Machine Learning

1. A) High R-squared value for train-set and High R-squared value for test-set.
2. B) Decision trees are highly prone to overfitting.
3. C) Random Forest
4. B) Sensitivity
5. B) Model B
6. A) Ridge and D) Lasso
7. A) Adaboost and D) Xgboost
8. D) All of the above
9. D) None of the above

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

Answer

Adjusted R-squared is the modified version of R-square,it shows wheather adding the new predictors will improve a regression model or not.

Adding the new predictors and comparing with the same model if R-squared value decreases than it is concluded that new predictors are not adding any value to model.

And if R-squared value increases than it can be conclude that new predictors are valuable to the model.

Machine Learning

Question 11. Differentiate between Ridge and Lasso Regression ?

Answer

Lasso and Ridge both are regularization technique used to deal with overfitted model.

Lasso will identify the features which are related to label and features which are not related to label. Lasso will simply remove the features which are donot contribute to the label internally. In other words we can say lasso as features selector.

Ridge is also regularization technique which work same as Lasso but it donot remove the less contributing features completely but it gives less importance to features which are contributing less to label.

Question12. What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modelling?

Answer

VIF (Variance inflation factor) is use to find the multicollinearity between independent variables or features.it show how one features is related with other features in the model. So by this we can avoid the overfitting problem in the model.

Generally vif value more than 5 is considered as suitable value to consider corelation between the features but many time it may vary to 10 or more than 10 depending on the model and data set provided.

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

Answer

Generally data is scaled before feeding it to the train the model.this is done to ensure that gradient decent move smothly towards it minima.when features are on the same scale than it helps gradient decent to move quickly to its minima.

For distance based algorithm also it is advised to scale the data before training it because features in data set have different magnitude so model can give higher priority to highly magnitude features.this can be avoided by scaleing the data so that all features may get equal importance.

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

Answer

There are many metrics used for checking the linear regression these are as follows

Machine Learning

1. Mean Squared Error (MSE).
2. Root Mean Squared Error (RMSE).
3. Mean Absolute Error (MAE)
4. Root Mean Squared Log Error (RMSLE)
5. R-squared
6. Adjusted R-squared

Question 15. From the following confusion matrix calculate sensitivity, specificity, precision, recall and accuracy

Answer

True Positives (TP) = 1000

True Negatives (TN) = 1200

False Positives (FP) = 50

False Negatives (FN) = 250

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy = (1000+1200)/(1000+1200+50+250)

Accuracy= 2200/2500

Machine Learning

Accuracy =0.88

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision = 1000/(1000+50)

Precision = 0.95

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall = 1000/1000+250

Recall = 0.80

Sensitivity=TP / TP + FN

Sensitivity=1000/1000+250

Sensitivity=0.80

Specificity =TN / TN + FP

Specificity=1200/1200+50

Specificity=0.96