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 model
C) Reinforcement learning D) All of the above

Ans: (B)

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

Ans: (D)

MACHINE LEARNING

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

A) Cross validation
B) Regularization
C) Kernel
D) SMOTE

Ans: (D)

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: (C)

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: (B)

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.

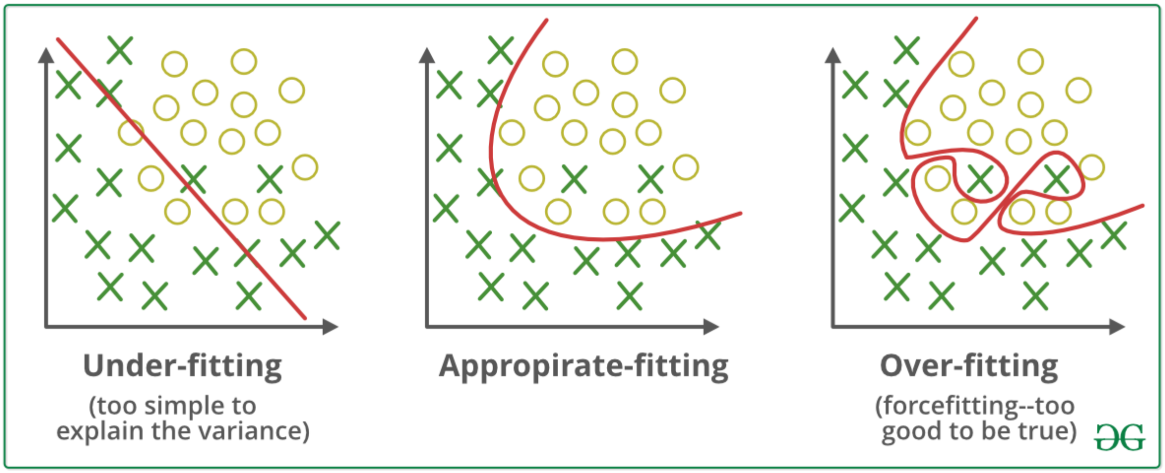
Ans: (A,B,C)

MACHINE LEARNING

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

13. Explain the term regularization?

Ans: Overfitting is a phenomenon that occurs when a Machine Learning model is constraint to training set and not able to perform well on unseen data.


Under-fitting
(too simple to explain the variance)
Appropirate-fitting
Over-fitting
(forcefitting--too good to be true)
ƏG

Regularization is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting.
The commonly used regularization techniques are -

1. L1 regularization or **LASSO(Least Absolute Shrinkage and Selection Operator)** regression
2. L2 regularization or **Ridge regression**
3. Dropout regularization

MACHINE LEARNING

14. Which particular algorithms are used for regularization?

Ans: Regularization is a kind of regression where the learning algorithms are modified to reduce overfitting. This may incur a higher bias but will lead to lower variance when compared to non-regularized models i.e. increases generalization of the training algorithm.

In a general learning algorithm, the dataset is divided as a training set and test set. After each epoch of the algorithm, the parameters are updated accordingly after understanding the dataset. Finally, this trained model is applied to the test set. Generally, the training set error will be less compared to the test set error. This is because of overfitting whereby the algorithm memorizes the training data and produces the right results on the training set. So the model becomes highly exclusive to the training set and fails to produce accurate results for other datasets including the test set. Algorithms like LASSO, Ridge, and Elastic-Net regression are used to reduce overfitting and increase the performance of the model on any general dataset.

Lasso Regression

Lasso, or Least Absolute Shrinkage and Selection Operator, is quite similar conceptually to ridge regression. It also adds a penalty for non-zero coefficients, but unlike ridge regression which penalizes sum of squared coefficients (the so-called L2 penalty), lasso penalizes the sum of their absolute values

(L1 penalty). As a result, for high values of λ, many coefficients are exactly zeroed under lasso, which is never the case in ridge regression. Under lasso, the loss is defined as:

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda\sum_{j=1}^{m}|\hat{\beta}_j|.$$

Ridge Regression

In Ridge Regression, the OLS loss function is augmented in such a way that we not only minimize the sum of squared residuals but also penalize the size of parameter estimates, in order to shrink them towards zero:

$$L_{ridge}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda\sum_{j=1}^{m}\hat{\beta}_j^2 = ||y - X\hat{\beta}||^2 + \lambda||\hat{\beta}||^2.$$

The only difference in ridge and lasso loss functions is in the penalty terms.

Elastic Net

Elastic Net first emerged as a result of critique on lasso, whose variable selection can be too dependent on data and thus unstable. The solution is to combine the penalties of ridge regression and lasso to get the best of both worlds. Elastic Net aims at minimizing the following loss function:

$$L_{enet}(\hat{\beta}) = \frac{\sum_{i=1}^{m}(y_i - x_i'\hat{\beta})^2}{2n} + \lambda(\frac{1-\alpha}{2}\sum_{j=1}^{m}\hat{\beta}_j^2 + \alpha\sum_{j=1}^{m}|\hat{\beta}_j|),$$

where α is the mixing parameter between ridge ($\alpha = 0$) and lasso ($\alpha = 1$).

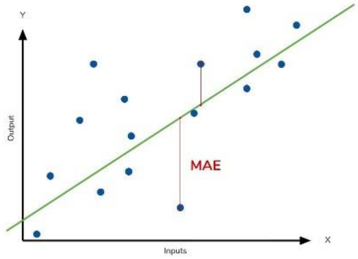
MACHINE LEARNING

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

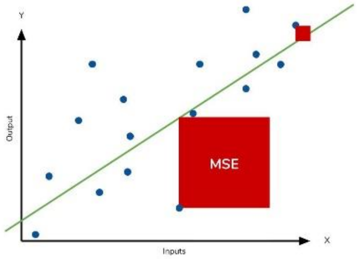
Ans: Since our model will produce an output given any input or set of inputs, we can then check these estimated outputs against the actual values that we tried to predict. We call the difference between the actual value and the model's estimate a **residual**. If our collection of residuals are small, it implies that the model that produced them does a good job at predicting our output of interest. Conversely, if these residuals are generally large, it implies that model is a poor estimator. We technically can inspect all of the residuals to judge the model's accuracy. Statisticians have developed summary measurements that take our collection of residuals and condense them into a single value that represents the predictive ability of our model. There are summary statistics like -

- Mean Absolute Error
- Mean Square Error
- Mean Absolute Percentage Error
- Mean Percentage Error

- The **mean absolute error** (MAE) describes the *typical* magnitude of the residuals. The picture below is a graphical description of the MAE. The green line represents our model's predictions, and the blue points represent our data.



- The **mean square error** (MSE) is just like the MAE, but *squares* the difference before summing them all instead of using the absolute value. Linear regression fits a line to the data by finding the regression coefficient that results in the smallest MSE. The following picture graphically demonstrates what an individual residual in the MSE might look like.



- The **mean absolute percentage error** (MAPE) is the percentage equivalent of MAE.

- The mean percentage error (MPE) equation is exactly like that of MAPE. The only difference is that it lacks the absolute value operation.