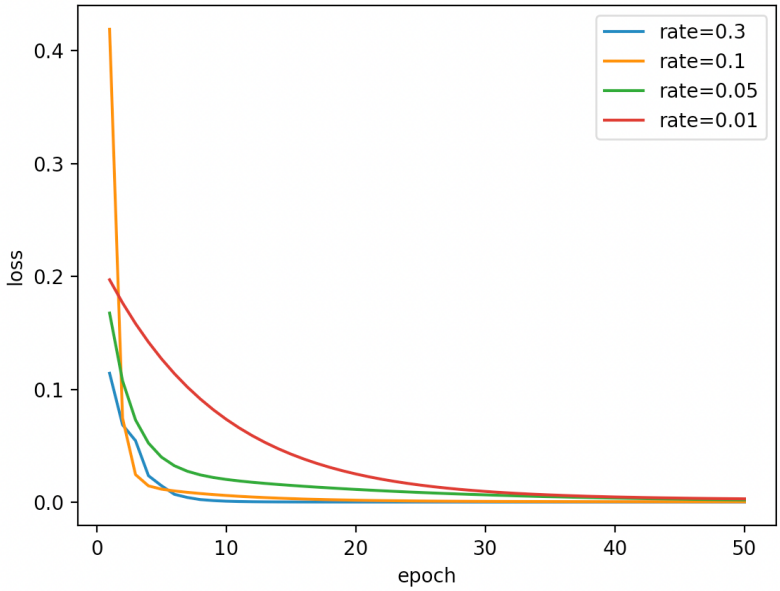
CE6023 Homework1 Report

學號：111526009 姓名：薛竣祐

1. (5%) 使用四種不同的 Learning Rate 進行 Training（方法參數需一致），作圖並討論其收斂過程（橫軸為 Iteration 次數，縱軸為 Loss 的大小，四種 Learning Rate 的收斂線請以不同顏色呈現在一張圖裡做比較）。

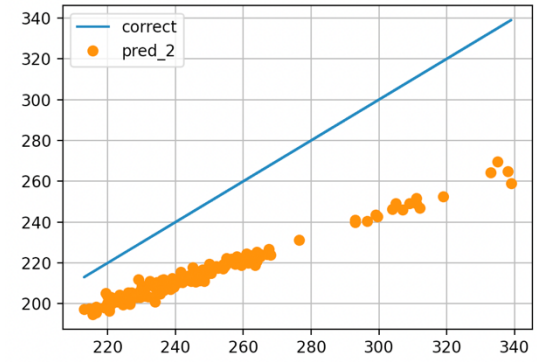
根據實驗發現，Learning Rate越大則學習速率普遍越快，且Learning Rate越低，學習曲線越平滑，但以50次epoch來說，收斂的結果差異不大。

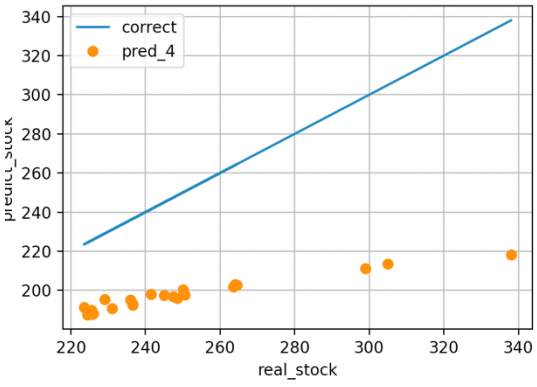


2. (5%) 比較取前 2 天和前 4 天的資料的情況下，於 Validation data 上預測的結果，並說明造成的可能原因。

觀察Validation data發現模型普遍低估股價。但取前兩天資料的預測相較取前四天資料的預測誤差較低。

猜測可能原因有二：一為時間較近之資料對結果影響較大，但模型未作此處理，只取前兩天資料可過濾較遠時間之資料；二為只取兩天資料可分出較多組的學習資料，考慮股票週末不開盤，且模型目的為預測隔天之收盤價，因此取前四天資料的話，一週只有1組資料可訓練，而取前兩天資料的話，一週有3組資料，資料量將有3倍左右。

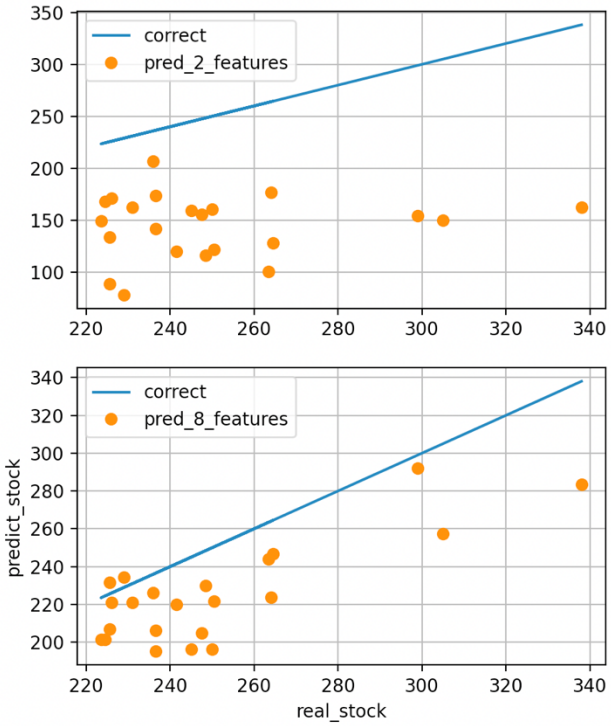




3. (5%) 比較只取部分特徵和取所有特徵的情況下，於 Validation data 上預測的結果，並說明造成的可能原因。

實驗發現使用所有特徵的狀況相較只取2種特徵時預測結果較為平滑、集中。
推測原因可能為特徵較多可以有更多的資訊來預測結果。

但後來測試，在較多epoch及較大dim size的狀況下，只選擇部分的Feature會有較佳的結果。

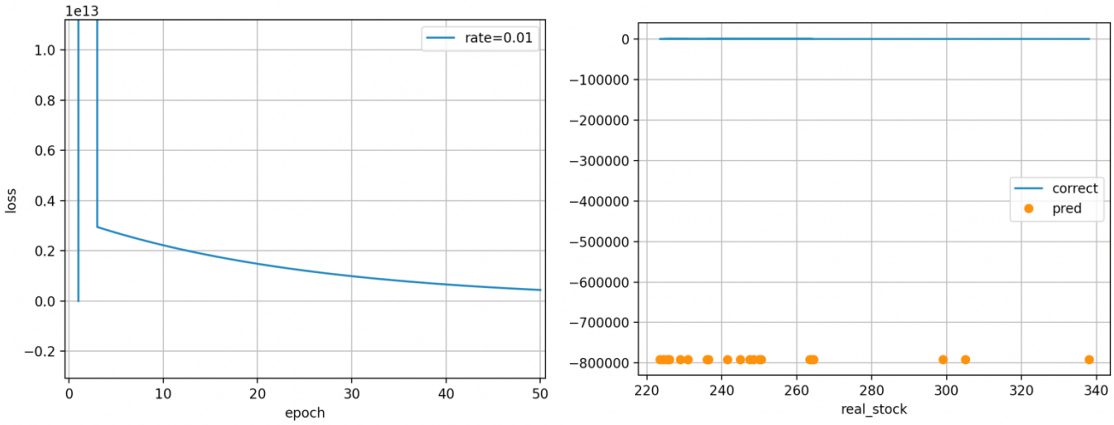


4. (5%) 比較資料在有無 Normalization 的情況下，於 Validation data 上預測的結果，並說明造成的可能原因。

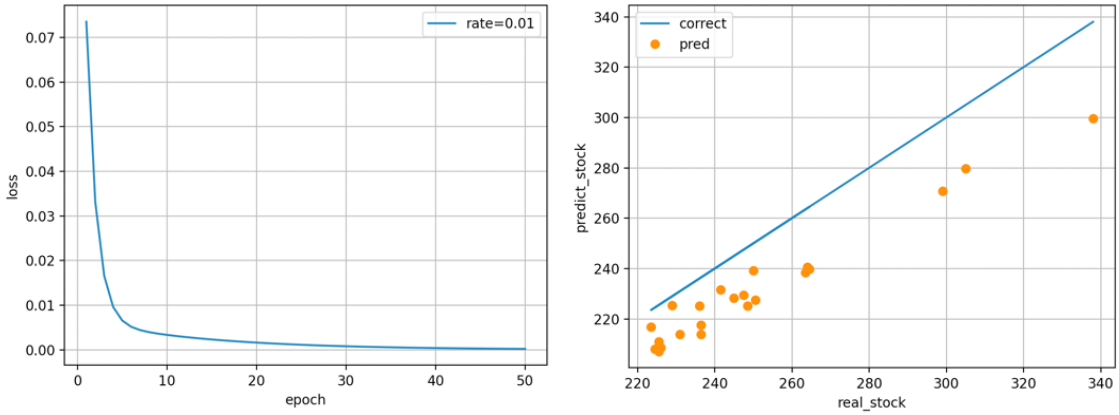
在沒有Normalization時，Loss計算出來的數值將會被極大的影響，根據圖表可以看出就算經過50 epoch，Loss值仍將近1e12，但有經過Normalization的data，Loss只有1e-3，且有逐漸收斂之趨勢，在預測上理所當然的差異相差甚大。
而造成此現象的主要原因為， Lost Function使用MSELoss的緣故，其中MSELoss的公式為[$loss(x_i, y_i) = (x_i - y_i)^2$]，在不同的資料範圍下，MSELoss接收資料若無經過正規化處理，將會隨資料數值大小而變動，因此必須將資料統一正規化過後，才可以去除資料數值大小的影響。

(下頁附圖)

無Normalization狀況下之Loss圖與Validation結果：



有Normalization狀況下之Loss圖與Validation結果：



5. (10%) 請說明你超越 Baseline 的 Model（最後選擇在 Kaggle 上提交的）是如何實作的（若你有額外實作其他 Model，也請分享是如何實作的）。

在助教提供的Regression Model基礎上，多做了以下的變化。

- 額外新增一層Linear層、Relu整流函式
  - Linear -> Relu -> Linear
- 只選取部分Feature
  - Open、High、Low、Close
- 使用Min–max Normalization
  - 使用pandas函式，df = (df – df.min()) / (df.max() – df.min())
- Epoch提高至300
- 使用Testing Data資訊來做正規化而非Training Data

另外也做了以下嘗試，但於Public Leaderboard上成績並未明顯提升：

- 使用LSTM Model
- 使用三層以上的Layer
- 過濾非連續日期(週五至週一)之資料

** 因為 Testing data 預測結果要上傳 Kaggle 後才能得知，所以在報告中並不要求同學們呈現 Testing data 的結果，至於什麼是 Validation data 請參考：https://youtu.be/D_S6y0Jm6dQ?t=1949