Machine Learning Interview Questions

Beginner 🔠

1. What is Machine Learning?
2. Explain the difference between supervised and unsupervised learning.
3. What is the difference between classification and regression in machine learning?
4. What are the Machine Learning Project steps?
5. Can you explain the steps involved in the data pre-processing process?
6. Explain the term 'bias' in the context of machine learning models.
7. What is the importance of feature scaling in machine learning?
8. Can you explain the concept of regularisation in machine learning?
9. What is the difference between L1 and L2 regularisation?
10. What is the purpose of a confusion matrix in classification tasks?
11. How do you handle a situation where the data is too imbalanced?
12. What is the purpose of the Support Vector Machine (SVM) algorithm?
13. What is the purpose of the F1 score metric in evaluating classification models?

Intermediate 🚀

1. How do you handle a large volume of data that cannot fit into memory?
2. How do you handle a situation where there are too many features compared to the number of observations?
3. How to Find the Best Fit Line in Linear Regression using Gradient Descent
4. Why is MSE preferred over MAE in Linear Regression?
5. How are nominal and ordinal categorical variables encoded in machine learning, and why are different techniques used for each?

6. Suppose you're building a linear regression model and observe that two of your independent variables are highly correlated with each other.
   What potential issues could this cause in your model, and how would you address them?
7. Can you explain the concepts of bias and variance in machine learning, and how they affect model performance?
   Additionally, how would you identify and address a model that is underfitting or overfitting based on these concepts?
8. You’re working on a marketing campaign for an e-commerce platform. The goal is to predict whether a user will click on a promotional email or not.
   You decide to use Logistic Regression. How would you approach this problem, and what are the key considerations when using Logistic Regression in this context?
9. You’ve built a classification model for detecting fraud transactions. The dataset is highly imbalanced, with only 1% of the transactions being fraudulent. Your model reports 99% accuracy.
   Would you consider this a good model? Why or why not? What alternative metrics would you use and why?
10. What is a Kernel in SVM (Support Vector Machine)?
11. You’re building a machine learning model to detect cancer from medical scans. The model occasionally predicts false negatives (i.e., says "no cancer" when cancer is present).
    Which evaluation metric would you prioritize in this case and why?

Advanced 💧

1. What happens when we use MSE as the cost function for Logistic Regression, which uses the sigmoid function?

2. You’ve built a decision tree classifier and noticed that it's perfectly fitting your training data but performing poorly on unseen data.
   You’ve already tried limiting the tree depth and pruning.
   Can you explain why decision trees are prone to overfitting, and how ensemble methods like Random Forest or Gradient Boosting help overcome this problem? Also, how does feature importance differ between a single decision tree and an ensemble?
3. You're working with a small and imbalanced dataset. You're using standard K-Fold Cross-Validation to evaluate your model.
   What challenges might this pose, and how can Stratified K-Fold help address them?
4. You're working on tuning a Random Forest model to predict loan default. Your hyperparameter space includes:

   - Number of estimators: [50, 100, 200, 500]
   - Max depth: [None, 10, 20, 50]
   - Min samples split: [2, 5, 10]
   - Max features: ['sqrt', 'log2', 0.5, 0.8]

   You initially use GridSearchCV to explore this space, but the process becomes extremely slow and computationally expensive. Your manager asks you to propose a faster tuning strategy that still delivers strong performance without exhaustively searching all combinations.

   How would you approach this challenge? What are the limitations of your current approach, and what alternative would you recommend? Discuss the trade-offs involved.

5. You’ve applied K-Means clustering to customer behaviour data. You're unsure if the number of clusters k is optimal. The dataset has outliers and features with different scales.

How would you choose the right number of clusters? What are the limitations of K-Means, and how would you address them? How would you evaluate clustering quality?

6. How Does DBSCAN Overcome the Limitations of K-Means Clustering?
7. Define how you would design an efficient and secure API to serve a trained machine learning model.

   Your answer should include:

   - ✓ How to load the model efficiently
   - ✓ How to structure input/output handling
   - ✓ How to validate incoming data
   - ✓ How to secure the API endpoint

8. Why Do We Need Kubernetes for Orchestration?
9. What is a CI/CD pipeline, and how does Jenkins facilitate it in production-grade ML or software deployment workflows?
10. What is Data Drift and Concept Drift?

•