Great Learning

POWER AHEAD

Supervised Learning Classification Interview Questions

1. What is XGBoost? What is the use of the Hessian in the XGBoost algorithm?
2. Explain the Decision Tree algorithm. How does the algorithm decide the root node?
3. What are the metrics to split the information in a decision tree?
4. What is the main disadvantage of Decision Trees? What effect do outliers have on Decision Trees?
5. Define entropy, how is it calculated? What are the maximum and minimum possible values of entropy?
6. Define Gini Index? Why does Scikit Learn use Gini Index and not Entropy in the Decision Tree?
7. Explain Random forest algorithm. How does the Random Forest reduce model variance?
8. Why is a Random forest Superior to a Decision Tree?
9. Suppose we produce 10 bootstrapped samples from a data set containing two classes 'fraud' and 'not fraud'. Next we fit a Decision Tree to each bootstrapped sample and for a specific value of X , produce 10 estimates of P(fraud/X) as 0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.65, 0.7 and 0.75.
10. What is the final classification based on the majority vote approach?
11. What is the working of logistic regression? Is it a Discriminative or Generative Algorithm?
12. Name evaluation metrics of regression/classification models and their Scikit learn functions?
13. Explain the working of ensemble methods such as bagging, boosting, and stacking.
14. Difference between precision/ recall/ f1 score. Which is the best metric?
15. What is the difference between K-means clustering and KNN?
16. How does Boosting differ from Bagging?
17. What is ADA boosting? How does ADA Boost differ from Gradient Boosting?
18. Explain the Gradient Descent Algorithm for finding the Global Minimum?
19. What are independent variables and categorical variables? Highlight the key differences.
20. How do you identify fraudulent cases when customers are returning the product, and what features you will use to build an ML model
21. If you have a large number of missing values in your dataset, what are the ways of dealing?
22. Why is the Naïve Bayes having the word Naïve?
23. Is it better to have too many false positives or too many false negatives? Explain.
24. Explain how k-fold cross-validation is implemented? What is the advantage of k-fold cross-validation relative to the validation set approach?
25. IF we have a very large dataset for an ML task, which classification algorithm will be more efficient in terms of Time Order Complexity, Naïve Bayes, or KNN. Why?
26. What is the working of Naïve Bayes? Is it a Discriminative or Generative Algorithm?
27. What does the K in KNN stand for and how can we find the optimum value of K?
28. How do we tune hyperparameters for Classification ML Algorithms using Grid Search?

29. What is the difference between Gini Impurity and Entropy in a Decision Tree?
30. Why is the word "Naive" used in the Naive Bayes classifier?
31. While calculating the probability of a given situation, what error can we run into in Naïve Bayes and how can we solve it?
32. What factors can attribute to the popularity of Logistic Regression?
33. What is the difference between the outputs of the Logistic model and the Logistic function?
34. Can we solve the multiclass classification problems using Logistic Regression? If Yes then How?
35. Why can’t we use Linear Regression in place of Logistic Regression for Binary classification?
36. How does a Random Forest Algorithm give predictions on an unseen dataset?
37. Since Ensemble Learning provides better output most of the time, why do you not use it all the time?
38. How does Ensemble Learning tackle the No-Free Lunch Dilemma?
39. What is the difference between Maximum Likelihood Estimation and Gradient Descent?
40. What would happen if there are untreated outliers in the dataset while modelling with Decision Trees?