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 over-fitting?
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: C)Random Forest

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: 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?
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)Restricting the max depth of the tree

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

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

10.)Explain how does the adjusted R-squared penalize the presence of unnecessarypredictors in the model?
Ans:

Adjusted R squred is a modified version of R-squared that has been adjusted for the number of predictors in the model. Adjusted R squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance.

11.)Differentiate between Ridge and Lasso Regression.
Ans:

Lasso stands for least absolute shrinkage and selection operator.It adds penalty term to the cost function.The difference between ridge and lasso is that it tends to make coefficients to absolute zero as compared to ridge which never sets the value of coefficient to absolute zero.

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) is a measure of colinearity among predictor variables within a multiple regression. It is calculated by taking an independent variable and regressing it against every other predictor in the model.

$$\text{VIF} = \frac{1}{1 - R_i^2}$$

Including highly correlated variables in your model can lead to overfitting. If we overfit, then the model performs extraordinarily well on the training data but doesn't generalize well when we try to use it on new data.

Small VIF values, VIF < 3, indicate low correlation among variables under ideal conditions. The default VIF cutoff value is 5; only variables with a VIF less than 5 will be included in the model. However, many sources say that a VIF of less than 10 is acceptable.

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

Ans:

To ensure that 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.Having features on a similar scale can help the gradient descent converge more quickly towards the minima.

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

Ans :

Five metrics gives us some hints about goodness of fit of our model.They are mean absolute error, root mean squared error, relative absolute error, relative squared error and coefficient of determination(R squared)

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

Actual/Predicted	True	False
True	1000 TP	50 FP
False	250 FN	1200 TN

Ans:

Accuracy = TP+TN/TP+TN+FP+FN = 1000+1200/1000+1200+250+50 =2200/2500 = 0.88

Recall = TP/TP+FN = 1000/1000+50 = 0.95

Precision = TP/TP+FP = 1000/1000+250 = 0.8

Sensitivity = TP/TP+FN = 1000/1000+50 = 0.95

Specificity = TN/FP+TN = 1200/50+1200 = 1200/1250 = 0.96