NYCU Introduction to Machine Learning, Homework 2

110550080, 何曉嫻

Part. 1, Coding (50%):

(15%) Logistic Regression

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

```
LR = LogisticRegression(learning_rate=0.02, iteration=5000)
```

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

```
Part 1: Logistic Regression
Weights: [-0.04937657 -0.39527957  0.83583958  0.01939349  0.02807955 -0.39894947], Intercept: -0.8256588387635152
```

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
```

(40%) Linear Regression Model - Gradient Descent Solution

1. (0%) Show the mean vectors mi (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]
```

2. (5%) Show the within-class scatter matrix SW of the training set.

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

3. (5%) Show the between-class scatter matrix SB of the training set.

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

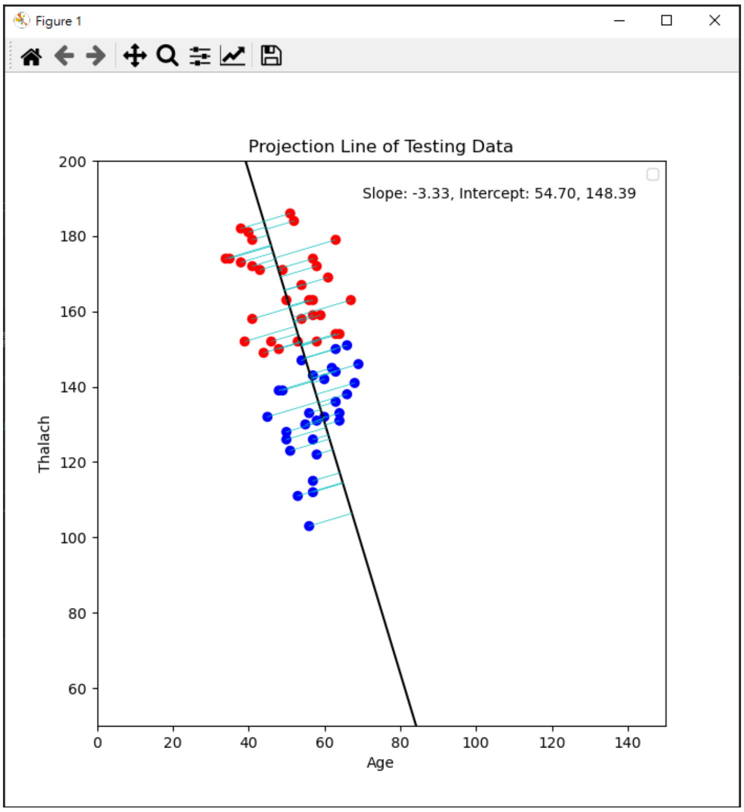
4. (5%) Show the Fisher's linear discriminant w of the training set.

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

5. (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
```

6. (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. Your result should look like the below image.



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.
   Sigmoid function compresses a single input into a range between 0 and 1 and produce either 0 or 1 as its output depending on whether it is being used for two-class classification. It is often used in two-class classification cases.
   Softmax function compresses multiple inputs into a range between 0 and 1 and produce any number of classes as its output depending on whether it is being used for multi-class classification. It is often used in multi-class classification cases.

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.
   When the MSE loss function is plotted with respect to weights of the logistic regression model, the curve obtained is not a convex curve which makes it very difficult to find the global minimum. Besides, the output of logistic regression is a probability between 0-1. The actual target value is either 0 or 1 in classification problems. So the $(y-p)^2$ will always be between 0 and 1, making it very difficult to track the progress of error value as it is hard to store high precision floating-point numbers.

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.
   (1) Accuracy
   (2) Confusion matrix
   (3) Precision and Recall

   3.2. (5%) Based on the previous question, how do you determine the predicted class of each sample?
   The logistic regression model outputs probabilities for each class $P1, P2, \ldots, Pc$ for a given input sample. These probabilities represent the likelihood of the sample belonging to each class. The predicted class is the one with the highest predicted probability. Mathematically, it is determined by $\hat{y} = \text{argmax}_i\ P_i$ argmax_i represents the index of the maximum value among outputs probabilities, and $\hat{y}$ is the predicted class..
   In two-class cases, use a threshold to determine the predicted class. For instance, if P_i>threshold → $\hat{y}$ =1, predict to class 1.

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?

In a class-imbalanced dataset, using metrics like accuracy might be misleading because a model could achieve high accuracy by simply predicting the majority class. First, accuracy might be high due to the dominance of the majority class. Second, Metrics like precision and recall might neglect minority classes.

To evaluate in a fair manner, we can use Confusion Matrix to understand the model's performance on each class. Or use F1-Score especially macro or weighted variants, to considers precision and recall in a balanced manner.

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%)

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

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

$$= \frac{d}{dx}\left(-t \times \ln\left(\frac{1}{1+e^{-x}}\right) - (1-t) \times \ln\left(1-\frac{1}{1+e^{x}}\right)\right) \qquad u = \frac{1}{1+e^{-x}}, \quad w = \frac{e^{x}}{1+e^{x}}$$

$$= -t\left(\frac{d}{du}\ln u \times \frac{d}{dx}\frac{1}{1+e^{-x}}\right) + (t-1)\left(\frac{d}{dw}\ln w \times \frac{d}{dx}\frac{e^{x}}{1+e^{x}}\right)$$

$$= -t \cdot \frac{1}{u} \cdot \frac{d}{dx}\frac{1}{1+e^{x}} + (t-1) + (t-1)\left(\frac{1}{w} \cdot \frac{d}{dx}\frac{e^{x}}{1+e^{x}}\right)$$

$$= -t\,(1+e^{x})\left(+(1+e^{x})^{-2} \cdot e^{x}\right) + (t-1)\,\frac{1+e^{x}}{e^{x}}\;\frac{-e^{x}(1+e^{x})+e^{x}(0+e^{x})}{(1+e^{x})^{2}}$$

$$= -t\,(1+e^{x})^{-1} \cdot e^{x} + (t-1)\,\frac{-1}{1+e^{x}}$$

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

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

$$= \sigma(x) - t \quad ※$$

4.2 (10%)

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

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

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

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

$$= -2t \cdot \frac{+e^{-x}}{+(1+e^{-x})^2} + \frac{\cancel{+2e^{-x}+2e^{-2x}}\ 2e^{-x}(\cancel{1+e^{-x}})}{+(1+e^{-x})^{\cancel{4}\,3}}$$

$$= \frac{-2t(1+e^{-x})\,e^{-x} + 2e^{-x}}{(1+e^{-x})^3}$$

$$= \frac{-2[\,t e^{-2x} + t e^{-x} + e^{-x}\,)}{(1+e^{-x})^3}$$

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