DataInterview.com

70 Machine Learning Interview Questions

Interview Questions asked in FAANGs, startups and consulting firms

ML Breath

1. What is the difference between supervised and unsupervised learning?
2. Can you explain the concept of overfitting and underfitting in machine learning models?
3. What is cross-validation? Why is it important?
4. Describe how a decision tree works. When would you use it over other algorithms?
5. How do you handle missing or corrupted data in a dataset?
6. What is the bias-variance tradeoff?
7. What is the difference between bagging and boosting?
8. How would you validate a model you created to generate a predictive analysis?
9. Can you explain the principle of a support vector machine (SVM)?
10. What are some of the advantages and disadvantages of a neural network?
11. How does the k-means algorithm work?
12. Can you explain the difference between L1 and L2 regularization methods?
13. What is principal component analysis (PCA) and when is it used?
14. Can you describe what an activation function is and why it is used in an artificial neural network?
15. How would you handle an imbalanced dataset?
16. Can you explain the concept of "feature selection" in machine learning?
17. What is the difference between stochastic gradient descent (SGD) and batch gradient descent?
18. Can you describe how a convolutional neural network (CNN) works?
19. How do you handle categorical variables in your dataset?
20. What is reinforcement learning? Can you give an example of where it could be used?

ML Breath Cont'd

21. Describe a situation where you had to handle missing data. What techniques did you use?
22. How would you evaluate a machine learning model's performance?
23. Which metrics would you use for binary classification? How about for multi-class classification or regression?
24. Describe a scenario where you chose one algorithm over another based on its performance characteristics.
25. How do you handle categorical variables when preparing data for machine learning?
26. How would you deploy a machine learning model in a production environment?
27. Describe a situation where you had to tune hyperparameters. Which methods did you use and why?
28. How do you ensure that your machine learning model is not just memorizing the training data?
29. Describe a situation where ensemble methods improved your model's performance.
30. How do you deal with large datasets that don't fit into memory?
31. What is the role of the cost function in machine learning algorithms?
32. What is the curse of dimensionality? How do you avoid this?

ML Depth

Decision Tree

1. What is entropy and how is it used in decision trees?
2. How do decision trees handle continuous numerical variables?
3. What is information gain and how does it relate to decision tree construction?
4. Explain the concept of pruning in decision trees.
5. What are the primary differences between the CART, ID3, and C4.5 decision tree algorithms?
6. How do decision trees deal with missing values during both training and prediction?

Random Forest

1. Explain the concept of bootstrapping in relation to random forests.
2. How does feature selection work in a random forest as compared to a single decision tree?
3. Why might a random forest be less prone to overfitting than a single decision tree?
4. How can you estimate the importance of a feature using a random forest?
5. What are the key hyperparameters to tune in a random forest model?

XGBoost

1. What is gradient boosting and how does XGBoost utilize it?
2. Explain the differences between XGBoost and a traditional gradient boosting machine (GBM).
3. How does XGBoost handle regularization?
4. What are the key advantages of using XGBoost over other boosting methods?
5. How does XGBoost handle missing values during training?

ML Depth

Neural Networks:

1. Describe the backpropagation algorithm.
2. How does a convolutional neural network (CNN) differ from a regular feedforward neural network?
3. What is dropout and why might you use it when training a neural network?
4. How does the vanishing/exploding gradient problem impact neural network training, and how can it be mitigated?
5. Explain the concept and purpose of an activation function in neural networks. Can you name a few common activation functions?
6. Describe the difference between batch normalization and layer normalization.

Regression Models:

1. What are the assumptions behind a linear regression model?
2. How do you handle multicollinearity in a regression model?
3. Explain the difference between ridge regression and lasso regression.
4. In logistic regression, how do you interpret the coefficients of the predictors?
5. What is the purpose of the R-squared statistic in a linear regression model?
6. Describe the difference between simple linear regression and multiple linear regression.
7. What are the key differences between a linear regression and a polynomial regression model?
8. How do you detect and handle outliers in regression analysis?

ML System Design

1. **[Netflix]** You're tasked with improving the recommendation engine for a streaming service. How would you design a system that suggests relevant shows or products to users based on their past behavior?

2. **[Google]** Design a scalable system to categorize billions of user photos into predefined categories (e.g., landscapes, portraits, events). How would you ensure minimal latency when a user uploads a new photo?

3. **[Apple]** How would you design a system to improve the accuracy of voice command recognition in noisy environments?

4. **[Google]** Imagine you're designing a new search algorithm for a social media platform. How would you design a system that ranks user-generated content in search results based on relevance, timeliness, and user engagement?

5. **[Facebook]** How would you design a machine learning system that predicts the click-through rate (CTR) of ads shown to users, ensuring that users find the ads relevant and not intrusive?

6. **[Amazon]** Design a real-time system to detect potentially fraudulent transactions on an e-commerce platform. How would you ensure the balance between blocking genuine transactions and letting fraudulent ones through?

7. **[YouTube]** How would you design an ML system to automatically detect and filter out harmful content or misinformation from a platform with billions of posts?

8. **[Netflix]** Design a system to predict the optimal bit rate for streaming content to users based on their internet speed, device type, and content preferences, ensuring minimal buffering.