

- 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

Ans. a) Least Square Error

- 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

Ans. a) Linear regression is sensitive to outliers

- 3) A line falls from left to right if a slope is \_\_\_\_\_ ?
- a) Positive
  - b) Negative
  - c) Zero
  - d) Undefined

Ans. b) Negative

- 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

Ans. c) Both of them

- 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

Ans. c) Low bias and high variance

- 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

Ans. b) Predictive modal

7) Lasso and Ridge regression techniques belong to \_\_\_\_\_ ?

- a) Cross validation
- b) Removing outliers
- c) SMOTE
- d) Regularization

Ans. d) Regularization

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

- a) Cross validation
- b) Regularization
- c) Kernel
- d) SMOTE

Ans. d) SMOTE

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

Ans. a) TPR and FPR

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

- a) True
- b) False

Ans. a) True

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

Ans. a) Construction bag of words from a email

- b) Apply PCA to project high dimensional data
- c) Removing stop words

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

Ans. 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.

13) Explain the term regularization?

Ans. Regularization describes a technique to prevent overfitting. Complex models are prone to picking up random noise from training data which might obscure the patterns found in the data. Regularization helps reduce the influence of noise on the model's predictive performance. The goal of a machine learning model is to find patterns in data and apply the knowledge about these patterns to make predictions on different data from a similar problem domain.

Assume you are trying to predict blood pressure using such factors as age, alcohol consumption, and history of smoking. Generally, higher age, drinking, and smoking are correlated with higher blood pressure. So, there is a pattern that the model can pick up. Nevertheless, there may be some old individuals in your dataset who smoke and drink but have below-average blood pressure. This deviation from the pattern cannot be explained by our predictors. We treat it as noise.

Machine learning models might pick up some of that noise in a process known as overfitting. Even single examples that deviate significantly from the pattern known have the potential to change the model's understanding of the pattern significantly. These examples are also known as outliers.

Imagine you have an outlier in your blood pressure training dataset (for example, an 80-year-old man who has been smoking and drinking his entire life but only has a blood pressure of 125/70).

If we fit a complex non-linear model without regularization, we might end up with a polynomial function that fits the training data perfectly well.

In other words, the model thinks that a blood pressure of 125/70 for an 80-year-old smoking alcoholic is part of the pattern. If it finds an individual with similar characteristics in the test set, it might predict unrealistically low blood pressure.

14) Which particular algorithms are used for regularization?

Ans. There are main regularization techniques are:

1. Ridge Regression (L2 Norm)
2. Lasso (L1 Norm)

Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called as L2 regularization. It is one of the types of linear regression in which a small amount of bias is introduced so that we can get better long-term predictions. In this technique, the cost function is altered by adding the penalty term to it. The amount of bias added to the model is called Ridge Regression penalty. We can calculate it by multiplying with the lambda to the squared weight of each individual feature.

Lasso regression is another regularization technique to reduce the complexity of the model. It stands for Least Absolute and Selection Operator. It is like the Ridge Regression except that the penalty term contains only the absolute weights instead of a square of weights. Since it takes absolute values, hence, it can shrink the slope to 0, whereas Ridge Regression can only shrink it near to 0. It is also called as L1 regularization.

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

Ans. The error term of a regression equation represents all of the variation in the dependent variable not explained by the weighted independent variables. A regression equation is the formula for a straight line — in this case, the best-fit line through a scatterplot of data. If there were no error, all the data points would be located on the regression line; to the extent they are not represents error; this is what the error term summarizes.