黃偉祥 X1136010

Loss function : (Y_true – Y_predict)^2 + $\frac{\lambda}{2}$*|w|^2

Gradient : $2\frac{2}{m}\sum_{i=0}^{m-1}\{(Y_i^{True} - Y_i^{Predict}) * X'\} + \lambda * w$

Update Weight = Weight – learning_rate * Gradient

Validation Loss = (Y^{validation} – Y^{validation}_predict)^2 / number of Y

Basic part:

Y_predict = X_{in} * W + noise

Weight_dimentions : {noise , X_{in}}

Advanced part:

X' = {noise , X_0 , X_1 , X_2 , X_3 , X_4 , X_5 , X_6}

$Y^{predict}$ = noise + X_0*W_0 + X_1*W_1 + X_2*W_2 + X_3*W_3 + X_4*W_4 + X_5*W_5 + X_6*W_6

Weight_dimentions : {noise , X_0,X_1,X_2,X_3,X_4,X_5,X_6}

In basic part the input of X is only {weight} variable, but in advanced part we have 7 variables which are {age , gender , height , weight , bodyfat , diastolic , systolic} as X.

First, I don't know why my python cannot recognize the function np.isnan(), so I change my data into pandas.DataFrame and use fillna() to replace missing value(np.nan).

Second, I realize that my model doesn't work so well, then I start to check how to make model perfect, but then I found that the main point is the outliers, so I remove the outliers using IQR and setting lower bound as Q1 – 1.5*IQR and upper bound as Q3 + 1.5*IQR.

Third, the second column of advanced part (gender) is given a string(F or M), then it cannot be doing mathematical operation, so I change it to 0.0 and 1.0.

Fourth, I tried to apply the non-linear basis functions to X, but only improve little in the result.

Last but not least, I realize that my model keep running even after convergences, and this is wasting a lot of time and may cause to overfitting, so I used the loss of validation dataset to have a early stopping point which are:

1. When the loss of validation dataset is not going down. (loss > previous_loss)
2. When the loss is going down very little. (previous_loss – loss < 0.0000001)

Last, I realize that each time of my result will be different, so I set the random seed and run a for loop to get the best random seed of my model.

In conclusion, I tried to apply some features engineering to the dataset and set some early stopping point to prevent overfitting and time waste.