

MACHINE LEARNING

1. In which of the following you can say that the model is overfitting?

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?

Ans. (B) Decision trees are highly prone to overfitting.

3. Which of the following is an ensemble technique?

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?

Ans. (B) Sensitivity/recall

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?

Ans. (B) Model B

6. Which of the following are the regularization technique in Linear Regression?

Ans. (A) Ridge (D) Lasso

7. Which of the following is not an example of boosting technique?

Ans. (B) Decision Tree (C) Random Forest

8. Which of the techniques are used for regularization of Decision Trees?

Ans. (A) Pruning (C) Restricting the max depth of the tree

9. Which of the following statements is true regarding the Adaboost technique?

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

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

Ans. The adjusted R-squared recognition of loss for the addition of variable and only increases, if the new predictor enhances the model above what would be obtained by probability. Conversely, it will decrease when a predictor improves the model less than what is predicted by chance.

11. Differentiate between Ridge and Lasso Regression

Ans.

Sn	Lasso Regression	Sn	Ridge Regression
1	When model is penalized for the sum of absolute value of weight will be reduced.	1	The model is penalized for sum of squared value of weight,

2	Many tend to be zeros	2	Weights not only tend to have small absolute value.
---	-----------------------	---	---

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

Ans. Variance Inflation Factor (VIF): -

This is a measure of the amount of multicollinearity in regression analysis, multicollinearity exists when there is a correlation between multiple independent variable(features) in a regression model. The default VIF cut-off is 5, but many source says VIF cut-off is 10.

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

Ans. Scale independent variable before feeding data to train model because it is made to easy for model to understand data trends. After scale data our machine process and learn well and working well.

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

Ans. Different metrics are –

1. R-square score
2. MSE
3. RMSE
4. MAE

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

Actual/Predicted	True	False
True	1000	50
False	250	1200

Ans.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+FP} = \frac{2200}{2500} = 0.88$$

$$\text{Sensitivity} = \text{recall} = \frac{TP}{TP+FN} = \frac{1000}{1000+250} = 0.8$$

$$\text{Specificity} = \frac{TN}{TN+FP} = \frac{1200}{1250} = 0.96$$

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{1000}{1050} = 0.95$$