# 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

1. Which of the following statement is true about outliers in linear regression?
   1. Linear regression is sensitive to outliers B) linear regression is not sensitive to outliers

C) Can’t say D) none of these

1. A line falls from left to right if a slope is ?
   1. Positive B) Negative

C) Zero D) Undefined

1. Which of the following will have symmetric relation between dependent variable and independent variable?
   1. Regression B) Correlation

C) Both of them D) None of these

1. Which of the following is the reason for over fitting condition?
   1. High bias and high variance B) Low bias and low variance C) Low bias and high variance D) none of these
2. If output involves label then that model is called as:
   1. Descriptive model B) Predictive modal

C) Reinforcement learning D) All of the above

1. Lasso and Ridge regression techniques belong to ?
   1. Cross validation B) Removing outliers

C) SMOTE D) Regularization

1. To overcome with imbalance dataset which technique can be used?
   1. Cross validation B) Regularization

C) Kernel D) SMOTE

1. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses to make graph?
   1. TPR and FPR B) Sensitivity and precision

C) Sensitivity and Specificity D) Recall and precision

1. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less.
   1. True B) False
2. Pick the feature extraction from below:

A) Construction bag of words from a email

B) Apply PCA to project high dimensional data

1. Removing stop words
2. Forward selection

# In Q12, more than one options are correct, choose all the correct options:

1. Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression?
   1. We don’t have to choose the learning rate.
   2. It becomes slow when number of features is very large. C) We need to iterate.

D) It does not make use of dependent variable.

# 

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

1. **Explain the term regularization?**
2. **Which particular algorithms are used for regularization?**

Regularization techniques to are used in linear regression models in order to minimize the adjusted loss function and prevent overfitting or underfitting. When a model is flexible and allows for tweaks, there is high variance, which in tun causes overfitting. Regularization aims to limit the flexibility of the models to certain extent. Some flexible type of models includes random forest. It increases flexibility of the model by reducing weights. Weights are nothing but values assigned to input features. A model cannot go beyond a certain weight otherwise it will overfit. Overfitting is nothing but a model exaggerating the importance of a certain input feature in relation to the output.

There are two popular types of regularization algorithms

1. **Ridge L2**

It modifies the overfitted or underfitted models by adding the penalty equivalent to the sum of the squares of the magnitude of coefficients

Cost function = Loss + λx Σ || w || ²

Loss = Sum of the squared residuals

λ = Penalty for the errors

W = slope of the curve / line

1. **Lasso Regression LI**

It modifies the overfitted underfitted by adding the penalty equivalent to the sum of the absolute values of coefficients

Cost function = Loss + 2 x E || w ||

Here , Loss = Sum of the squared residuals λ = Penalty for the errors w = slope of the curve / line

|  |  |
| --- | --- |
| **Ridge Regularization** | **Lasso Regularization** |
| Useful when we have many variables with relatively smaller data samples | Preferred when we are fitting a linear model with fewer variables |
| The model does not encourage convergence towards zero but is likely to make them closer to zero and prevent overfitting | It encourages the coefficients of the variables to go towards zero because of the shape constraint which is of an absolute value. |

1. **Explain the term error present in linear regression equation?**

In case of a tracking a stock’s price over time, the error term is the difference between the expected price at a particular time and the price that was actually observed. In instances where the price is same as the anticipated value at a particular time, the price will fall on the trend line and the error term will be zero. If there were no error, all the data points would be located *on* the regression line, and there will be n representation for the term ‘error’

The Points that do not fall directly on the trend line state that the dependent variable, i.e. the price, is influenced by more than just the independent variable, representing the passage of time. The error term therefore is defined by any influence being exerted on the price variable, such as changes in market sentiment, other than the independent variable.

## **Difference Between Error and Residuals**

Although the error term and residual are often used synonymously, there is an difference between the two. An error term is generally unobservable while a residual is observable and calculable, ehich makes it much easier to quantify and visualize. In effect, while an error term represents the way observed data differs from the actual population, a residual represents the way observed data differs from sample population data.