MACHINE LEARNING

In Q1 to Q5, only one option is correct, Choose the correct option:

1. In which of the following you can say that the model is overfitting?
   A) High R-squared value for train-set and High R-squared value for test-set.
   B) Low R-squared value for train-set and High R-squared value for test-set.
   C) High R-squared value for train-set and Low R-squared value for test-set.
   D) None of the above

2. Which among the following is a disadvantage of decision trees?
   A) Decision trees are prone to outliers.
   B) Decision trees are highly prone to overfitting.
   C) Decision trees are not easy to interpret
   D) None of the above.

3. Which of the following is an ensemble technique?
   A) SVM
   B) Logistic Regression
   C) Random Forest
   D) Decision tree

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?
   A) **Accuracy**
   B) Sensitivity
   C) Precision
   D) None of the above.

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?
   A) Model A
   B) Model B
   C) both are performing equal
   D) Data Insufficient

In Q6 to Q9, more than one options are correct, Choose all the correct options:

6. Which of the following are the regularization technique in Linear Regression??
   A) **Ridge**
   B) R-squared
   C) MSE
   D) Lasso

7. Which of the following is not an example of boosting technique?
   A) Adaboost
   B) Decision Tree
   C) Random Forest
   D) Xgboost.

8. Which of the techniques are used for regularization of Decision Trees?
   A) **Pruning**
   B) L2 regularization
   C) **Restricting the max depth of the tree**
   D) All of the above

9. Which of the following statements is true regarding the Adaboost technique?
   A) We initialize the probabilities of the distribution as 1/n, where n is the number of data-points
   B) A tree in the ensemble focuses more on the data points on which the previous tree was not performing well
   C) It is example of bagging technique
   D) None of the above

Q10 to Q15 are subjective answer type questions, Answer them briefly.

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

Ans :- Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the

MACHINE LEARNING

model more than would be expected by chance. It decreases when a predictor improves the model by less than expected. Typically, the adjusted R-squared is positive, not negative. It is always lower than the R-squared.

Adding more independent variables or predictors to a regression model tends to increase the R-squared value, which tempts makers of the model to add even more variables. This is called overfitting and can return an unwarranted high R-squared value. Adjusted R-squared is used to determine how reliable the correlation is and how much it is determined by the addition of independent variables

11. Differentiate between Ridge and Lasso Regression.

Ans:- Difference between Lasso and Ridge Regression are :-

Ridge and Lasso regression uses two different penalty functions for regularisation. Ridge regression uses L2 on the other hand lasso regression go uses L1 regularisation 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).

Dimension reduction of feature space with lasso

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.

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

Ans :- The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.

Consider the following linear regression model:

$Y = \beta0 + \beta1 \times X1 + \beta2 \times X2 + \beta3 \times X3 + \varepsilon$

For each of the independent variables X1, X2 and X3 we can calculate the variance inflation factor (VIF) in order to determine if we have a multicollinearity problem.

Here's the formula for calculating the VIF for X1:

VIF (variance inflating factor) formula for the first variable in the model

R2 in this formula is the coefficient of determination from the linear regression model which has:

X1 as dependent variable

MACHINE LEARNING

X2 and X3 as independent variables
In other words, R2 comes from the following linear regression model:

$X1 = \beta0 + \beta1 \times X2 + \beta2 \times X3 + \varepsilon$

And because R2 is a number between 0 and 1:

When R2 is close to 1 (i.e. X2 and X3 are highly predictive of X1): the VIF will be very large
When R2 is close to 0 (i.e. X2 and X3 are not related to X1): the VIF will be close to 1.Therefore the range of VIF is between 1 and infinity.

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

Ans :- Scaling of the data comes under the set of steps of data pre-processing when we are performing machine learning algorithms in the data set. As we know most of the supervised and unsupervised learning methods make decisions according to the data sets applied to them and often the algorithms calculate the distance between the data points to make better inferences out of the data.

in the machine learning algorithms if the values of the features are closer to each other there are chances for the algorithm to get trained well and faster instead of the data set where the data points or features values have high differences with each other will take more time to understand the data and the accuracy will be lower. So if the data in any conditions has data points far from each other, scaling is a technique to make them closer to each other or in simpler words, we can say that the scaling is used for making data points generalized so that the distance between them will be lower.

MACHINE LEARNING

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

Ans :- R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale.

Image of a large R-squared.After fitting a linear regression model, you need to determine how well the model fits the data. Does it do a good job of explaining changes in the dependent variable? There are several key goodness-of-fit statistics for regression analysis. For instance, small R-squared values are not always a problem, and high R-squared values are not necessarily good..

Assessing Goodness-of-Fit in a Regression Model

Residuals are the distance between the observed value and the fitted value.

Linear regression identifies the equation that produces the smallest difference between all the observed values and their fitted values. To be precise, linear regression finds the smallest sum of squared residuals that is possible for the dataset.

Statisticians say that a regression model fits the data well if the differences between the observations and the predicted values are small and unbiased. Unbiased in this context means that the fitted values are not systematically too high or too low anywhere in the observation space.

However, before assessing numeric measures of goodness-of-fit, like R-squared, you should evaluate the residual plots. Residual plots can expose a biased model far more effectively than the numeric output by displaying problematic patterns in the residuals. If your model is biased, you cannot trust the results. If your residual plots look good, go ahead and assess your R-squared and other statistics.

MACHINE LEARNING

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

Actual/Predicted	True	False
True	1000	50
False	250	1200

Ans :- TP=1000
FP=50
FN=250
TN=1200

a) ACCURACY = TP+TN/TP+TN+FP+FN
1000+1200÷1000+1200+250+50
2200÷2500= 0.88 or 88%.

b) PRECISION = TP/TP+FP
1000÷1050= 0.95 OR 95%

c) SENSITIVITY= TP/TP+FN
1000÷1250= 0.8 OR 80%

d) SPECIFICITY = TN/TN+FP
1200÷1250 =0.96 OR 96%