University of Roehampton London

CMP020L012A	Data Analytics	2022/23

Week-6
Linear Regression
Solution
Answer the following Multiple-Choice Questions

1. A Regression model in which more than one predictor attribute is used to predict the label is called:
   a) Multiple Regression Model
   b) Single Regression Model
   c) Dependent Model
   d) None of the above

2. In the regression equation, $y=b_0 + b_1x$, where b_0 is equivalent to the:
   a) y intercept
   b) slope of the line
   c) Independent attribute
   d) None of the above

3. In the regression equation, $y=b_0 + b_1x$, where b_1 is:
   a) y intercept
   b) slope of the line
   c) independent attribute
   d) None of the above

s

4. What is Regression?
   a) It is a technique to fix data
   b) It is a technique to predict values
   c) It is a technique to fix outliers
   d) All of the above

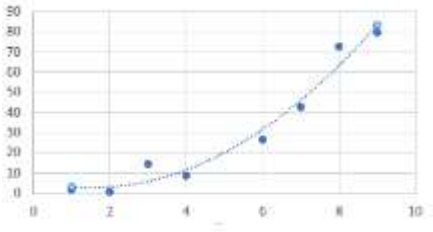
5. Which of the following is a method we use to find the best fit line for data in Linear Regression?
   a) Least Square Error
   b) Maximum Likelihood
   c) Logarithmic Loss
   d) None of them

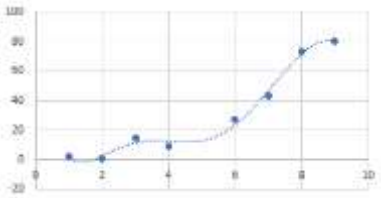
6. Which of the following evaluation metric can be used to evaluate a model while modelling a continuous output variable?
   a) ROC
   b) Accuracy
   c) Mean-Squared-Error
   d) Log-loss

7. The statement 'Overfitting depends on the flexibility of the model' is:

a) True
b) False

8. Which of the following models would be a better fit for the data?





a) First
b) Second
c) Both
d) None

9. Overfitting is more probable when ___.

a) The number of Datapoints are lower
b) The number of Datapoints are higher
c) Both
d) None

10. You fit linear regression models for the same data, where the first one gives an RMSE value of 3.78 and the second one returns a value of 6.33. Which of these is a better model?

a) second
b) First
c) Both are same
d) None

11. Select among the following scenarios where linear regression algorithm can be applied

a) You want to predict the sales of a retail store based on its size, given the dataset of sales of retails stores and their sizes
b) You have collected data from a house rental website like commonfloor.com. The data has the rental prices of apartments and customer ratings as HIGH or LOW. You want to predict the customer rating, given the rental price of a new house.
c) You want to predict the customer likely to leave the network provider
d) You have a dataset of BMI (body mass index) and the fat percentage of the customers of a fitness center. Now, the fitness center wants to predict the fat percentage of a new customer, given his BMI.

12. The Regression technique produces a straight line as the relationship between x(input) and y(output) called:

a) Hypothesis Function
b) Linear Regression
c) Related Regression
d) None of the above

13. MSE stands for:
- a) Minimum Squared Error
- b) Maximum Squared Error
- c) Mean Squared Error
- d) None of the above

14. MSE give the difference between _____
- a) Predicted Value and True Value
- b) True Value and False Value
- c) True Value and Wrong Value
- d) None of the above

15. Candidate solutions/models can be multiple.
- a) True
- b) False

16. The Dependent variable is:
- a) The feature of the dataset
- b) The parameters (predictors) values of the dataset
- c) The value we want to predict
- d) None of the above

17. What is Outliers?
- a) Extreme datapoints in the dataset
- b) Regression technique
- c) Value that are correlated to each other
- d) The trend in the dataset

18. Polynomial Regression is used for what?
- a) Classify Binary data
- b) Handle with Non-Linear separable data
- c) Find the best Linear line
- d) Handle Linear and separable data

University of Roehampton London

Answer the following Question

Assuming, you are given with a sample dataset set containing height and weight data as follows:

Height	Weight
43	41
44	45
45	49
46	47
47	44

You are now to calculate the MSE of each datapoint in the sample set for the two given regression lines where, the first one is y=9.2+0.8x and the second one is y= 8+1.5x to determine one of them as the better solution/model for this problem.

Answer:

Step 1: Find the new Y' values:

- 9.2 + 0.8(43) = 43.6
- 9.2 + 0.8(44) = 44.4
- 9.2 + 0.8(45) = 45.2
- 9.2 + 0.8(46) = 46
- 9.2 + 0.8(47) = 46.8

Step 2: Find the error (Y – Y'):

- 41 – 43.6 = -2.6
- 45 – 44.4 = 0.6
- 49 – 45.2 = 3.8
- 47 – 46 = 1
- 44 – 46.8 = -2.8

Step 3: Square the Errors:

- $-2.6^2 = 6.76$
- $0.6^2 = 0.36$
- $3.8^2 = 14.44$
- $1^2 = 1$
- $-2.8^2 = 7.84$

Step 4: Add all the squared errors up: 6.76 + 0.36 + 14.44 + 1 + 7.84 = 30.4.
Step 5: Find the mean squared error:
30.4 / 5 = 6.08.

Similarly, you do the same for the second regression line/equation. And certainly, the minimum MSE helps us to determine the better solution among the two.

Kimia Aksir