MACHINE LEARNING ASSIGNMENT

- 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 1. 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

Answer 2. 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

Answer 3. 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

Answer 4. B) Correlation

5. Which of the following is the reason for over fitting condition?
   A) High bias and high variance B) Low bias and low variance
   B) Low bias and high variance D) none of these

Answer 5. B) 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

Answer 6. B) Predictive modal

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

Answer 7. D) Regularization

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

Answer 8. 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

Answer 9. 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

Answer 10. B) False

11. Pick the feature extraction from below:
    A) Construction bag of words from an email
    B) Apply PCA to project high dimensional data
    C) Removing stop words
    D) Forward selection

Answer 11. B) Apply PCA to project high dimensional data

- 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 11. A) We don't have to choose the learning rate.
B) It becomes slow when number of features is very large.

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

13. Explain the term regularization?

Answer 13.

Regularization:

This is a form of regression, that constrains / regularizes or shrinks the coefficient estimates towards zero. In other words, this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.

A simple relation for linear regression looks like this. Here Y represents the learned relation and β represents the coefficient estimates for different variables or predictors(X).

$Y \approx \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta pXp$

The fitting procedure involves a loss function, known as residual sum of squares or RSS. The coefficients are chosen, such that they minimize this loss function.

$$\mathrm{RSS} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2.$$

Now, this will adjust the coefficients based on your training data. If there is noise in the training data, then the estimated coefficients won't generalize well to the future data. This is where regularization comes in and shrinks or regularizes these learned estimates towards zero.

14. Which particular algorithms are used for regularization?

Answer 14.

Algorithms used for regularization are:

1. Ridge Regression (L2 Regularization)

2. Lasso Regression (L1 Regularization)

Ridge Regression (L2 Regularization):

Ridge regression is often referred to as regularisation or L2 norm.

When using this technique, we add the sum of weight's square to a loss function and thus create a new loss function which is denoted thus:

$$\text{Loss} = \sum_{j=1}^{m}\left(Yi - Wo - \sum_{i=1}^{n} Wi\,Xji\right)^2 + \lambda\sum_{i=1}^{n} Wi^2$$

As seen above, the original loss function is modified by adding normalized weights. Here normalized weights are in the form of squares.

You may have noticed parameters λ along with normalized weights. λ is the parameter that needs to be tuned using a cross-validation dataset. When you use λ=0, it returns the residual sum of square as loss function which you chose initially. For a very high value of λ, loss will ignore core loss function and minimize weight's square and will end up taking the parameters' value as zero.

Now the parameters are learned using a modified loss function. To minimize the above function, parameters need to be as small as possible. Thus, L2 norm prevents weights from rising too high.

Lasso Regression (L1 Regularization):

Also called lasso regression and denoted as below:

$$\text{Loss} = \sum_{j=1}^{m}\left(Yi - Wo - \sum_{i=1}^{n} Wi\,Xji\right)^2 + \lambda\sum_{i=1}^{n}|Wi|$$

This technique is different from ridge regression as it uses absolute weight values for normalization. λ is again a tuning parameter and behaves in the same as it does when using ridge regression.

As loss function only considers absolute weights, optimization algorithms penalize higher weight values.

In ridge regression, loss function along with the optimization algorithm brings parameters near to zero but not actually zero, while lasso eliminates less important features and sets respective weight values to zero. Thus, lasso also performs feature selection along with regularization.

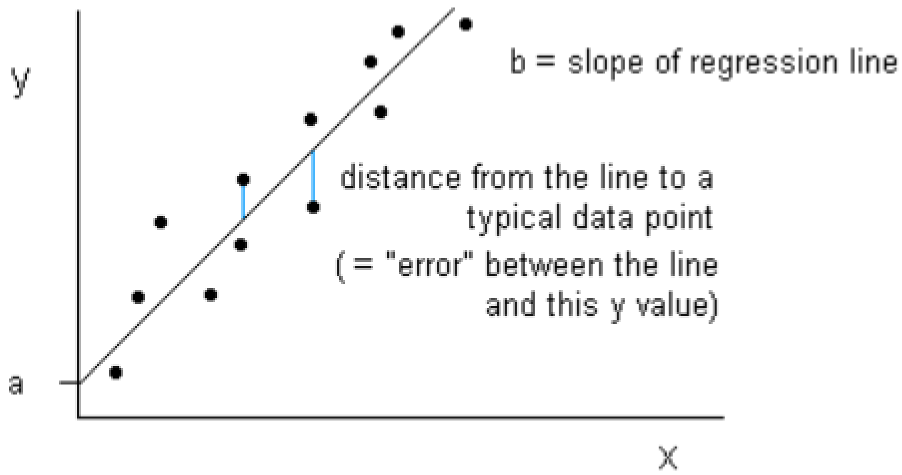
15. Explain the term error present in linear regression equation?

Answer 15. 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.

In econometric theory, the classical normal linear regression model (CNLRM) involves finding the best fitting linear model for observed data that shows the relationship between two variables.

For example, let's say you were running a study on the way the number of exams in a certain college affect the amount of red bull purchased from college vending machines. You could collect data which told you how many exams were given and how much red bull was purchased on a dozen or more days during the semester. This data can be plotted as a scatter plot, with exams (Ex) per given day on the x axis and red bull purchased (RB) per given day on the y axis. Then you would look for the line $y = β0 + β1\ x$ that best fit the data.

"Best fit" here means that the error term, the distance from each point to the line, is minimized. Since the relationship between variables is probably not completely linear and because there are other factors outside the scope of our study (sales on red bull, sales on other caffeine drinks, difficult physics homework sets, etc.) the graph won't actually go through all our data points. The distance between each point and the linear graph (shown as black arrows on the above graph) is our error term. So we can write our function as $RB=β0 + β1\ Ex + ε$ where β0 and β1 are constants and ε is an (non constant) error term.


y
b = slope of regression line
distance from the line to a typical data point
(= "error" between the line and this y value)
a
x