NYCU Introduction to Machine Learning, Homework 2

Part. 1, Coding (50%):

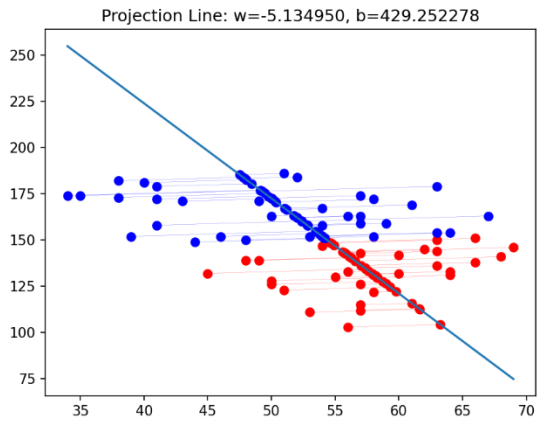
Show the hyperparameters (learning rate and iteration) used in Logistic Regression

```
LR = LogisticRegression(learning_rate=0.0001, iteration=100000)
```

Output

```
Part 1: Logistic Regression
Weights: [-0.05401261 -0.57194471  0.81540993 -0.02539069  0.02665697 -0.46607183], Intercept: -0.05272493538773201
Accuracy: 0.7540983606557377
Part 2: Fisher's Linear Discriminant
Class Mean 0: [ 56.75925926 137.7962963 ], Class Mean 1: [ 52.63432836 158.97761194]
With-in class scatter matrix:
[[ 19184.82283029 -16006.39331122]
 [-16006.39331122 106946.45135434]]
Between class scatter matrix:
[[ 17.01505494 -87.37146342]
 [-87.37146342 448.64813241]]
w:
[ 0.28737344 -0.95781862]
Accuracy of FLD: 0.6557377049180327
```

Plot the 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.

   Sigmoid takes in a single value and maps it to another value between 0~1. It operates independently on the probability of a specific class. While Softmax

takes a vector of values and converts them into a probability distribution over multiple classes, so the sum of these probabilities for all classes adds up to 1.

As a result, sigmoid is used for binary classification. The prediction is determined by comparing the probability with a threshold. In contrast, softmax is used for multi-class classification, where the class with the highest probability becomes the output prediction.

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.

The sigmoid function in logistic regression introduces non-linear transformation by converting the linear combination of inputs and weights into a probability value between 0~1. This non-linearity adds complexity to the relationship between the weight parameters and the error. If MSE is applied to logistic regression and plotted in terms of the model's weights, the resulting curve is not convex. That is, it has multiple local minima, making it more difficult to determine the optimal weight values.

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.

I. Accuracy: the ratio of correctly classified instances to the total number of instances
II. Precision: the ratio of true positive predictions to the total positive predictions (true positives + false positives)
III. Recall: the ratio of true positive predictions to the total actual positives (true positives + false negatives)

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

Among the probability distribution over all classes (P1, P2, ..., Pc), select the one with the highest probability as the predicted class.

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?

Among the metrics I mentioned above, “Accuracy” will have some problems in an imbalanced dataset. Taking the scenario described in the statement for example, predicting all samples as class-1 can achieve a high accuracy score. However, this approach would perform poorly for class-2 and class-3. This can be solved by adjusting class weight. By increasing the weight of the minority classes and decreasing the weight of the majority ones, we can make the model more sensitive to the minority classes.

“Precision” and “Recall” don’t have such problem since they have taken the minority classes into account.

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%)

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

$$= \frac{\partial}{\partial x}(-t \cdot \ln(\sigma(x))) - \frac{\partial}{\partial x}((1-t)\cdot \ln(1-\sigma(x)))$$
$$= -t \cdot \left(\frac{1}{\sigma(x)} \cdot \sigma(x)\cdot(1-\sigma(x))\right) - (1-t)\cdot\left(\frac{-1}{1-\sigma(x)}\cdot\sigma(x)\cdot(1-\sigma(x))\right)$$
$$= -t + t\sigma(x) + \sigma(x) - t\sigma(x)$$
$$= \sigma(x) - t \quad \#$$

4.2. (10%)

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

$$= 2(t-\sigma(x))\cdot\frac{\partial}{\partial x}(t-\sigma(x))$$
$$= -2(t-\sigma(x))\cdot\sigma(x)\cdot(1-\sigma(x)) \quad \#$$

$$\Delta\sigma'(x) = \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right)$$
$$= \frac{-(-e^{-x})}{(1+e^{-x})^2}$$
$$= \frac{1}{1+e^{-x}}\cdot\frac{(1+e^{-x})-1}{1+e^{-x}}$$
$$= \sigma(x)\cdot(1-\sigma(x))$$