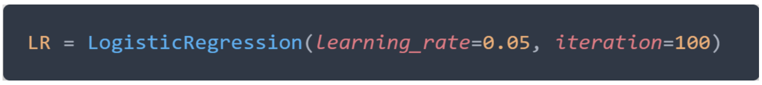
NYCU Intro. to ML
HW2 Report

110550093, 蔡師睿

Part I. Coding

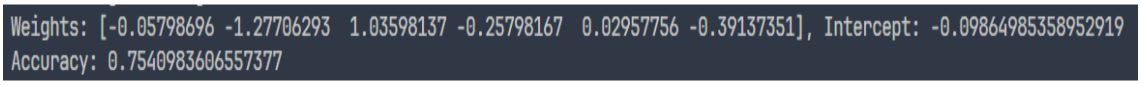
Logistic Regression

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



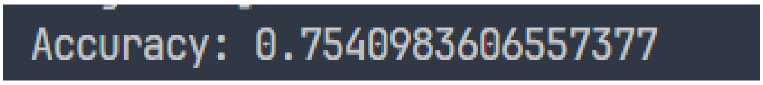
```
LR = LogisticRegression(learning_rate=0.05, iteration=100)
```

2. Show the weights and intercept of your model.



```
Weights: [-0.05798696 -1.27706293  1.03598137 -0.25798167  0.02957756 -0.39137351], Intercept: -0.09864985358952919
Accuracy: 0.7540983606557377
```

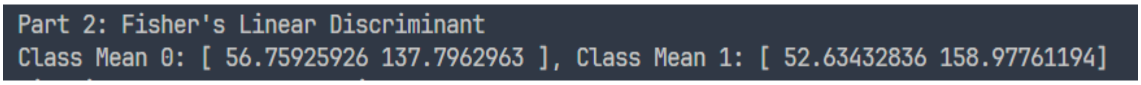
3. Show the accuracy score of your model on the testing set. The accuracy score should be greater than 0.75.



```
Accuracy: 0.7540983606557377
```

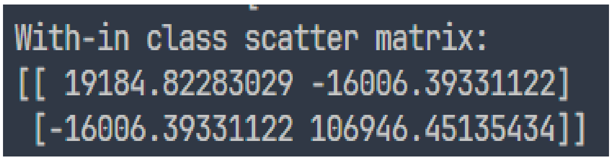
Fisher's Linear Discriminant (FLD)

4. Show the mean vectors m_i (i=0, 1) of each class of the training set.



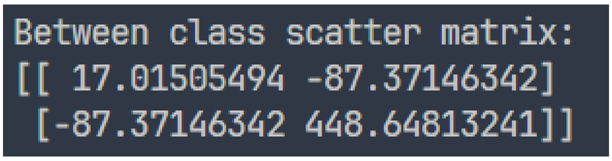
```
Part 2: Fisher's Linear Discriminant
Class Mean 0: [ 56.75925926 137.7962963 ], Class Mean 1: [ 52.63432836 158.97761194]
```

5. Show the within-class scatter matrix S_W of the training set.



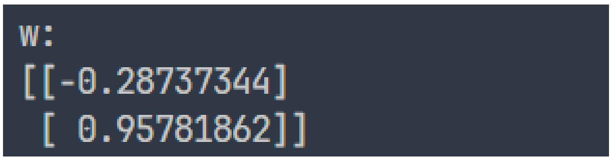
```
With-in class scatter matrix:
[[ 19184.82283029 -16006.39331122]
 [-16006.39331122 106946.45135434]]
```

6. **Show the between-class scatter matrix S_B of the training set.**



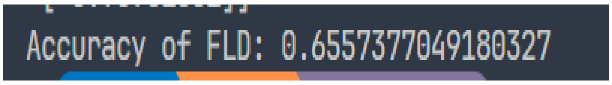
```
Between class scatter matrix:
[[ 17.01505494 -87.37146342]
 [-87.37146342 448.64813241]]
```

7. **Show the Fisher's linear discriminant w of the training set.**



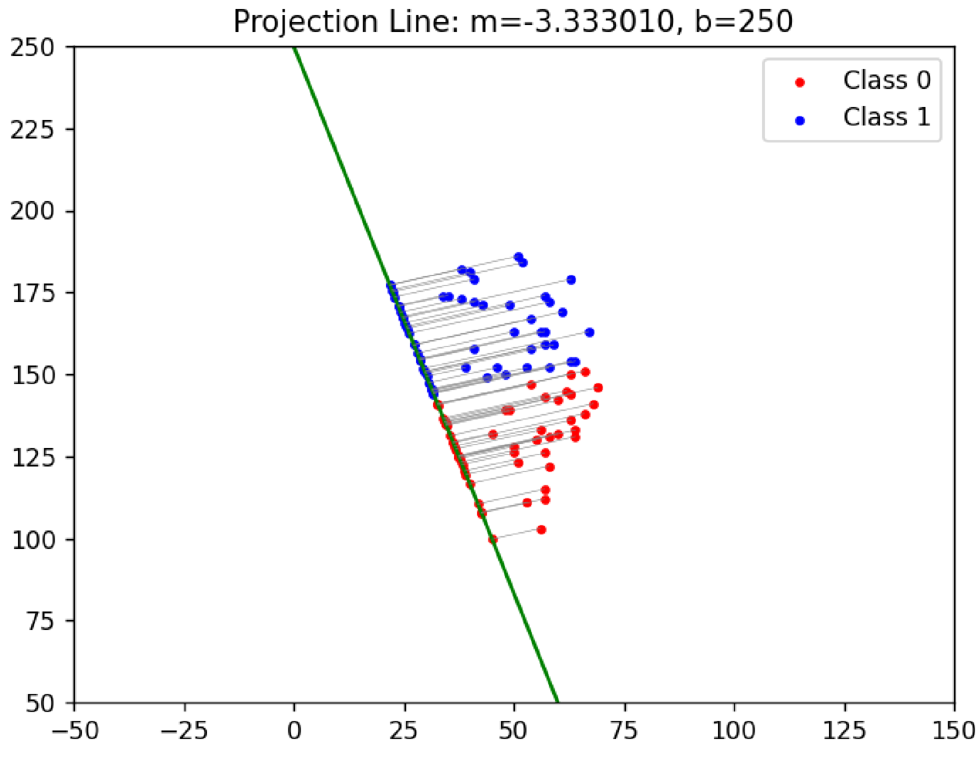
```
w:
[[-0.28737344]
 [ 0.95781862]]
```

8. **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
```

9. **Plot the projection line (x-axis: age, y-axis: thalach).**



Part II. Questions

1. What's the difference between the sigmoid function and the softmax function? In what scenarios will the functions be used?

The equations have distinct forms. They are as follows:

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

$$Softmax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}$$

As we can see, sigmoid function receives only one input, and returns the number between 0 and 1 which represents the probability of X of belonging to a certain class. While the other hand the input of softmax function is vectorized, and the meaning of its output is a probability distribution over K different classes, where K is the number of classes.

To summarize, sigmoid function is used for binary classification, while softmax function is used for multi-class classification problems.

2. 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?

We use sigmoid function in logistic regression, and therefore, to compute its gradient, we have to do the partial derivative based on its loss function. If we use MSE as our loss function, the equation will be:

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

, where t denotes the ground truth label, and $\sigma(x)$ denotes the sigmoid function. The resulting derivative simplifies to $\frac{2\sigma^2(x)}{e^x}(\sigma(x) - t)$ (refer to eq.7 on page 6 for more details). When x is extremely small, which approaches negative infinity, the gradient becomes 'nan', causing the model to fail to converge.

However, if we use cross-entropy loss, the partial derivative will be given by:

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

The resulting derivative simplifies to $\sigma(x) - t$ (refer to eq.6 on page 6 for more details). We can thereby demonstrate that cross-entropy loss is more suitable in this case compared to MSE loss.

3. 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. What are the metrics that are commonly used to evaluate the performance of the classifier?

1. Accuracy

$$accuracy = \frac{1}{n}\sum_{i=1}^{n} f(x), \quad where\ f(x) = \begin{cases} 1, & if\ output = label \\ 0, & otherwise \end{cases}$$

2. F1-score

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

3. ROC AUC score: ROC curve is a grphical representation of a classifier's problem. It's a chart that visualizes the tradeoff between true positive rate and false positive rate. The ROC AUC score is the area under the ROC curve. The more top-left your curve is the higher the area and hence higher ROC AUC score.

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

1. Probability thresholding: If it's a binary classification problem and the output of the model is the probabilities of the two classes which sum is equal to one, we can apply threshold. For example, if the threshold is set at 0.5, any sample predicted with a probability higher than 0.5 is categorized into that specific class.
2. Argmax function: If it's a multi-class classification problem and the output of the model is the probabilities of each classes which sums is equal to one, apply argmax function to determine the class with the most higher probability as the predicted class for that sample.

3.3. 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?

Yes, there is a problem in the accuracy evaluation metric mentioned above. For example, if a model can only predict class-1 no matter what the input is, the accuracy of the dataset described in the question will become 90% high. Obviously, accuracy is not a proper evaluation metric to evaluate the performance of the classifier in this case.

If we want to evaluate the model prediction performance in a fair manner, we can introduce the evaluation metrics which consider true positive, false positive, true negative, and false negative. F1-score and ROC AUC score mentioned above are these proper evaluation metrics.

4. Calculate the results of the partial derivatives for the following equations.

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

Sigmoid function:

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

First, compute the partial derivatives of $\sigma(x)$ and $1-\sigma(x)$:

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

$$\frac{\partial}{\partial x}(1-\sigma(x)) = \frac{\partial}{\partial x}(1-\frac{1}{1+e^{-x}}) = \frac{\partial}{\partial x}(\frac{1}{1+e^{x}}) = \frac{-e^{x}}{(1+e^{x})^2} \tag{3}$$

Second, we can calculate the partial derivatives of $\ln(\sigma(x))$ and $\ln(1-\sigma(x))$:

$$\begin{aligned}\frac{\partial}{\partial x}\ln(\sigma(x)) &= \left(\frac{\partial}{\partial\sigma(x)}\ln(\sigma(x)\right)\left(\frac{\partial\sigma(x)}{\partial x}\sigma(x)\right) = \frac{1}{\sigma(x)}\frac{e^{-x}}{(1+e^{-x})^2}\\ &= (1+e^{-x})\frac{e^{-x}}{(1+e^{-x})^2} = \frac{e^{-x}}{1+e^{-x}}\end{aligned} \tag{4}$$

$$\begin{aligned}&\frac{\partial}{\partial x}\ln(1-\sigma(x)) = \left(\frac{\partial}{\partial(1-\sigma(x))}\ln(1-\sigma(x)\right)\left(\frac{\partial(1-\sigma(x))}{\partial x}(1-\sigma(x)\right)\\ &= \frac{1}{1-\sigma(x)}\frac{-e^{x}}{(1+e^{x})^2} = \frac{1+e^{-x}}{e^{-x}}\frac{-e^{x}}{(1+e^{x})^2} = (1+e^{x})\frac{-e^{x}}{(1+e^{x})^2} = \frac{-1}{1+e^{-x}}\end{aligned} \tag{5}$$

Finally, get the result of the partial derivative from eq. 4 and 5:

$$\begin{aligned}
&\frac{\partial}{\partial x}(-t\ln(\sigma(x)) - (1-t)\ln(1-\sigma(x))) \\
&= -t\frac{\partial}{\partial x}\ln(\sigma(x)) - (1-t)\frac{\partial}{\partial x}\ln(1-\sigma(x)) \\
&= -t\frac{e^{-x}}{1+e^{-x}} - (1-t)\frac{-1}{1+e^{-x}} \\
&= \frac{1-t(1-e^{-x})}{1+e^{-x}} = \frac{1}{1+e^{-x}} - t \\
&= \sigma(x) - t
\end{aligned} \tag{6}$$

Result: $\frac{\partial}{\partial x}(-t\ln(\sigma(x)) - (1-t)\ln(1-\sigma(x))) = \sigma(x) - t$

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

Based on eq. 1 and 2:

$$\begin{aligned}
&\frac{\partial}{\partial x}((t-\sigma(x))^2) = \frac{\partial}{\partial x}\left(t - \frac{1}{1+e^{-x}}\right)^2 = 2\left(t - \frac{1}{1+e^{-x}}\right)\left(\frac{-e^{-x}}{(1+e^{-x})^2}\right) \\
&= -2\left(\frac{te^{-x}}{(1+e^{-x})^2} - \frac{e^{-x}}{(1+e^{-x})^3}\right) = \frac{-2e^{-x}}{(1+e^{-x})^2}\left(t - \frac{1}{1+e^{-x}}\right) \\
&= \frac{2\sigma^2(x)}{e^x}\left(\sigma(x) - t\right)
\end{aligned} \tag{7}$$

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