Data Analytics with Python

Week 8 Assignment Solution

Q1. Which of the following methods do we use to best fit the data in Logistic Regression?

A) Least Square Error

B) Maximum Likelihood

C) Jaccard distance

D) Both A and B

Solution: B

Logistic regression uses maximum likely hood estimate for training a logistic regression.

Q2. Which of the following evaluation metrics can not be applied in case of logistic regression output to compare with target?

A) AUC-ROC

B) Accuracy

C) Logloss

D) Mean-Squared-Error

Solution: D

Since, Logistic Regression is a classification algorithm so it's output can not be real time value so mean squared error can not use for evaluating it.

Q3. Let f(x) denote the logistic function. The range of f(x) for any real value of x is

A. (0, 1)

B. (-1, 1)

C. All positive integers

D. All negative integers

Solution: A

For values of *x* in the range of real number from −∞ to +∞ Logistic function will give the output between (0,1)

Q4. Which of the following option is true?

A. Linear Regression errors values has to be normally distributed but in case of Logistic Regression it is not the case

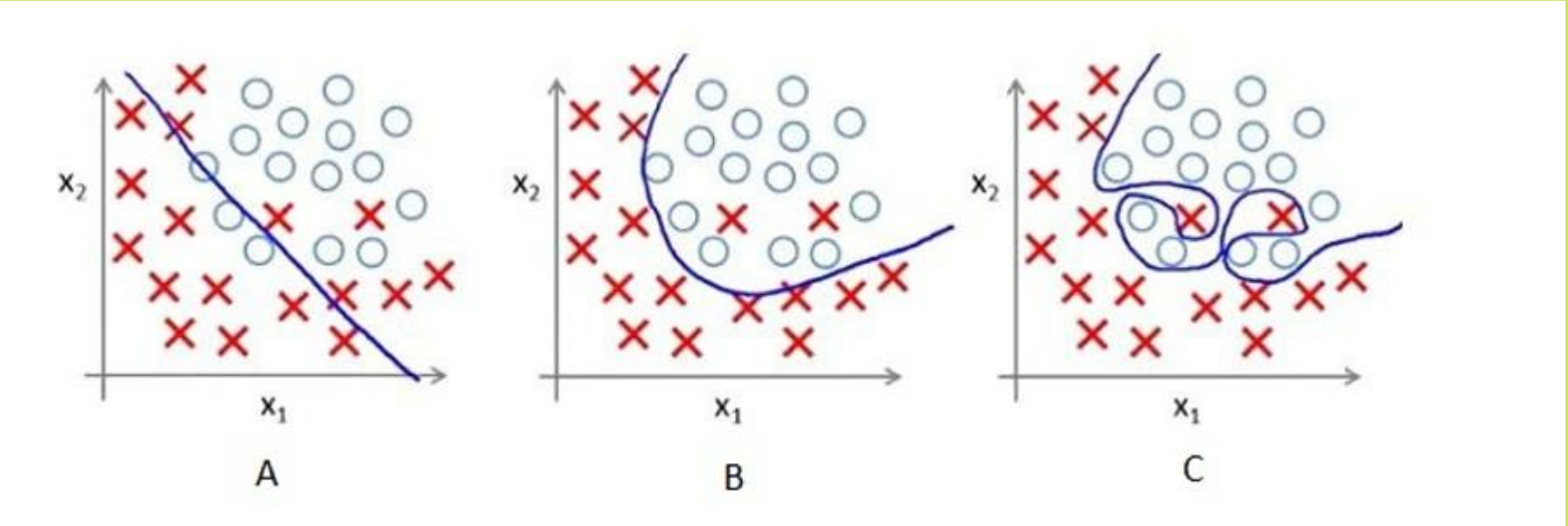
B. Logistic Regression errors values has to be normally distributed but in case of Linear Regression it is not the case

C. Both Linear Regression and Logistic Regression error values have to be normally distributed

D. Both Linear Regression and Logistic Regression error values have not to be normally distributed

Solution: A

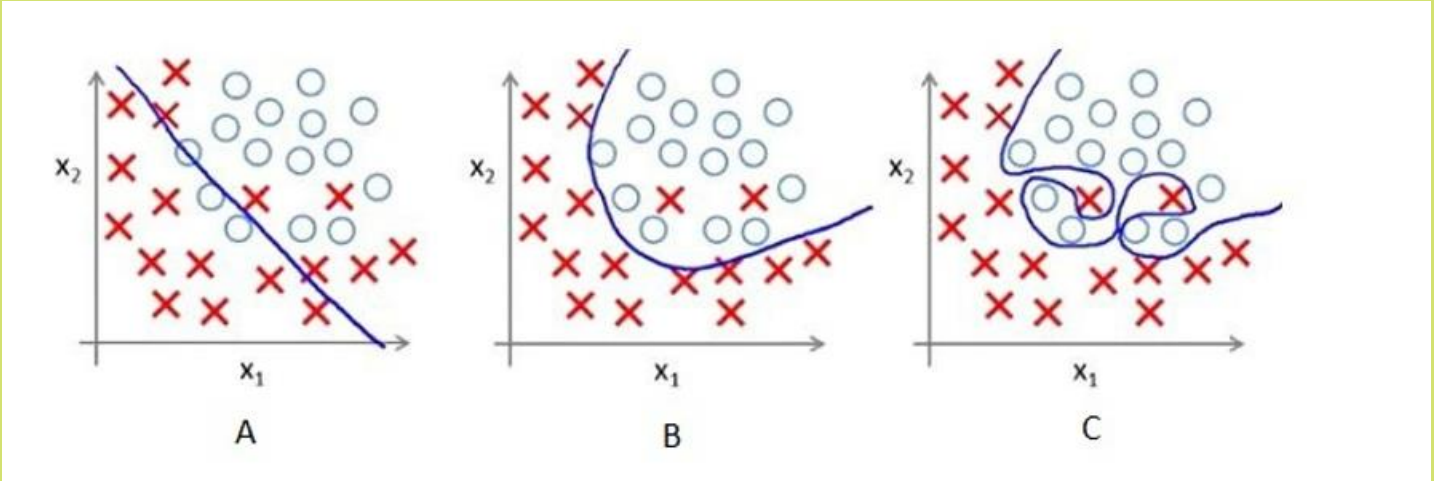
Q5. For the figure given below, which decision boundary is overfitting the training data?



Solution: C

Since in figure 3, Decision boundary is not smooth that means it will over-fitting the data.

Q6. Select the correct alternatives from the following based on the figure



1. The training error in first plot is maximum as compared to second and third plot.
2. The best model for this regression problem is the last (third) plot because it has minimum training error (zero).
3. The second model is more robust than first and third because it will perform best on unseen data.
4. The third model is overfitting more as compared to first and second.
5. All will perform same because we have not seen the testing data.

Solution: 1, 3 and 4

Q7. For categorical data with ‘n’ categories, the number of dummy variables will be_____

A. n

B. n-1

C. n+1

D. 2n

Solution: B

For 'n' categories, we need 'n-1' dummy variables to represent them.

Q8. In binary logistic regression

A. The dependent variable is continuous

B. The dependent variable is divided into two equal subcategories

C. The dependent variable consists of two categories

D. There is no dependent variable

Solution: C

In binary logistic regression, the dependent variable consists of two categories, usually referred to as the "success" category and the "failure" category.

Q9.

The recall, also known as sensitivity or true positive rate, is a measure of the proportion of actual positives that are correctly identified by a classifier

Given
number of false negatives = 5
number of true positives = 20

The recall can be calculated as:

Recall = True positives / (True positives + False negatives)

Recall = 20 / (20 + 5)

Recall = 20 / 25

Recall = 0.8

Q10

.

The F-measure, also known as the F1 score, is a harmonic mean of precision and recall. It is a measure of a classifier's accuracy that considers both precision and recall.

Given
Precision = 0.6
Recall = 0.4

The f-measure is calculated as:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$F1\ Score = 2 \times \frac{0.6 \times 0.4}{0.6 + 0.4}$$

$$F1\ Score = 2 \times 0.24$$

$$\boldsymbol{F1\ Score = 0.48}$$