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

Ans- 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

Ans- A

3. A line falls from left to right if a slope is_____?
   A) Positive B) Negative
   C) Zero D) Undefined

Ans- 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

Ans- 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

Ans- C

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- D

7. Lasso and Ridge regression techniques belong to_____?
   A) Cross validation B) Removing outliers
   C) SMOTE D) Regularization

Ans- D

8. To overcome with imbalance dataset which technique can be used?
   A) Cross validation B) Regularization
   C) Kernel D) SMOTE

Ans- D

MACHINE LEARNING

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

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

A) True
B) False

Ans- 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

Ans- A

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.

Ans- A, B and C

MACHINE LEARNING

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

13. Explain the term regularization?

Ans- *Regularization is one of the basic and most important concept in the world of Machine Learning. The word regularize means to make things "regular" or "acceptable". This is exactly why we use it for. Regularizations are techniques used to reduce the error by fitting a function appropriately on the given training set and avoid overfitting. It is a form of regression, that constrains/ regularizes or shrinks the coefficient estimates towards zero. This technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.*

In other words we can say, It helps to reduce the variance of the model, without a substantial increase in the bias.
If there is variance is the model that means that the model won't fit well for dataset different than training data. The tuning parameter λ (lambda) controls this bias and variance tradeoff. When the value of λ is increased up to a certain limit, It reduces the variance without loosing any important properties in the data. But after a certain limit, the model will start losing some important properties which will start loosing some important properties which will increase the bias in the data. Thus, the selection of good value of λ is the key. The value of λ is selected using cross-validation methods. A set of λ is selected and cross-validation error is calculated for each value of λ and that value of λ is selected for which the cross-validation error is minimum.

There are basically 3 types algorithm we used under Regularization:-

- *LASSO(Least Absolute Shrinkage and Selection Operator) Regression (L1 Form)*
- *RIDGE REGRESSION (L2 Form)*
- *ELASTIC NET(Less popular)*

---

14. Which particular algorithms are used for regularization?

Ans. *There are three main regularization techniques, namely:*

- *LASSO(Least Absolute Shrinkage and Selection Operator) Regression (L1 Form)Ridge Regression (L2 Norm)*
- *RIDGE REGRESSION (L2 Form)*
- *ELASTIC NET(Less popular)*

Ridge and Lasso can be used for any algorithms involving weight parameters, including neural nets. Elastic-net is primarily used in any kind of neural networks e.g. ANN, DNN, CNN or RNN to moderate the learning.

Let's take a closer look at each of the techniques :-

MACHINE LEARNING

- *L1 Form – Lasso method is a type of method which does not give importance to the data which has no relationship with the label. Lasso regression penalizes the model based on the sum of magnitude of the coefficients.*

  The regaularization term is given by:

  Regularization $= \lambda * \Sigma |\beta j|$

  Where, λ *is the shrinkage factor(learning rate)*

- *L2 Form – Ridge method is the method which treats feature according to lits importance i.e, its strength of relationship with the label. Ridge Regression penalizes the model based on the sum of the squares of magnitude of the*

  Coefficients. The regularization term is given by

  Regularization $= \lambda * \Sigma |\beta j^2|$

  Where, λ *is the shrinkage factor(learning rate)*

---

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

Ans. *In a linear regression model over the time, the term error is the difference between the expected data at a particular time and the price that was actually observed. The error term in a regression equation represents the effect of the variables that were removed from the equation.*

The error term is also known as the "residual", "disturbance", or "remainder term".

An error term is a residual variable produced by a statistical model, created when the model does not fully represent the actual relationship between the independent variables and the dependent variables. As a result of this incomplete relationship, the error term is the amount at which the equation may differ during empirical analysis.

An error term appears to indicate the uncertainty in the model and is a residual variable that accounts for a lack of perfect goodness of fit. Heteroskedastic refers to a condition in which the variance of the residual term, or error term, in a regression model varies widely.

An error term represents the margin of error within a statistical model; it refers to the sum of the deviations within the regression line, which provides an explanation for the difference between the theoretical value of the model and the actual observed results. The regression line is used as a point of analysis when attempting to determine the correlation between one independent variable and one dependent variable.

Error Term Use in a Formula

An error term essentially means that the model is not completely accurate and results in differing results during real-world applications.

For example, assume there is a multiple linear regression function that takes the following form:

$Y = \alpha X + \beta \rho + \epsilon$

where:

α, β=Constant parameters

X, ρ=Independent variables

ϵ=Error term

MACHINE LEARNING

When the actual Y differs from the expected or predicted Y in the model during an empirical test, then the error term does not equal 0,
which means there are other factors that influence Y.
