**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

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: 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. 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. Differentiate between Ridge and Lasso Regression.

Ans: 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.

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.

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

Ans: 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. R2value 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. Generally, if VIF is less than 4, the feature is acceptable to be a part of model otherwise it is dropped.

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

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

Ans: MSE- mean squared error. As the name suggests it is the average value of squares of the errors made by model on a dataset.

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.

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

|  |  |  |
| --- | --- | --- |
| Actual/Predicted | True | False |
| True | 1000 | 50 |
| False | 250 | 1200 |

Ans: 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) = 0.88

**Statistics**

1. Which of the following can be considered as random variable?

Ans: d) All of the mentioned

2. Which of the following random variable that take on only a countable number of possibilities?

Ans: (a)Discrete

3. Which of the following function is associated with a continuous random variable?

Ans: (a)pdf ( Probability density function)

4. The expected value or \_\_\_\_\_\_\_ of a random variable is the center of its distribution.

Ans: (c)Chi-squared distribution

5. Which of the following of a random variable is not a measure of spread?

Ans: (c)Mean

6. The \_\_\_\_\_\_\_\_\_ of the Chi-squared distribution is twice the degrees of freedom.

Ans: (a)Varience

7. The beta distribution is the default prior for parameters between \_\_\_\_\_\_\_\_\_\_\_\_

Ans: c) 0 and 1

8. Which of the following tool is used for constructing confidence intervals and calculating standard errors for difficult statistics?

Ans: b) bootstrap

9. Data that summarize all observations in a category are called \_\_\_\_\_\_\_\_\_\_ data.

Ans: b) summarized

10. What is the difference between a boxplot and histogram?

Ans: Histograms are bar charts that show the frequency of a numerical variable’s values and are used to approximate the probability distribution of the given variable.

Boxplot gather other information like the quartiles, the range, and outliers. Boxplots are especially useful when you want to compare multiple charts at the same time because they take up less space than histograms.

11. How to select metrics?

Ans: The metric(s) chosen to evaluate a machine learning model depends on various factors:

a) To find a regression or classification task.

b) To find the business objective like precision vs recall

c) To know distribution of the target variable.

There are a number of metrics that can be used, including adjusted r-squared, MAE, MSE, accuracy, recall, precision, f1 score, and the list goes on.

12. How do you assess the statistical significance of an insight?

Ans: First, you would state the null hypothesis and alternative hypothesis. Second, you would calculate the p-value, the probability of obtaining the observed results of a test assuming that the null hypothesis is true. Last, you would set the level of the significance (alpha) and if the p-value is less than the alpha, you would reject the null — in other words, the result is statistically significant.

13. Give examples of data that does

Ans: Any type of categorical data won’t have a gaussian distribution or lognormal distribution.

Exponential distributions — eg. the amount of time that a car battery lasts or the amount of time until an earthquake occurs. not have a Gaussian distribution, nor log-normal.

14. Give an example where the median is a better measure than the mean.

Ans: When there are a number of outliers that positively or negatively skew the data.

15. What is the Likelihood?

Ans: The probability of some of the observed outcomes under specific parameter values is regarded as the likelihood of the set of parameter values under certain observed outcomes.