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 ofthe 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 themodel?
11. Differentiate between Ridge and Lasso Regression.
12. What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modelling?
13. Why do we need to scale the data before feeding it to the train the model?

MACHINE LEARNING

14. What are the different metrics which are used to check the goodness of fit in linear regression?
15. From the following confusion matrix calculate sensitivity, specificity, precision, recall and accuracy.

Actual/Predicted	True	False
True	1000	50
False	250	1200

9. (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
10. In linear regression model R-squared on training data increases as we increase the number of predictors because as we increase the number of predictors we are adding more information to the model so the r-squared increases on the training dataset. But this can lead to overfitting of the model. To prevent the overfitting in spite of r squared we use adjusted r-squared. Its formula is given as follows: As we can see from the above formula if R-squared remains constant and number of predictors increase then adjusted R-Squared decreases. So, if R-squared is not increased significantly on adding predictors, adjusted R Squared will decrease. So, in this way adjusted r-squared will penalize the presence of unnecessary predictors.
11. Ridge Regression: It is technique of regularization in linear regression. It regularizes the model by using L2 regularization which tries to minimize the sum of squares of the magnitude of the coefficients along with the error. The cost function in ridge regression looks like this:
Cost function= MSE+ α * (sum of square of coefficients)^2
Where, MSE is the mean squared error. α is the regularization constant. The regularization increases with increment in α. Lasso Regression: It is also technique of regularization in linear regression. It regularizes the model by using L1 regularization which tries to minimize the sum of the magnitude of the coefficients along with the error. The cost function in lasso regression looks like this:
Cost function= MSE+ α * |sum of magnitude of coefficients|

Where, MSE is the mean squared error. α is the regularization constant. The regularization increases with increment in α.
12. VIF: VIF stands for Variance Inflation Factor. VIF determines the strength of the correlation between the independent features. It is predicted by taking a variable and regressing it against every other feature in the dataset. R2 value is determined to find out how well an independent feature is described by the other independent features. A high value of R2 means that the feature is highly correlated with the other features. This is captured by the VIF which is denoted below: Where Ri 2 = R2 score of ith feature against all other features So, the closer the R2 value to 1, the higher the value of VIF and the higher the multicollinearity with the particular independent variable. Generally, if VIF is less than 4, the feature is acceptable to be a part of model otherwise it is dropped.

13. The following are the reasons to scale the data before feeding it to train the model: · The gradient descent algorithm which is used to reach the optimal solution in most of the cases, it reached the optimal solution much faster if all the features are at the same scale. That's why scaling helps to reach the optimal solution. · If the features in the training dataset are on different scales, then during training the features with large scales will be favored over there in order to minimize the loss. That's why we do Scaling to puts all the features on the same scale. 14. Some of the metrics used to check the goodness of fit in linear regression are: · MSE- mean squared error. As the name suggests it is the average value of squares of the errors made by model on a dataset. MSE = (sum (Ytrue – Yexpected)2)/n · R-squared. It is defined as the variance explained by the model/Total variance of the dataset. · Adjusted R-squared: It takes in to account both the R-squared as well as the number of predictors in the model. That is it considers both the variance explained by the model as well as the number of predictors used by the model to explain that variance.

Fp: False positives = 250 Fn: False negatives = 50 Tp: True positives = 1000 Tn: True negatives = 1200
Sensitivity: Tp/(Tp+Fn) = 1000/1050 = 0.9523
Specificity: Tn/(Tn+Fp) = 1200/1450 = 0.8275
Precision: Tp/(Tp+Fp) = 1000/(1000+250) = 0.8
Recall: Tp/(Tp+Fn) = 1000/(1000+50) = 0.9523
Accuracy: (Tp+Tn)/(Tp+Tn+Fp+Fn) = (1000+1200)/(1000+1200+250+50) = 0.8