

MACHINE LEARNING

Name: Darshan M.Katgeri

In Q1 to Q11, only one option is correct, choose the correct option:

1. Which of the following methods do we use to find the best fit line for data in Linear Regression?

- A) Least Square Error B) Maximum Likelihood
- C) Logarithmic Loss D) Both A and B

Answer: A

2. Which of the following statement is true about outliers in linear regression?

- A) Linear regression is sensitive to outliers
- B) linear regression is not sensitive to outliers
- C) Can't say D) none of these

Answer: A

3. A line falls from left to right if a slope is _____?

- A) Positive B) Negative C) Zero D) Undefined

Answer: B

4. Which of the following will have symmetric relation between dependent variable and independent variable?

- A) Regression B) Correlation C) Both of them D) None of these

Answer: B

5. Which of the following is the reason for over fitting condition?

- A) High bias and high variance B) Low bias and low variance
- C) Low bias and high variance D) none of these

Answer: A

6. If output involves label then that model is called as:

- A) Descriptive model B) Predictive modal
- C) Reinforcement learning D) All of the above

Answer: B

7. Lasso and Ridge regression techniques belong to _____?

- A) Cross validation B) Removing outliers C) SMOTE D) Regularization

Answer: D

8. To overcome with imbalance dataset which technique can be used?

- A) Cross validation B) Regularization C) Kernel D) SMOTE

Answer: D

9. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses _____ to make graph?

- A) TPR and FPR B) Sensitivity and precision
C) Sensitivity and Specificity D) Recall and precision

Answer: A

10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less.

- A) True B) False

Answer: B

11. Pick the feature extraction from below:

- A) Construction bag of words from a email
B) Apply PCA to project high dimensional data
C) Removing stop words
D) Forward selection

Answer: B

In Q12, more than one options are correct, choose all the correct options:

12. Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression?

- A) We don't have to choose the learning rate.
B) It becomes slow when number of features is very large.
C) We need to iterate.
D) It does not make use of dependent variable.

Answer: B, D

Q13 and Q15 are subjective answer type questions, Answer them briefly.

13. Explain the term regularization?

Answer.13:

Understanding of the terms generalization, underfitting and overfitting would be helpful before explanation of regularization.

Underfitting: Choosing too simple a model is called underfitting. Underfit model performs badly even on training data. The prediction score is very less.

Overfitting: Overfitting occurs when one fits the model too closely to the particularities of the training set and obtain a model that works well with training set but not able to generalize new data.

Regularization: The term regularization means to explicitly restricting a model to avoid overfitting. The problem with overfitting model is that the model may predict with great accuracy for training data, but on test data the model may not perform well. This discrepancy between performance on the training set and the test set is a clear sign of overfitting. That is the model is biased towards training

data and may have high variance with respect to test data. On the other hand, if the prediction is low on training data as well as on the test data then it is a case of under fitting the model.

Generally, in ML the objective must be to arrive at a generalized model. Usually, we have to build models that make accurate prediction using training set. A generalized model is a good model when the model has low bias and low variance. This may be achieved by introducing a small amount of bias in the train model by readjusting the slope of the line. Lasso and Ridge regression techniques may be used in regularization.

14. Which particular algorithms are used for regularization?

Answer.14:

There are two algorithms that are extensively used for regularization, namely,

- i. Lasso regression (uses L1-norm)(least absolute shrinkage and selection operator)
- ii. Ridge regression (uses L2-norm): Each feature should have as little effect on outcome as possible (.i.e the line should have small slope).
- iii. Elasticnet (combination of Lasso and Ridge)

15. Explain the term error present in linear regression equation?

Answer.15:

The linear regression equation is given by

$$y = \beta x + c + e \dots\dots\dots(i)$$

where β is slope of line, c is the y-intercept and e is the error term.

The error is defined as the difference between the actual value and the predicted value.

$$\text{Error} = \text{actual value} - \text{predicted value}$$

The objective is to have minimum error (or to minimize the error)
