural Network 表現最好(test_acc = 55.42%)，另外兩種分類器的準確度為：Logistic Regression--53.01%, KNN--54.21%

   - Why ?

     線性回歸的 Logistic Regression 可能在這次的實驗裡有先天性的弱勢，某種程度上為 Logistic Regression 的上位替換的 Neural Network，較能適應股票分析之類的較複雜的計算。

   - Can this result remain if the dataset is different ?

     若參數調校得當，我認為 Neural Network 會在其他 dataset 也勝過 Logistic Regression，KNN 則由於在本次實驗中不太穩定，因此不太能斷言。

3. How did you improve your classifiers ?

   除了針對個別分類器尋找最適合的參數外，也可以在基礎的 dataset 上做努力，例如計算 N 天內 Close Price 的平均值、當天的 Open Price 與 Close Price 的價差、隔天的 Open Price 之間的價差、RSI 指標等等，採用經濟學上常用的分析工具最為訓練資料，也能增強分類器的準確度。