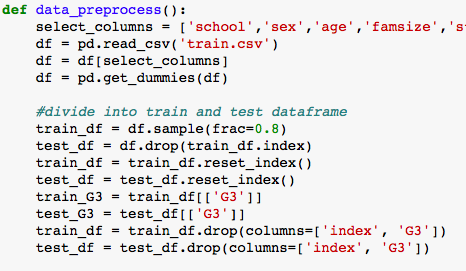
Deep Learning Homework1

StudentID: 107062650

Name: 曾昱榮

Part1 (a):

我將train.csv檔所有資料存到df這個dataframe，再使用DataFrame.sample()的函式將df內80%的資料存到train\_df，而剩下的20%存到test\_df。



Part1 (c):

Part1 (f):

-

Linear Regression with Pseudo-inverse

RMSE: 11.621016643974272

-

Linear Regression with regularization λ = 1.0

RMSE: 11.621107867670135

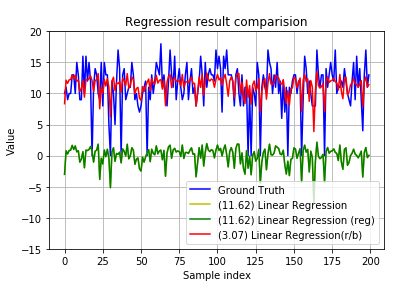
-

Linear Regression with regularization λ = 1.0 and bias term

RMSE: 3.0677115773854338

-

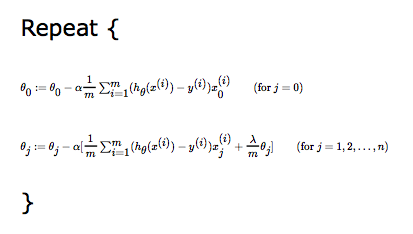
有加bias term的model預測的G3值會較接近Ground Truth，從下圖可以看出，bias term可以上下調整Value的誤差值，所以有bias項的(d)(e) model會有較接近Ground Truth的G3預測值。



Part2 (b):



上圖為regularized logistic regression的cost function，下圖為經過上圖公式針對偏微分更新feature的權重theta值的公式。值是learning rate，值越高訓練越快，但太高也有可能導致theta更新幅度過大而越過minimum的情形。而gradient descent還需設定好iteration的次數，次數只需達到cost的值收斂即可。



Part2 (e):

|  |  |  |  |
| --- | --- | --- | --- |
|  | threshold = 0.1 | threshold = 0.5 | threshold = 0.9 |
| Linear regression | Acc: 0.805  Precision: 0.804 | Acc: 0.835  Precision: 0.843 | Acc: 0.315  Precision: 0.848 |
| Logistic regression | Acc: 0.815  Precision: 0.814 | Acc: 0.84  Precision: 0.857 | Acc: 0.28  Precision: 0.95 |

threshold從0.5變0.9，Accuracy大幅下降的原因為，在原始train.csv的資料裡G3大於10分的為大多數，所以大多數的G3都被分類至label 1，只有少部分被分到label 0，所以預測當分類器在分類類別時，預測1會較容易預測成功而Accuracy也會較高。而當threshold變為0.9時，因預測為1的門檻變太高，而導致預測1的數量大幅下降，然而G3真實類別又大部分為1，所以針對真實類別為1的數據預測準確度也大幅下降，所以accuracy變得相當低。