MACHINE LEARNING

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
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
3. A line falls from left to right if a slope is_____?
   A) Positive B) Negative
   C) Zero D) Undefined
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
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
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
7. Lasso and Ridge regression techniques belong to_____?
   A) Cross validation B) Removing outliers
   C) SMOTE D) Regularization
8. To overcome with imbalance dataset which technique can be used?
   A) Cross validation B) Regularization
   C) Kernel 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
10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less.
    A) True B) False
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

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.

MACHINE LEARNING

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

13. Explain the term regularization?

14. Which particular algorithms are used for regularization?

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

MACHINE LEARNING

Answers

1. (A)
2. (A)
3. (B)
4. (B)
5. (C)
6. (B)
7. (D)
8. (D)
9. (A)
10. (B)
11. (A)
12. (B)

SUBJECTIVE TYPE

13. Regularization: In Machine Learning Regularization is a technique that makes slight modifications to the learning algorithm such that the model generalizes better. This in turn improves the model's performance on the unseen data as well.

    Sometimes the machine learning model performs well with the training data but does not perform well with the test data. It means the model is not able to predict the output when dealing with unseen data by introducing noise in the output, and hence the model is called overfitting. Overfitting is a phenomenon that occurs when a Machine Learning model is constrained to the training set and not able to perform well on unseen data. Here, then, regularization comes into the picture to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting.

    This technique can be used in such a way that it will allow maintaining all variables or features in the model by reducing the magnitude of the variables. Hence, it maintains accuracy as well as a generalization of the model. It mainly regularizes or reduces the coefficient of features toward zero.

    In simple words, in the regularization technique, we reduce the magnitude of the features by keeping the same number of features.

14. Algorithms used for regularization: There are mainly two types of regularization techniques, which are given below:

    - Ridge Regression
    - Lasso Regression

MACHINE LEARNING

❖ *Ridge Regression*

- Ridge regression 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.
- Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called **L2 regularization**.
- 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 the **Ridge Regression penalty**. We can calculate it by multiplying the lambda by the squared weight of each individual feature.
- The equation for the cost function in ridge regression will be:

$$\sum_{i=1}^{M}(y_i - y'_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{n}\beta_j * x_{ij}\right)^2 + \lambda\sum_{j=0}^{n}{\beta_j}^2$$

- In the above equation, the penalty term regularizes the coefficients of the model, and hence ridge regression reduces the amplitudes of the coefficients that decrease the complexity of the model.
- As we can see from the above equation, if the values of λ tend to zero, the equation becomes the cost function of the linear regression model. Hence, for the minimum value of λ, the model will resemble the linear regression model.
- A general linear or polynomial regression will fail if there is high collinearity between the independent variables, so to solve such problems, Ridge regression can be used.
- It helps to solve the problems if we have more parameters than samples.

❖ *Lasso Regression*

- Lasso regression is another regularization technique to reduce the complexity of the model. It stands for Least Absolute and Selection Operator.
- It is similar to 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 L1 regularization. The equation for the cost function of Lasso regression will be:

$$\sum_{i=1}^{M}(y_i - y'_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{n}\beta_j * x_{ij}\right)^2 + \lambda\sum_{j=0}^{n}|\beta_j|$$

MACHINE LEARNING

- Some of the features in this technique are completely neglected for model evaluation.
- Hence, the Lasso regression can help us to reduce the overfitting in the model as well as the feature selection.

Key Difference between Ridge Regression and Lasso Regression: -

- **Ridge regression** is mostly used to reduce the overfitting in the model, and it includes all the features present in the model. It reduces the complexity of the model by shrinking the coefficients.
- **Lasso regression** helps to reduce the overfitting in the model as well as feature selection.

15. Term Error present in Linear Regression Equation: An error term in statistics is a value that represents how observed data differs from actual population data. It can also be a variable that shows how a given statistical model differs from reality. The error term is often written ε.

    The classical linear regression model involves finding the best fitting linear model for observed data that shows the relationship between two variables.

    After collecting the data this data can be plotted as a scatter plot, with input on the x-axis and output on the y-axis. Then we would look for the line $y = \beta_0 + \beta_1 x$ that fits the data.

The general linear regression model can be stated by the equation:

$$yi = \beta0 + \beta1X1i + \beta2X2i + \cdots + \beta kXki + \varepsilon i,$$

where, β0 is the intercept, βi's are the slope between Y and the appropriate Xi, and ϵ (pronounced epsilon), is the error term that captures errors in the measurement of Y and the effect on Y of any variables missing from the equation that would contribute to explaining variations in Y.

This equation is the theoretical population equation and therefore uses Greek letters. The equation we will estimate will have the Roman equivalent symbols. This is parallel to how we kept track of the population parameters and sample parameters before. The symbol for the population mean was μ and for the sample mean X_bar and for the population standard deviation was σ and for the sample standard deviation was s. The equation that will be estimated with a sample of data for two independent variables will thus be:

$$yi = b0 + b1x1i + b2x2i + ei$$

As with our earlier work with probability distributions, this model works only if certain assumptions hold. These are that the Y is normally distributed, the errors are also normally distributed with a mean of zero and a constant standard deviation, and the error terms are independent of the size of X and independent of each other.

This is the estimated value of y. It is the value of y obtained using the regression line. $\tilde{y}$ is not generally equal to y from the data. The term $y0 - \tilde{y} = e0$ is called the "error" or residual.

It is not an error in the sense of a mistake. The error term was put into the estimating equation to capture missing variables and errors in measurement that may have occurred in the dependent variables. The absolute value of residual measures the vertical distance between the actual value of y and the estimated value of y. In other words, it measures the vertical distance between the actual data point and the predicted point on the line as can be seen on the graph at point X0.

MACHINE LEARNING

If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for y.

If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for y.

In the graph, y0− ỹ = e0 is the residual for the point shown. Here the point lies above the line and the residual is positive. For each data point the residuals, or errors, are calculated

yi− ỹi =ei for i=1,2, 3..., ni=1,2, 3..., n, where, n is the sample size. Each |e| is a vertical distance.

The sum of the errors squared is the term obviously called **Sum of Squared Errors (SSE)**.

Using calculus, you can determine the straight line that has the parameter values of b0b0 and b1b1 that minimizes the **SSE**. When you make the **SSE** a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation:

y^=b0+b1xy^=b0+b1x, where,

$$b_0 = \bar{y} - b_1\bar{x} \text{ and } b_1 = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{\text{cov}(x,y)}{s_x^2}$$

MACHINE LEARNING