MACHINE LEARNING

1. In which of the following you can say that the model is overfitting?

ANS- (d) All of the mentioned

2. Which among the following is a disadvantage of decision trees?

ANS- B) Decision trees are highly prone to overfitting.

3. Which of the following is an ensemble technique?

ANS- A) SVM

4. Suppose you are building a classification model for detection of a fatal disease where detection of the disease is most important. In this case which of the following metrics you would focus on?

ANS- B) Sensitivity

5. The value of AUC (Area under Curve) value for ROC curve of model A is 0.70 and of model B is 0.85. Which of these two models is doing better job in classification?

ANS- B) Model B

6. Which of the following are the regularization technique in Linear Regression??

ANS- A) Ridge & D) Lasso

7. Which of the following is not an example of boosting technique?

ANS- A) Adaboost & D) Xgboost

8. Which of the techniques are used for regularization of Decision Trees?

ANS- A) Pruning & C) Restricting the max depth of the tree

9. Which of the following statements is true regarding the Adaboost technique?

ANS- B) A tree in the ensemble focuses more on the data points on which the previous tree was not performing well

10. Explain how does the adjusted R-squared penalize the presence of unnecessary predictors in the model?

ANS. The adjusted R-squared compensates for the addition of variables and **only** increases if the new predictor enhances the model above what would be obtained by probability. Conversely, it will decrease when a predictor improves the model less than what is predicted by chance.

11. Differentiate between Ridge and Lasso Regression

ANS-

Ridge and Lasso regression uses two different penalty functions for regularisation. Ridge regression uses L2 on the other hand lasso regression go uses L1 regularization technique. In ridge regression, the penalty is equal to the sum of the squares of the coefficients and in the Lasso, penalty is considered to be the sum of the absolute values of the coefficients. In lasso regression, it is the shrinkage towards zero using an absolute value (L1 penalty or regularization technique) rather than a sum of squares(L2 penalty or regularization technique).

Since we know that in ridge regression the coefficients can't be zero. Here, we either consider all the coefficients or none of the coefficients, whereas Lasso regression algorithm technique, performs both parameter shrinkage and feature selection simultaneously and automatically because it nulls out the co-efficients of collinear features. This helps to select the variable(s) out of given n variables while performing lasso regression easier and more accurate.

There is an another type of regularization method, which is ElasticNet, this algorithm is a hybrid of lasso and ridge regression both. It is trained using L1 and L2 prior as regularizer. A practical advantage of trading-off between the Lasso and Ridge regression is that it allows Elastic-Net Algorithm to inherit some of Ridge's stability under rotation.

12. What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modelling?

ANS- A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity

Most research papers consider a VIF (Variance Inflation Factor) > 10 as an indicator of multicollinearity, but some choose a more conservative threshold of **5 or even 2.5**.

13. Why do we need to scale the data before feeding it to the train the model?

ANS- To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model.

14. What are the different metrics which are used to check the goodness of fit in linear regression?

ANS - There are 3 main metrics for model evaluation in regression:

1. R Square/Adjusted R Square

2. Mean Square Error(MSE)/Root Mean Square Error(RMSE)

3. Mean Absolute Error(MAE)

1. R Square/Adjusted R Square

R Square measures how much variability in dependent variable can be explained by the model. It is the square of the Correlation Coefficient(R) and that is why it is called R Square.

2. Mean Square Error(MSE)/Root Mean Square Error(RMSE)

While R Square is a relative measure of how well the model fits dependent variables, Mean Square Error is an absolute measure of the goodness for the fit.

3. Mean Absolute Error(MAE)

Mean Absolute Error(MAE) is similar to Mean Square Error(MSE). However, instead of the sum of square of error in MSE, MAE is taking the sum of the absolute value of error.