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

Ans = C) High R-squared value for train-set and Low R-squared value for test-set.

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.

Ans = B) Decision trees are highly prone to overfitting

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

Ans = 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.

Ans = C) Precision

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

Ans = B) Model B

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

Ans = A) Ridge D) Lasso

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

Ans = B) Decision Tree c) Random Forest

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

Ans = A) Pruning B) L2 regularization

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

Ans = A) We initialize the probabilities of the distribution as 1/n, where n is the number of data-points

C) It is example of bagging technique

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 = The adjusted R-squared is a modified version of R-squared that accounts for predictors that are not significant in a regression model. In other words, the adjusted R-squared shows whether adding additional predictors improve a regression model or not. To understand adjusted R-squared, an understanding of R-squared is required. R-squared comes with an inherent problem – additional input variables will make the R-squared stay the same or increase (this is due to how the R-squared is calculated mathematically). Therefore, even if the additional input variables show no relationship with the output variables, the R-squared will increase. An example that explains such an occurrence is provided below.

Essentially, the adjusted R-squared looks at whether additional input variables are contributing to the model. Consider an example using data collected The idea is that you can change the value of one independent variable and not the others. However, when independent variables are correlated, it indicates that changes in one variable are associated with shifts in another variable. The stronger the correlation, the more difficult it is to change one variable without changing another. It becomes difficult for the model to estimate the relationship between each independent variable and the dependent variable *independently* because the independent variables tend to change in unison.

11. Differentiate between Ridge and Lasso Regression.

Ans = Ridge :- Similar to the lasso regression, ridge regression puts a similar constraint on the coefficients by introducing a penalty factor. However, while lasso regression takes the magnitude of the coefficients, ridge regression takes the square.

Lasso :- In this shrinkage technique, the coefficients determined in the linear model from equation above are shrunk towards the central point as the mean by introducing a penalization factor called the alpha α (or sometimes lamda) values. Alpha (α) is the penalty term that denotes the amount of shrinkage (or constraint) that will be implemented in the equation. With alpha set to zero, you will find that this is the equivalent of the linear regression model from equation 1.2, and a larger value penalizes the optimization function. Therefore, lasso regression shrinks the coefficients and helps to reduce the model complexity and multi-collinearity.

Considering the geometry of both the lasso (left) and ridge (right) models, the elliptical contours (red circles) are the cost functions for each. Relaxing the constraints introduced by the penalty factor leads to an increase in the constrained region (diamond, circle). Doing this continually, we will hit the center of the ellipse, where the results of both lasso and ridge models are similar to a linear regression model.

However, both methods determine coefficients by finding the first point where the elliptical contours hit the region of constraints. Since lasso regression takes a diamond shape in the plot for the constrained region, each time the elliptical regions intersect with these corners, at least one of the coefficients becomes zero. This is impossible in the ridge regression model as it forms a circular shape and therefore values can be shrunk close to zero, but never equal to zero.

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 regrerssion results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity. A variance inflation factor (VIF) provides a measure of multicollinearity among the independent variables in a multiple regression model. Detecting multicollinearity is important because while multicollinearity does not reduce the explanatory power of the model, it does reduce the statistical significance of the independent variables.

Multicollinearity creates a problem in the multiple regression model because the inputs are all influencing each other. Therefore, they are not actually independent, and it is difficult to test how much the combination of the independent variables affects the dependent variable, or outcome, within the regression model.

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 real life, if we take an example of purchasing apples from a bunch of apples, we go close to the shop, examine various apples and pick various apples of the same attributes. Because we have learned about the attributes of apples and we know which are better and which are not good also we know which attributes can be compromised and which can not. So if most of the apples consist of pretty similar attributes we will take less time in the selection of the apples which directly affect the time of purchasing taken by us. The moral of the example is if the apples every apple in the shop is good we will take less time to purchase or if the apples are not good enough we will take more time in the selection process which means that if the values of attributes are closer we will work faster and the chances of selecting good apples also strong.

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

Ans = MSE : - The Mean Squared Error measures how close a regression line is to a set of data points. It is a risk function corresponding to the expected value of the squared error loss. Mean square error is calculated by taking the average, specifically the mean, of errors squared from data as it relates to a function. A larger MSE indicates that the data points are dispersed widely around its central moment (mean), whereas a smaller MSE suggests the opposite. A smaller MSE is preferred because it indicates that your data points are dispersed closely around its central moment (mean). It reflects the centralized distribution of your data values, the fact that it is not skewed, and, most importantly, it has fewer errors (errors measured by the dispersion of the data points from its mean).

RMSE :- Root mean square error or root mean square deviation is one of the most commonly used measures for evaluating the quality of predictions. It shows how far predictions fall from measured true values using Euclidean distance. To compute RMSE, calculate the residual (difference between prediction and truth) for each data point, compute the norm of residual for each data point, compute the mean of residuals and take the square root of that mean. RMSE is commonly used in supervised learning applications, as RMSE uses and needs true measurements at each predicted data point.

In machine learning, it is extremely helpful to have a single number to judge a model's performance, whether it be during training, cross-validation, or monitoring after deployment. Root mean square error is one of the mo st widely used measures for this. It is a proper scoring that is intuitive to understand and compatible with some of the most common statistical assumptions.

15. From the following confusion matrix calculate sensitivity, specificity, precision, recall and accuracy. Actual/Predicted True False True 1000 50 False 250 1200

Ans = 1000 = True positive

50 False Positive

250 = False Negative

1200 True Negative

1) Accuracy

TP + TN / (TP + TN + FP + FN)

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

= 2200 / 2500

= 0.88

2) Recall

= TP / (TP + FN)

= 1000 / (1000 + 250)

= 1000/1250

= 0.8

3) Precision

= TP / (TP + FP)

= 1000 / (1000 + 50)

= 1000 / 1050

= 0.95

4) Specificity

= TN / (TN + FP)

= 1200 / (1200 + 50)

= 1200 / 1250

= 0.96