Homework Assignment 3

Yuanjian Zhou
ECON 425

March, 5, 2020

Problem 1 Logistic Regression

Step 1

Instruction: Generate data (provided) and split it into training and testing subsets. You will need to write your codes to do the split. Then, the code will display the splitting results.
Results: Split codes are in source code file, graphic results see below.(Because there are lots of random generation process, the graphs for each trial will be different)

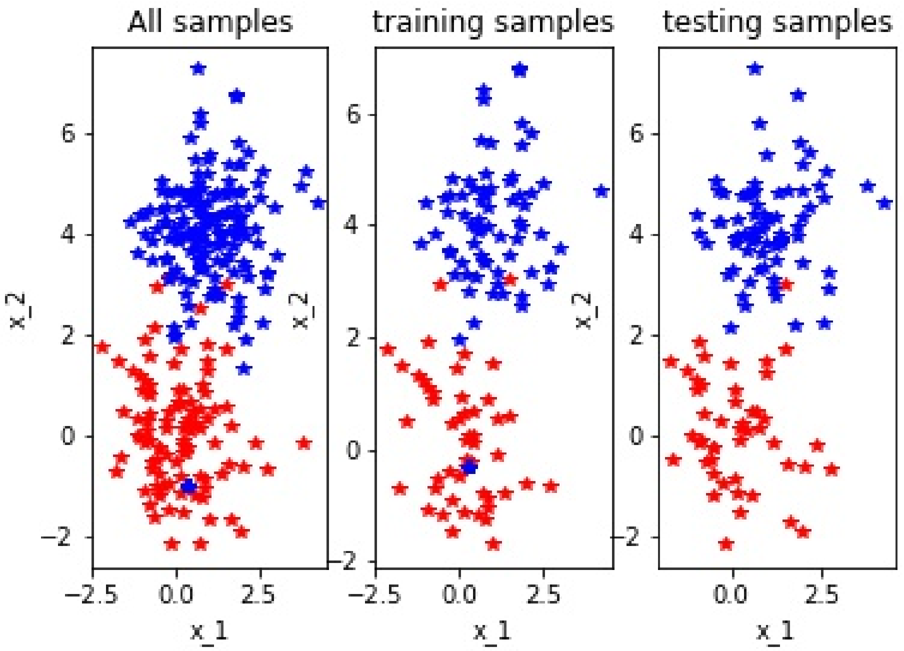

Figure 1: Data Split

Step 2

Instruction: Train a logistic regression model using the training data. To do so, you will need to use the functions we provided in the folder 'codeLogit'. Remember there are two different implementations.

Results: Training codes are seen in source code file. To compare the performance of two implementations, I calculate the score of train set for both, which means the ratio of the numbers of the correctly predicted over the number of all examples. The score for sklearn model is **0.97** and the score for self-developed model is **0.78**.

Step 3

Instruction: Apply the learned model to get the binary classes of testing samples. This step should be modified according to the implementation of the second step.

Results: The results of binary classes of testing samples for both models are printed in source code file.

Step 4

Instruction: Compare the predictions with the ground-truth labels and calculate average errors and standard deviation
Results: The average error and standard deviation for sklearn model is **0.023** and **0.15**, and the average error and standard deviation for self-developed model is **0.192** and **0.39**

Problem 2 Confusion Matrix

Instruction: Manually compute and report the confusion matrix and accuracy. For each of the three categories, calculate its precision and recall rates.
Results::See confusion matrix below: Then we calculate the accuracy and precision, recall for each class.

	$\hat{y}$=Cat	$\hat{y}$=Dog	$\hat{y}$=Monkey
y=Cat	1	3	1
y=Dog	3	3	2
y=Monkey	2	2	3

$$\begin{aligned}
Accuracy &= \frac{7}{20} = 0.35 \\
Precision_{Cat} &= \frac{1}{6} \approx 0.17 \\
Precision_{Dog} &= \frac{3}{8} = 0.375 \\
Precision_{Monkey} &= \frac{3}{6} = 0.5 \\
Recall_{Cat} &= \frac{1}{5} \approx 0.2 \\
Recall_{Dog} &= \frac{3}{8} = 0.375 \\
Recall_{Monkey} &= \frac{3}{7} \approx 0.14
\end{aligned} \tag{1}$$

Problem 3 Comparative Studies

Instruction: Please use above function in the script, and report the confusion matrix of both logistic regression implementations.
Results: see codes in source code file.

sklearn model:

	$\hat{y} = 0$	$\hat{y} = 1$
y=0	48	3
y=1	0	79

self-developed model:

	$\hat{y} = 0$	$\hat{y} = 1$
y=0	26	25
y=1	0	79

Statistics:

	sklearn	self-developed
Accuracy	0.977	0.808
Precision of 0	1	1
Precision of 1	0.963	0.759
Recall of 0	0.941	0.51
Recall of 1	1	1