Machine Learning HW1

110511010 楊育陞

Part I

Results

1 Fitting curve of the third input feature

我使用 feature x3 作為橫軸，考慮所有其他點，畫出模型產生出的預測並對照資料 target。下圖中橘色點為模型的預測值，藍色點為資料 target，可以些微的發現，M 越大模型會更貼近資料。

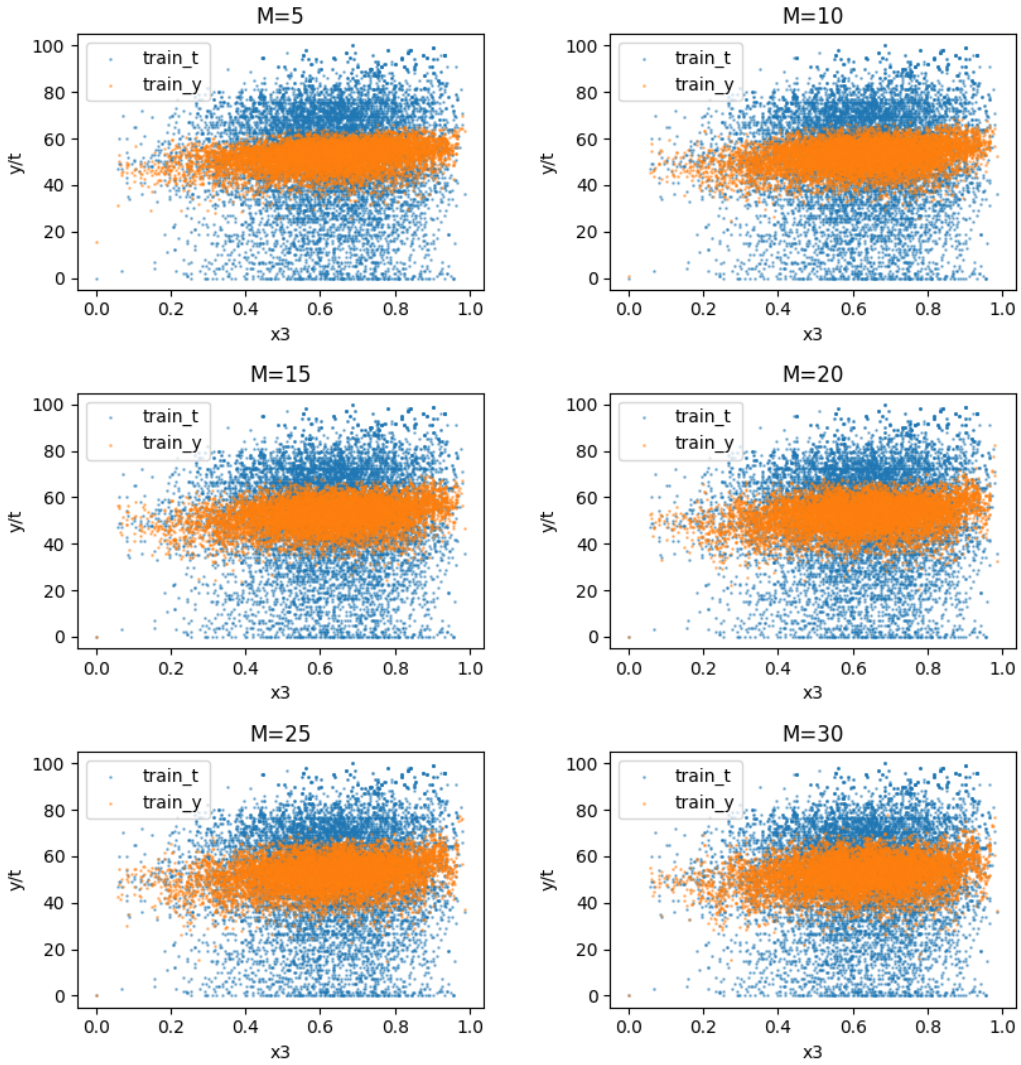

Figure 1: Fitting Curve of feature x3

2 MSE and Accuracy

觀察 MSE 與 Accuracy，其中隨著 M 越大，training 的誤差會越小，但是 test 到某個程度開始會有極大誤差。因此需要做 Regularization(Ridge Regression)，對模型參數做壓制，避免模型有太大的 bias，更泛用於真正的資料分佈。

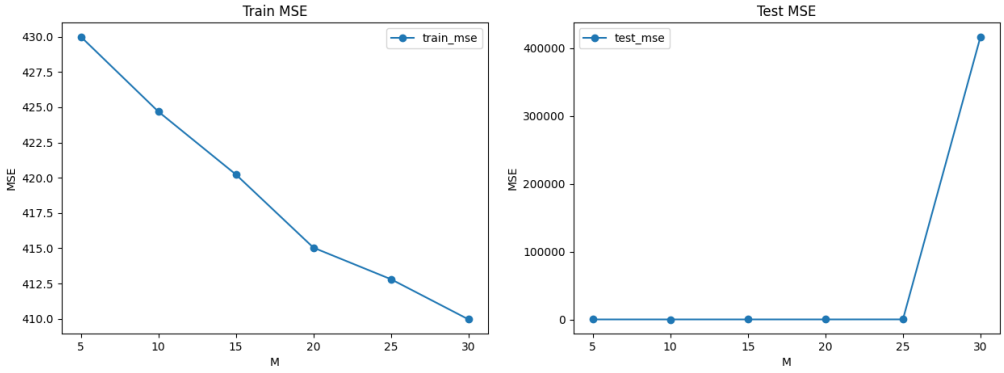

Figure 2: MSE of Normal Regression: Train/Test

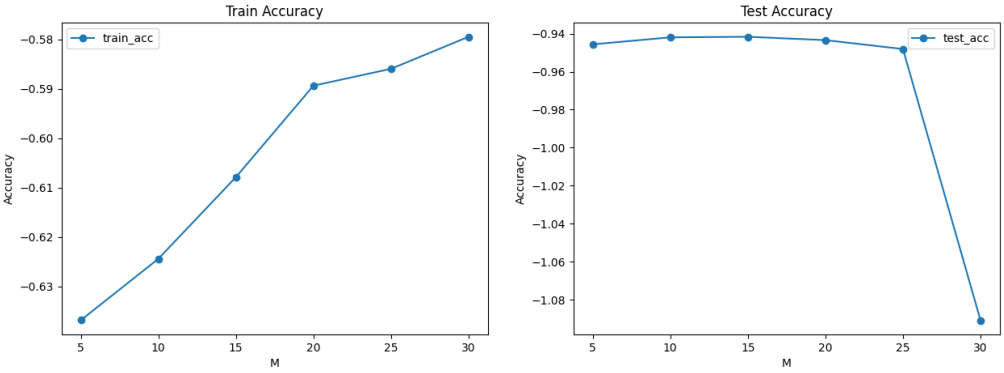

Figure 3: Accuracy of Normal Regression: Train/Test

3 5-fold cross-validation

使用 cross validation，我將資料依序切成五等分，在分次做 validation，並在同一個 m 中各個 validation 的 accuracy 做加總，找出總 accuracy 最大的 m 為 5。

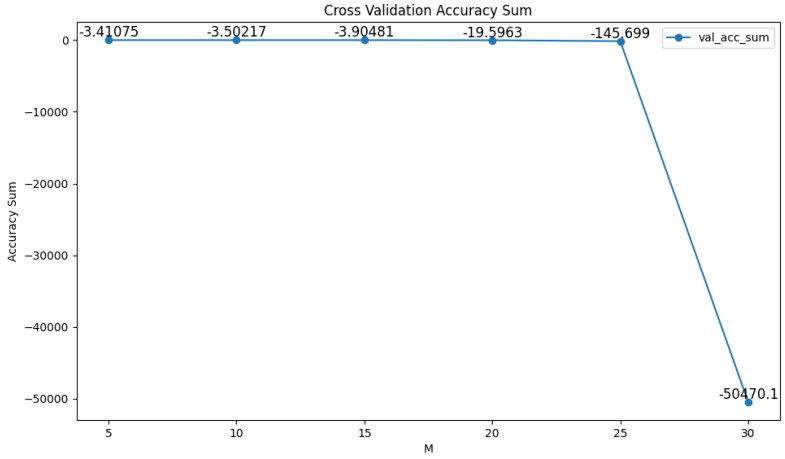

Figure 4: Cross Validation Accuracy Sum for Ms

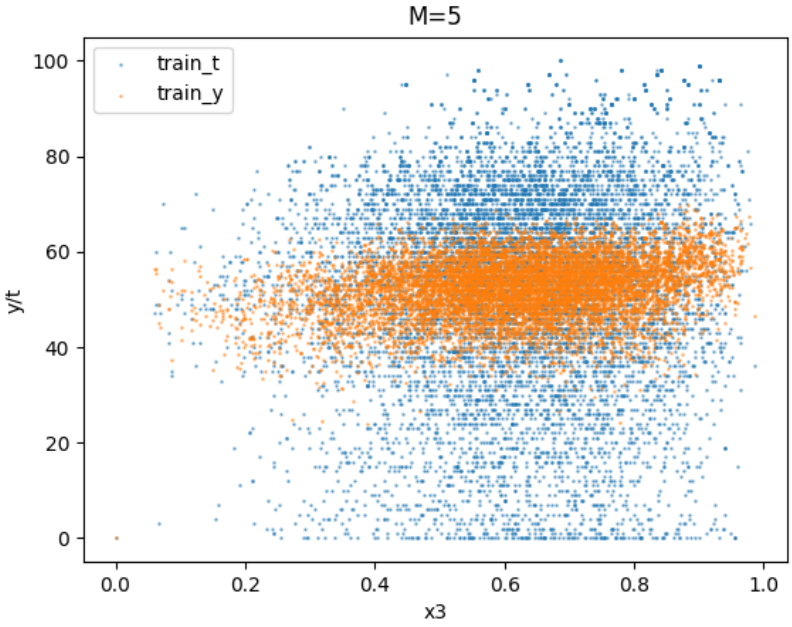

Figure 5: Best M=5 using Cross Validation

4 Regularization

做 Regularization 後可以發現，模型不會因為 m 有太大的誤差。

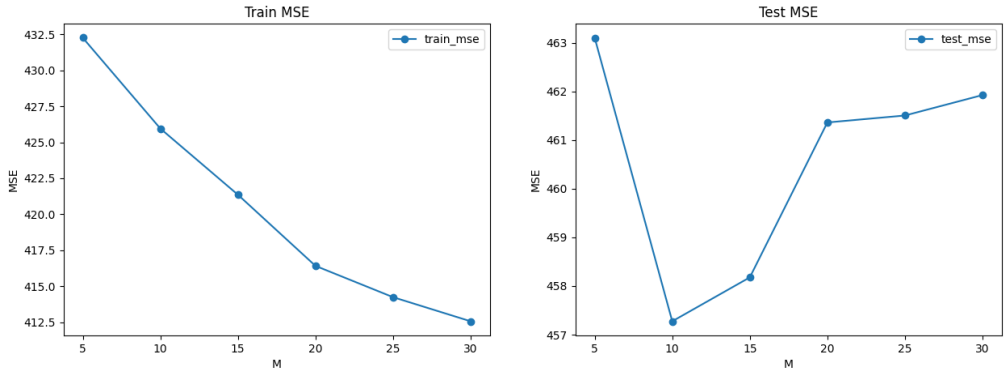

Figure 6: MSE of Ridge Regression: Train/Test

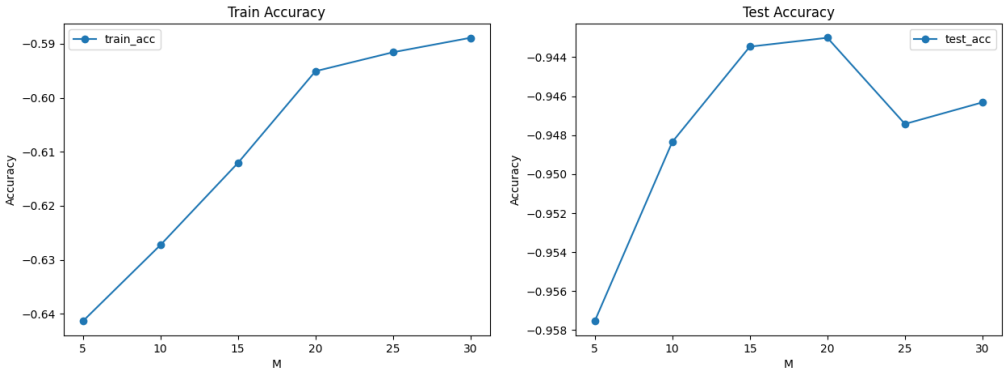

Figure 7: Accuracy of Ridge Regression: Train/Test

Part II

Further Discussion

5 影響最大的 feature

我計算各個 feature 對預測的相關係數，發現最相關的 feature 為 feature5(index=4)，且為負相關。

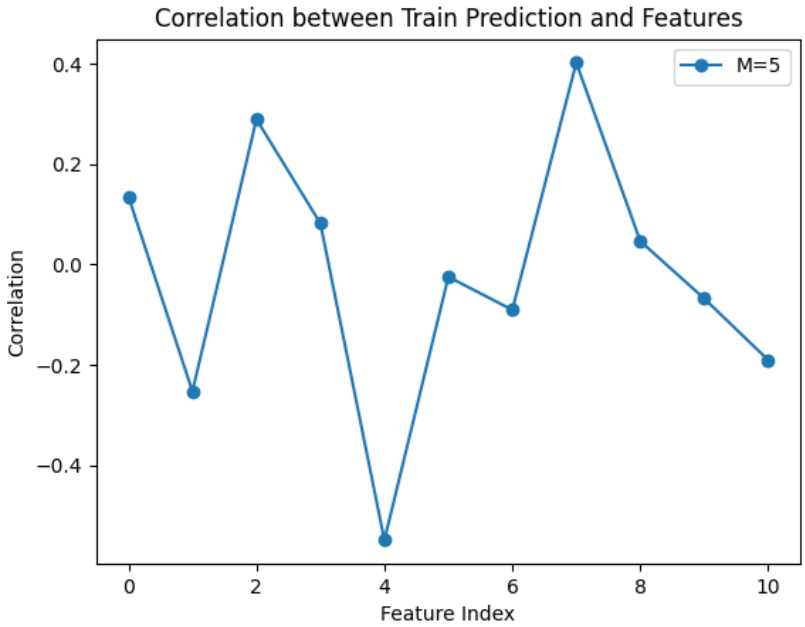

Figure 8: M=5 不同 feature 與 prediction 的相關係數

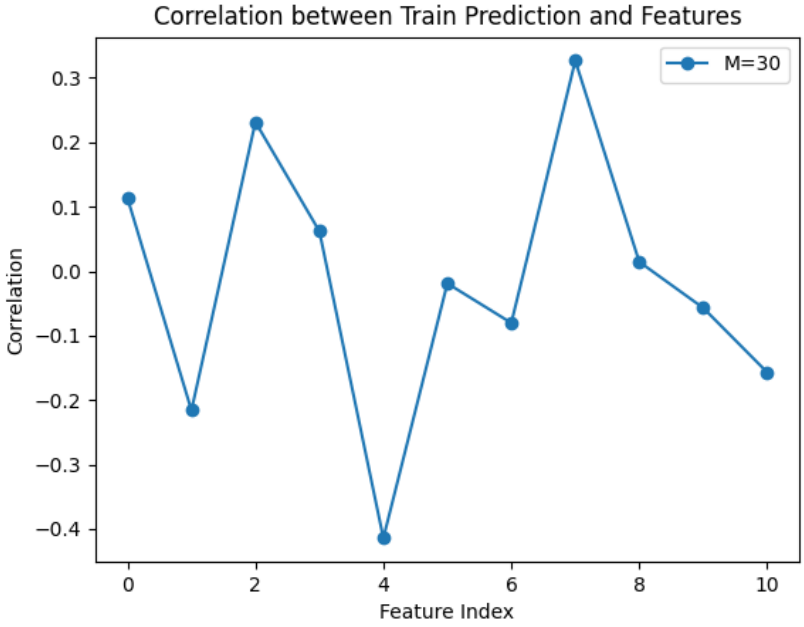

Figure 9: M=30 不同 feature 與 prediction 的相關係數