NYCU Pattern Recognition, Homework 2

Deadline: Apr. 08, 23:59

Part. 1, Coding (70%):

In this coding assignment, you are requested to implement 1) logistic Regression and 2) Fisher's Linear Discriminant by using only Numpy, then train your model on the provided dataset and finally evaluate the performance on testing data. Please train your logistic regression model using Gradient Descent, not the closed-form solution.

(20%) Logistic Regression Model

Requirements:

- Use Gradient Descent
- Use CE (Cross-Entropy) as your loss function.
- Use Softmax for this multiclass classification task.

Criteria:

1. (0%) Show the learning rate, epoch, and batch size that you used.
2. (5%) What's your training accuracy?
3. (5%) What's your testing accuracy?
4. (5%) Plot the learning curve of the training. (x-axis=epoch, y-axis=loss)
5. (5%) Show the confusion matrix on testing data.

(30%) Fisher's Linear Discriminant (FLD) Model

Requirements:

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

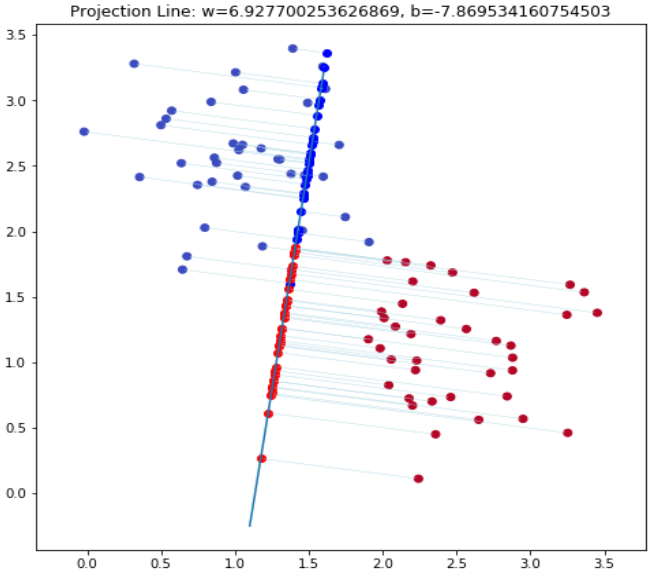
Criteria:

6. (2%) Compute the mean vectors m_i (i=1, 2, 3) of each class on training data.
7. (2%) Compute the within-class scatter matrix S_W on training data.
8. (2%) Compute the between-class scatter matrix S_B on training data.
9. (4%) Compute the Fisher's linear discriminant w on training data.
10. (8%) Project the testing data to get the prediction using the shortest distance to the class mean. Report the accuracy score and draw the confusion matrix on testing data.
11. (8%) Project the testing data to get the prediction using K-Nearest-Neighbor. Compare the accuracy score on the testing data with K values from 1 to 5.
12. (4%)

    1) Plot the best projection line on the training data and show the slope and intercept on the title *(you can choose any value of intercept for better visualization)*

    2) colorize the training data with each class

    3) project all training data points on your projection line. Your result should look like the below image (This image is for reference, not the answer)



(20%) Train your own model

Requirements:

- Using another dataset that we provided (a real-world dataset).
- Train your model (FLD or Logistics Regression model that you implemented above).
- Try different parameters and feature engineering to beat the baseline.
- Save your testing predictions in the CSV file.

Criteria:

13. Explain how you chose your model and what feature processing you have done in detail. Otherwise, no points will be given.

Point	Accuracy
20	testing acc > 0.921
15	0.91 < testing acc <= 0.921
8	0.9 < testing acc <= 0.91
0	testing acc <= 0.9

Part. 2, Questions (30%):

(6%) 1. Discuss and analyze the performance

a) between Q10 and Q11, which approach is more suitable for this dataset. Why?

b) between different values of k in Q11. (Which is better, a larger or smaller k? Does this always hold?)

(6%) 2. Compare the sigmoid function and softmax function.

(6%) 3. Why do we use cross entropy for classification tasks and mean square error for regression tasks?

(6%) 4. In Q13, we provide an imbalanced dataset. Are there any methods to improve Fisher Linear Discriminant's performance in handling such datasets?

(6%) 5. 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.)

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

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