NYCU Introduction to Machine Learning, Homework 2

110550167 侯博軒

Part. 1, Coding (50%):

In this coding assignment, you are requested to implement Logistic Regression and Fisher's Linear Discriminant by using only Numpy. After that, train your model on the provided dataset and evaluate the performance on the testing data.

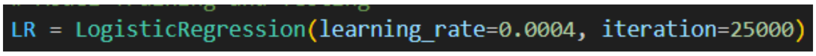
(15%) Logistic Regression

Requirements:

- Use Gradient Descent to update your model
- Use CE (Cross-Entropy) as your loss function.

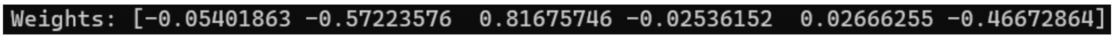
Criteria:

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

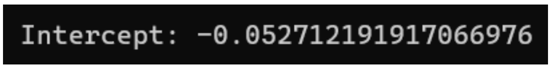


```
LR = LogisticRegression(learning_rate=0.0004, iteration=25000)
```

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

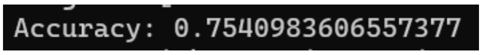


```
Weights: [-0.05401863 -0.57223576  0.81675746 -0.02536152  0.02666255 -0.46672864]
```



```
Intercept: -0.052712191917066976
```

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
```

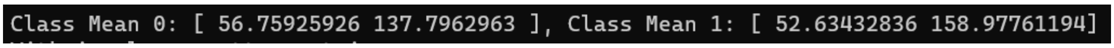
(35%) Fisher's Linear Discriminant (FLD)

Requirements:

- Implement FLD to reduce the dimension of the data from 2-dimensional to 1-dimensional.

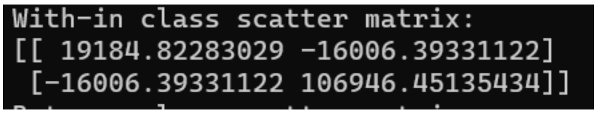
Criteria:

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



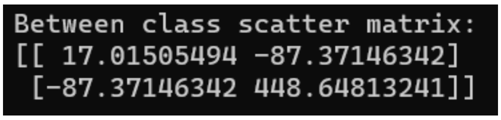
```
Class Mean 0: [ 56.75925926 137.7962963 ], Class Mean 1: [ 52.63432836 158.97761194]
```

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



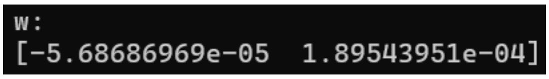
```
With-in class scatter matrix:
[[ 19184.82283029 -16006.39331122]
 [-16006.39331122 106946.45135434]]
```

6. (5%) Show the between-class scatter matrix SB of the training set.



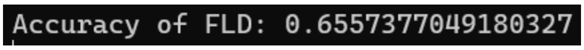
```
Between class scatter matrix:
[[ 17.01505494 -87.37146342]
 [-87.37146342 448.64813241]]
```

7. (5%) Show the Fisher’s linear discriminant w of the training set.



```
w:
[-5.68686969e-05  1.89543951e-04]
```

8. (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.



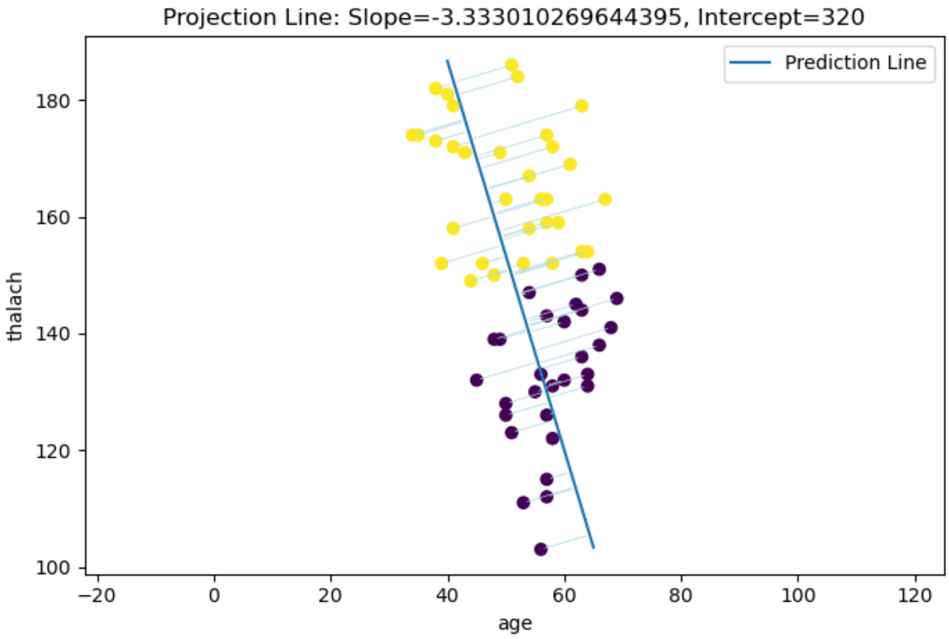
```
Accuracy of FLD: 0.6557377049180327
```

9. (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.

   (1) One key difference between the two functions is that the sigmoid function produces independent probabilities according to the raw input vector, while softmax function produces correlated probabilities according to the raw input vector, the sum of its output probabilities are always 1.

   (2) According to the above mentioned characteristic of the two activation function, sigmoid function is better used on binary classification problems where the output only has 2 opposite options.
   On the other hand, softmax function would be preferred when dealing with multi-class classification problem, where the output has multiple options and while the likelihood of one class increase the likelihood of other classes would decrease accordingly.

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.

We use MSE as our lost function on the linear regression homework and cross-entropy on logistic regression because we are dealing with two different tasks with regards to different aspects of accuracy. In linear regression we want the output to be close to our ground truth, as for logistic regression we want our model to give a higher probability on the correct label. According to these two different criteria we discuss why one function is better than the other:

The cross-entropy function is a measure of the difference between two probability distributions. It is particularly useful for classification problems because it penalizes the model more heavily for making incorrect predictions with high confidence.

The MSE function is a measure of the difference between two continuous variables. It is particularly useful for regression problems because it penalizes the model more heavily for large errors. However, it may be not sensitive enough to deal with un-continuous variables, which in the logistic regression case is the 2 class labels of the input. Resulting in the lack of loss when the model gives a wrong prediction with high confidence, slowing down the training process. Also, we want the prediction to be accurate instead of being numerically close to the true label. Therefore, we should not choose MSE as our lost function for classification problems.

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?

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 – Measure the overall correctness of the classifier by calculating the ratio of correctly predicted labels over total amount of samples.

   (2) F1 score – The F1 score ranges from 0 to 1, with higher values indicating better performance.
   F1 Score = 2*(Precision * Recall) / (Precision + Recall)
   Precision = true positive prediction / total positive prediciton
   Recall = true positive prediction / total true positive

   (3) Confusion Matrix – For a m class classification, the confusion matrix is a m*m matrix indexed through (true label, predicted label).

2. (5%) Based on the previous question, how do you determine the predicted class of each sample?

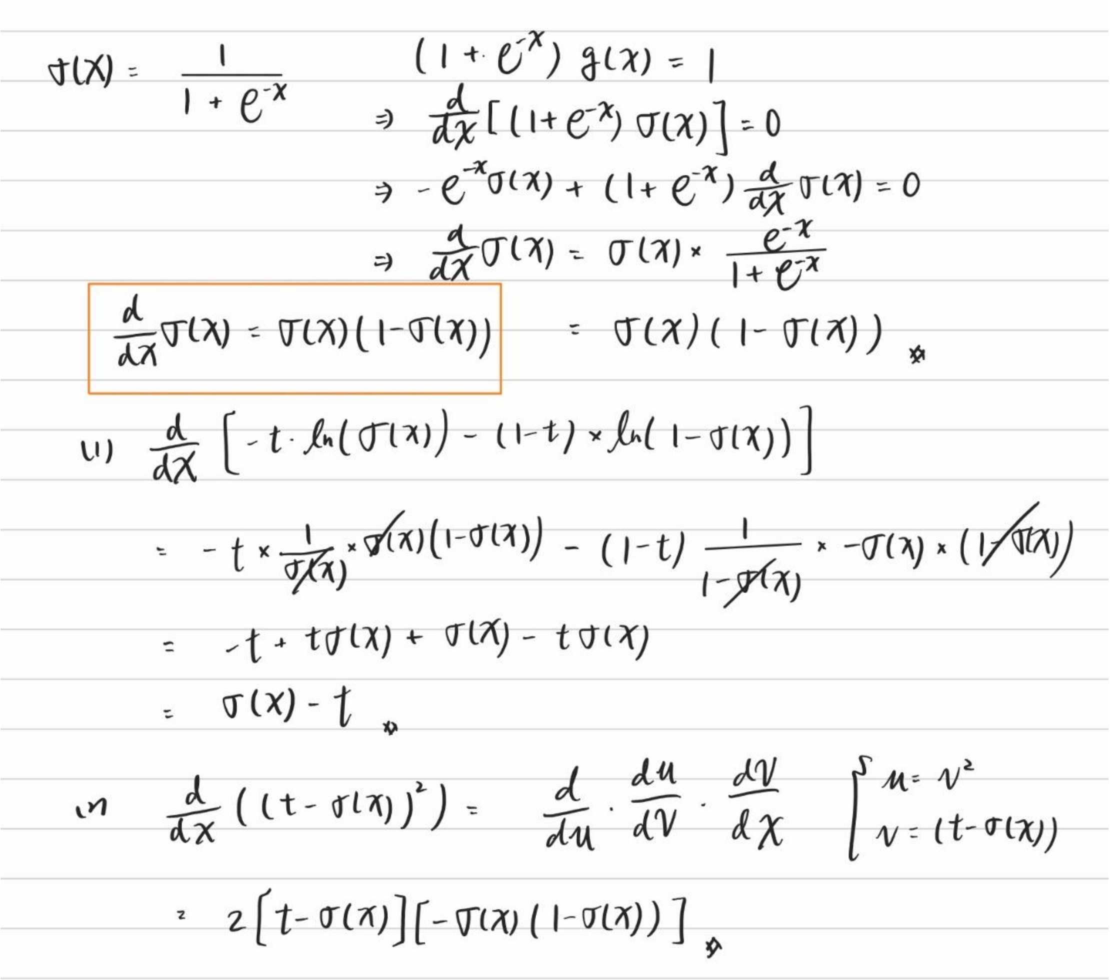
   Based on the prediction probability P1, P2, .. Pc, the class with the highest probability is chosen as the predicted class of the testing sample.

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 imbalance dataset, using metrics like accuracy can be misleading because the classifier may achieve high accuracy by simply predicting the majority class. To address this problem, we have to choose cautiously on which evaluation metric to use.

   The F1 score manner could give a fairer insight into imbalance dataset evaluation. Because F1 score considers both false positives and false negatives, it is less likely to be inflated by the true negatives of the majority class in imbalance dataset. The confusion matrix could also be a good indicator when the dataset is imbalance. From indents of the matrix we can also calculate the false negative rate. Showing the error on smaller classes if they were to be predicted as the major class.

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.)



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

$$(1+e^{-x})\, g(x) = 1$$

$$\Rightarrow \frac{d}{dx}\left[(1+e^{-x})\,\sigma(x)\right] = 0$$

$$\Rightarrow -e^{-x}\sigma(x) + (1+e^{-x})\frac{d}{dx}\sigma(x) = 0$$

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

$$\boxed{\frac{d}{dx}\sigma(x) = \sigma(x)(1-\sigma(x))} \qquad = \sigma(x)(1-\sigma(x))$$

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

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

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

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

(2) $$\frac{d}{dx}\left((t-\sigma(x))^2\right) = \frac{d}{du}\cdot\frac{du}{dv}\cdot\frac{dv}{dx} \qquad \begin{cases} u = v^2 \\ v = (t-\sigma(x)) \end{cases}$$

$$= 2\left[t-\sigma(x)\right]\left[-\sigma(x)(1-\sigma(x))\right]$$