# MACHINE LEARNING

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

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

   A) Linear regression is sensitive to outliers

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

   B) Negative

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

   B) Correlation

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

   C) Low bias and high variance

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

   B) Predictive modal

7. Lasso and Ridge regression techniques belong to _____?

   C) Regularization

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

   D) SMOTE

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

   D) Sensitivity and Specificity

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

    B) False

11. . Pick the feature extraction from below:

   B) Apply PCA to project high dimensional data

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.

13.  Explain the term regularization?

Answer: **Regularization** refers to techniques that are used to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting or underfitting.
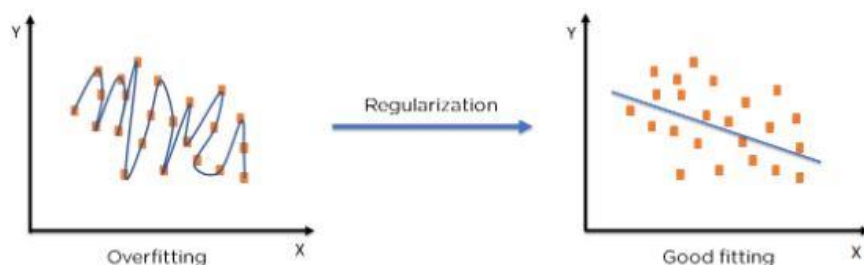


Regularization on an over-fitted model

There are two main types of regularization techniques: Ridge Regularization and Lasso Regularization.

**Ridge regression** is the method used for the analysis of multicollinearity in multiple regression data. It is most suitable when a data set contains a higher number of predictor variables than the number of observations. The second-best scenario is when multicollinearity is experienced in a set

**Lasso regression** is also called Penalized regression method. This method is usually used in machine learning for the selection of the subset of variables. It provides greater prediction accuracy as compared to other regression models. Lasso Regularization helps to increase model interpretation

14. Which particular algorithms are used for regularization?

ANSWER)

**LASSO REGRESSION**

• t works the same as ridge regression when it comes to assigning the penalty for coefficient,

• It removes the coefficient and the variables with the help of this process and limits the bias

- regularization term penalizes absolute value of the coefficients
- sets irrelevant values to 0
- might remove too many features in your model

**RIDGE REGRESSION**

- It is used highly for the treatment of multicollinearity in regression, it means when an independent variable is correlated in such a way that both resemble each other,
- It causes high variance among the independent variables, we can change the value of the independent variable but it will cause a loss of information.

---

- penalizes the size (square of the magnitude) of the regression coefficients
- enforces the $B$ (slope/partial slope) coefficients to be lower, but not 0
- does not remove irrelevant features, but minimizes their impact

**ELASTIC REGRESSION**

Elastic net is a penalized linear regression model that includes both the L1 and L2 penalties during training. Using the terminology from "The Elements of Statistical Learning," a hyperparameter "alpha" is provided to assign how much weight is given to each of the L1 and L2 penalties

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

ERROR TERM IN LINEAR REGRESSION

The standard error of the regression (S), also known as the standard error of the estimate, represents the average distance that the observed values fall from the regression line. Conveniently, it tells you how wrong the regression model is on average using the units of the response variable.

An error term in statistics is a value which represents how observed data differs from actual population data. It can also be a variable which represents how a given statistical model differs from reality. The error term is often written ε.

$$y' = a + bx+e$$

**e** – is the difference between actual value and model predicted value

The Mean Squared Error (MSE) is perhaps the simplest and most common loss function, often taught in introductory Machine Learning courses. To calculate the MSE, you take the difference between your model's predictions and the ground truth, square it, and average it out across the whole dataset.20-May-2019

MSE formula = (1/n) * Σ(actual – forecast)$^2$
Where:

- n = number of items,
- Σ = <u>summation notation</u>,
- Actual = original or observed y-value,
- Forecast = y-value from regression.