IE343: Statistical Machine Learning

Assignment 2

Professor: Seyoung Yun

(TA : Mingyu Kim)

1 Instruction

This assignment is to implement "Fisher's linear discriminant analysis" (called LDA) and "logistic regression". The task is binary classification $x \in \mathbb{R}^2, y \in \{0, 1\}$. You write a code for implementing the '_fit' function in the LDA `(./App/fishers_linear_discriminant.py)` and the logistic regression `(./App/logistic_regressior.py)`. To help your understanding, we provide a complete source code of the least square method for the same classficiation task `(./Least_square.py)`. Some of details of source codes are as follows :

* There are "App", "Data" and "Result" directories and `(./Least_square.py, ./LDA.py, ./logistic_regression.py)`. Students have to run these '.py' files to make sure that LDA and logistic regression have no errors. After running these files, you can get a BCE and its standard deviation.
* When it comes to "App" directory, there are "Pre_processing" directory and several source codes such as data_import, evaluation. The important thing is that you should complete the "_fit" methods which can properly work `(./App/fishers_linear_discriminant.py)` `(./App/logistic_regressior.py)`.
* LDA method requires gaussian distribution. 'Maximum likelihood estimation" method helps you to obtain the parameters of gaussian distribution `(./App/Pre_processing/gaussian.py)`. In order for LDA to work properly, threshold is very important. In this assignment, we can determine this value by 'minimizing the expected loss", which is already implemented at "_threshold" method. When you run `./LDA.py`, minimizing expected loss will be automatically executed.
* Logistic regression requires numerically approximate solutions named as "gradient descent". This method is impelmented in `(./App/Pre_processing/optimizer.py)`. Before using this method, you should explicitly figure out the gradient of "weights" w.r.t loss function.
* To help your understanding, all _fit methods are partially written. Most of written sections are for data handling. Therefore, you can concentrate on writing the main-algorithms.
* In this assignment, we only consider a binary classification task. However, if you want to use these methods for multiple classification task, all _fit methods requires one-hot encoded target variables. You can easily transform discrete target variables to one-hot vectors by using the phaser `(./App/Pre_processing/label_transformer.py)`.

In this assignment, we need to describe the model that best fits the entire data set. For evaluation, we provide the Binary Cross Entropy Method (BCE). Like assignment 1, average BCE and its standard errors for the entire data set are also provided. Furthermore, you need to visualize training and testing data with the decision boundary. `(./Results/result_(i).png)`
Therefore, you answer the questions in English and submit a report as a PDF file with your completed source code. The report should includes theoretical background for LDA and logistic regression and your opinion about each model. (Please write a report concisely). In addition, you have to complete this source codes in the attached python files.

2 Data

As we mentioned before, all data is generated by specific data generation process, but it is not released for students. There are 10 data sets. First 5 data sets consists of 50 training points and 20 testing data points. First 5 data sets was generated without outliers, but last 5 data sets consists of 60 training points and 25 testing points, which have outliers. In this assignment, students can study how much the model is sensitive to outliers. The classifiers should be trained by only one of data sets. After training your model, you should compare binary 'y' values to fitted values in terms of binary cross entropy (BCE) `(By default, all metrics are given)`.

* The first 5 data sets have 50 training data and 20 test data, but the last 5 data sets have 60 training data and 25 test data with outliers.
* This data comes from a specific data generation process, but it is not released.
* You shoud use same numbered data sets while training and testing.

3 Implementation

This code should be written using the 'Numpy' package. It is not allowed to use other packages. (Scipy, Tensorflow, pytorch and etc.) You have to write down appropriate methods ('_fit' method) in `(./App/fishers_linear_discriminant.py)` `(./App/logistic_regressior.py)`. For your understanding, we provide least square method for classification. You can see how it works by running `(./Least_square.py)`. Except for "_fit" methods, you don't need to modify any source code.

4 Question

A. Complete the '_fit' method in `(./App/fishers_linear_discriminant.py)`. In case of LDA, m_0, m_1 are the average values of X in class 0 and class 1. cov_inclass is the summation of covariance of X in class 0 and covariance of X in class 1. You should attach the picture indicating '_fit' method.

B. Complete the '_fit' method in `(./App/logistic_regressior.py)`. In case of logistic regression, *grad* is the gradient of weights w.r.t the loss function (cross-entropy). You should attach the picture indicating '_fit' method.

C. Write down mathematical derivation of linear discriminative analysis and its assumptions. You also need to describe how to figure out parameters, "w".

D. Write down the reason $p(\mathcal{C}_1|x) = \sigma(w^T x)$ in logistic regression and necessarity of assumptions. Also, you need to describe loss function(likelihood) and its gradients. Lastly, if we have gradients, explain how to figure out the parameters, "w" by using gradient descent

E. After applying LDA, write average BCE and its standard error and make a plot with train, test data and decision boundary in any data set. Remember that you need to attach two kinds of results and plots, such as, with outliers as well as without outliers. You need to check how LDA works when you train and test the model using data set $5 \sim 9$. Compare the result from data set $0 \sim 4$ with the result from data set $5 \sim 9$. Also explain your opinion about LDA method in terms of sensitivity.

F. After applying logistic regression, write average BCE and its standard error and make a plot train, test data and decision boundary in any data set. Remember that you should attach two kinds of results and plots with outliers as well as without outliers. You need to check how logistic regression works when you train and test the model using data set $5 \sim 9$. Compare the result from data set $0 \sim 4$ with the result from data set $5 \sim 9$. Also explain your opinion about logistic regression in terms of sensitivity.

G. If possible, please suggest appropriate polynomial engineering to improve the performance of logistic regression. Which dimensions of features is the best for this data set. Also, please describe your opinions with graphs, average BCE and its error.