# NYCU Introduction to Machine Learning, Homework 2

110612117 張仲瑜

## Part. 1, Coding (50%):

### (15%) Logistic Regression

1. (0%) Show the hyperparameters (learning rate and iteration) that you used.

```
LR = LogisticRegression(learning_rate=0.00000205, iteration=100000)
LR.fit(X_train, y_train)
y_pred = LR.predict(X_test)
accuracy = accuracy_score(y_test , y_pred)
```

2. (5%) Show the weights and intercept of your model.

```
Part 1: Logistic Regression
Weights: [-0.05552877 -1.3593208   0.99628603  0.18042485  0.03320151 -0.64334277], Intercept: -0.18668346910337374
Accuracy: 0.7540983606557377
```

3. (10%) Show the accuracy score of your model on the testing set. The accuracy score should be greater than 0.75.

```
Accuracy: 0.7540983606557377
```

### (35%) Fisher’ s Linear Discriminant (FLD)

```
Class Mean 0: [ 56.75925926 137.7962963 ], Class Mean 1: [ 52.63432836 158.97761194]
```

4. (5%) Show the within-class scatter matrix $S_W$ of the training set.

```
With-in class scatter matrix:
[[ 19184.82283029 -16006.39331122]
 [-16006.39331122 106946.45135434]]
```

5. (5%) Show the between-class scatter matrix $S_B$ of the training set.

```
Between class scatter matrix:
[[ 17.01505494 -87.37146342]
 [-87.37146342 448.64813241]]
```

6. (5%) Show the Fisher’ s linear discriminant $w$ of the training set.

```
W:
[ 0.28737344 -0.95781862]
```

7. (10%) Obtain predictions for the testing set by measuring the distance between the projected value of the testing data and the projected means of the training data for the two classes. Show the accuracy score on the testing set. The accuracy score should be greater than 0.65.

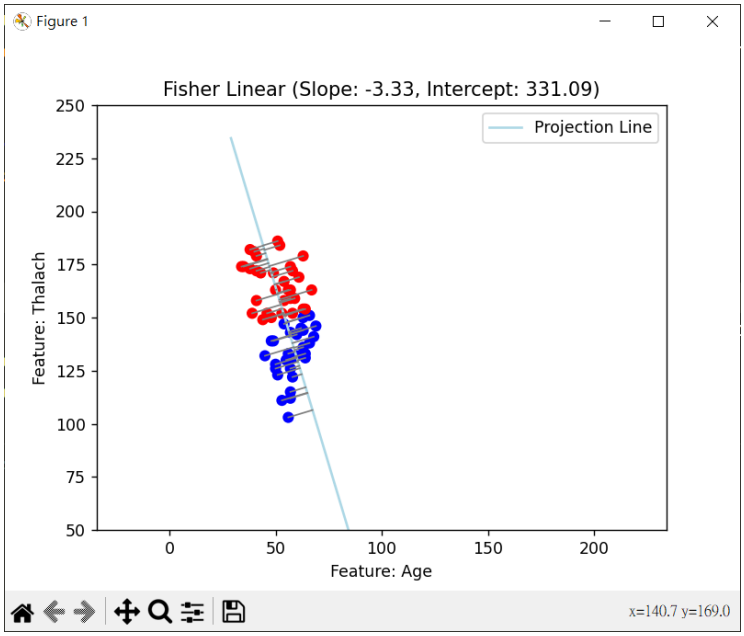
```
Accuracy of FLD: 0.6557377049180327
```

8. (10%) Plot the projection line (x-axis: age, y-axis: thalach).

   1) Plot the projection line trained on the training set and show the slope and intercept on the title (you can choose any value of intercept for better visualization).

   2) Obtain the prediction of the testing set, plot and colorize them based on the prediction.

   3) Project all testing data points on your projection line.




Part. 2, Questions (50%):

1. (5%) What' s the difference between the sigmoid function and the softmax function? In what scenarios will the two functions be used? Please at least provide one difference for the first question and answer the second question respectively.

Ans:

1.

sigmoid function: Single output, which is commonly used to indicate the probability of one of a class in two. The output range is between 0 and 1.

formula : $\sigma(x) = \frac{1}{1+e^{-x}}$

Softmax function: Multiple output, which is commonly used to indicate the probability distribution over multiple classes. Every number in the output distribution is between 0 and 1, and the sum of all the numbers are 1.

formula : $softmax(x_i) = \frac{e^{x_i}}{\sum_{j-1}^{k} e^{x_j}}$ where k is the number of classes.

2.

The sigmoid function is used for binary classification since the output of it is constrained between 0 and 1, which can be used to indicate the probability of the data belonging to class 1, and (1 – output ) as the probability belong of the data belonging to class 2. In contrast, the softmax function is applied when solving the multiclass classification problem, which can output the probability distribution over multiple classes.

2. (10%) In this homework, we use the cross-entropy function as the loss function for Logistic Regression. Why can’ t we use Mean Square Error (MSE) instead? Please explain in detail.

ANS:

1.

The goal of Logistic regression and FLD is to deal with the classification, in other words, we mainly care about the whether the classification is correct instead of the absolute probability difference between the prediction and the ground truth. The former can be achieved by cross entropy since the penalty of misclassification is gigantic due to the formula. In comparison, MSE can’ t give a effective feedback when the misclassification happened, so it’ s better to select the cross entropy as the loss function.

Eg. No matter the ground truth = 0 or 1, the MSE between the prediction = 0.5 and the ground truth the same.

2.

Due to the sigmoid function, the optimization problem will be non-convex, while MSE is designed for convex optimization. Knowing above, the ineffective and inefficient process of selecting MSE as the loss function can be predicted. Choose cross entropy will be better.

3. (15%) In a multi-class classification problem, assume you have already trained a classifier using a logistic regression model, which the outputs are P1, P2, ... Pc, how do you evaluate the overall performance of this classifier with respect to its ability to predict the correct class?

3.1. (5%) What are the metrics that are commonly used to evaluate the performance of the classifier? Please at least list three of them.

ANS：Confusion Matrix、Accuracy、Precision、Recall rate

Confusion matrix divides all the predictions into four types.

| | Actual Pos | Actual Neg |
|---|---|---|
| Predicted Pos | True Pos(TP) | False Pos(FP) |
| Predicted Neg | False Neg(FN) | True Neg(TN) |

Accuracy = (TP + TN) / (TP + TN + FN + TN), which represents the ratio of correctly predicted instances to the total instances.

Precision = TP / (TP + FP), which represent the accuracy of the positive prediction.

Recall Rate = TP / (TP + FN), which represent the sensitivity when the actual positive happens.

3.2. (5%) Based on the previous question, how do you determine the predicted class of each sample?

ANS：

If we take Accuracy above as the deterministic metric, the predicted class of each sample will be $\boldsymbol{argmax_i(P_i)}$, since $\boldsymbol{P_i}$ represent (TP + TN) / (TP + TN + FN + TN) here.

3.3. (5%) In a class imbalance dataset (say 90% of class-1, 9% of class-2, and 1% of class-3), is there any problem with using the metrics you mentioned above and how to evaluate the model prediction performance in a fair manner?

ANS：

Take accuracy for example. In the imbalance dataset above, if a model just simply predicts all of the data as class-1, the accuracy will also be 90%. However, the model does not do anything in real prediction aspect.

In a fair manner, it would be better to consider Precision also to increase the penalty of FN, which can better evaluate the effectiveness of the model.

4. (20%) Calculate the results of the partial derivatives for the following equations. (The first one is binary cross-entropy loss, and the second one is mean square error loss followed by a sigmoid function. σ is the sigmoid function.)

4.1. (10%)

$$\frac{\partial}{\partial x}\left(-t * \ln\left(\sigma(x)\right) - (1-t) * \ln\left(1-\sigma(x)\right)\right)$$

4.2. (10%)

$$\frac{\partial}{\partial x}\left(\left(t-\sigma(x)\right)^2\right)$$

**ANS:**

1. $\frac{\partial}{\partial x}(-t \cdot \ln(\sigma(x)) - (1-t) \cdot \ln(1-\sigma(x)))$

$= \frac{-t}{\sigma(x)} \cdot \sigma(x)(1-\sigma(x)) - \frac{(1-t)}{1-\sigma(x)} \cdot (-\sigma(x)(1-\sigma(x)))$

$= -t(1-\sigma(x)) + (1-t)\sigma(x)$

$= -t + t\sigma(x) + (1-t)\sigma(x)$

$= \sigma(x) - t$ — Ans

$\sigma(x) = \frac{1}{1+e^{-x}}$

$\frac{d\sigma(x)}{dx} = \frac{-(-e^{-x})}{(1+e^{-x})^2}$

$= \frac{1}{1+e^{-x}} \times \frac{e^{-x}}{1+e^{-x}}$

$= \frac{1}{1+e^{-x}}\left(\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right)$

$= \sigma(x)(1-\sigma(x))$

2. $\frac{\partial}{\partial x}((t-\sigma(x))^2)$

$= 2 \cdot (t-\sigma(x)) \cdot -\sigma(x)(1-\sigma(x))$

$= -2\sigma(x)(t-\sigma(x))(1-\sigma(x))$ — Ans