AI & ML Interview Questions with Answers

# I. GENERAL AI & ML CONCEPTS

* Q: What is the difference between AI, Machine Learning, and Deep Learning?

A: AI is a broad field of creating intelligent machines. Machine Learning (ML) is a subset of AI that uses statistical methods to enable machines to learn from data. Deep Learning (DL) is a subset of ML that uses neural networks with many layers.

* Q: What are the main types of Machine Learning?

A: Supervised Learning, Unsupervised Learning, Semi-supervised Learning, and Reinforcement Learning.

* Q: Explain the difference between supervised and unsupervised learning.

A: Supervised learning uses labeled data to train models, while unsupervised learning finds hidden patterns in unlabeled data.

* Q: What is overfitting and underfitting?

A: Overfitting occurs when a model learns the training data too well and performs poorly on new data. Underfitting happens when a model is too simple to capture the underlying pattern of the data.

* Q: How do you evaluate a machine learning model?

A: Using metrics such as accuracy, precision, recall, F1-score for classification; MSE, RMSE, MAE for regression.

* Q: What is the bias-variance trade-off?

A: It's the balance between a model's ability to minimize bias (error due to erroneous assumptions) and variance (error due to complexity).

* Q: Explain cross-validation and why it's important.

A: Cross-validation splits the data into training and validation sets multiple times to ensure the model generalizes well to unseen data.

* Q: What are precision, recall, F1-score, and accuracy?

A: Accuracy is the ratio of correct predictions. Precision is the ratio of true positives to all predicted positives. Recall is the ratio of true positives to all actual positives. F1-score is the harmonic mean of precision and recall.

* Q: What is the difference between classification and regression?

A: Classification predicts categorical outcomes, while regression predicts continuous values.

* Q: What are some real-world applications of AI/ML?

A: Examples include recommendation systems, fraud detection, medical diagnosis, autonomous vehicles, and chatbots.

# II. DATA PREPROCESSING & FEATURE ENGINEERING

* Q: How do you handle missing data in a dataset?

A: Options include removing rows/columns with missing values, imputing with mean/median/mode, or using algorithms that support missing values.

* Q: What is normalization and standardization?

A: Normalization scales values between 0 and 1. Standardization rescales data to have a mean of 0 and a standard deviation of 1.

* Q: What is one-hot encoding? When would you use it?

A: One-hot encoding converts categorical variables into binary vectors. It's used when the categorical variable has no ordinal relationship.

* Q: How do you handle categorical variables?

A: Techniques include one-hot encoding, label encoding, ordinal encoding, and using embeddings.

* Q: What is feature selection and why is it important?

A: Feature selection is the process of selecting the most relevant features for model training. It reduces overfitting, improves accuracy, and decreases training time.

* Q: What is dimensionality reduction? Explain PCA.

A: Dimensionality reduction reduces the number of features. PCA (Principal Component Analysis) transforms features into a smaller number of uncorrelated variables.

* Q: What is the curse of dimensionality?

A: As the number of features increases, the data becomes sparse, making it hard for models to generalize.

* Q: What is feature scaling and when should it be applied?

A: Feature scaling standardizes the range of features. It's essential for algorithms like SVM, KNN, and gradient descent-based models.

* Q: How do you deal with imbalanced datasets?

A: Techniques include resampling (oversampling/undersampling), SMOTE, or using algorithms that handle imbalance well.

* Q: Explain the role of EDA (Exploratory Data Analysis) in ML.

A: EDA helps understand data distributions, relationships, outliers, and guides feature engineering and model selection.

# III. MACHINE LEARNING ALGORITHMS

* Q: How does the Decision Tree algorithm work?

A: It splits data into subsets based on feature values to create a tree of decisions leading to predictions.

* Q: What is the difference between bagging and boosting?

A: Bagging builds multiple models independently (e.g., Random Forest). Boosting builds models sequentially to correct previous errors (e.g., XGBoost).

* Q: Explain how the K-Nearest Neighbors algorithm works.

A: KNN classifies data based on the majority class of its k closest neighbors in the feature space.

* Q: What is the intuition behind Support Vector Machines?

A: SVM finds the hyperplane that best separates classes with the maximum margin.

* Q: How does Naive Bayes classifier work?

A: It applies Bayes’ theorem with the assumption of feature independence.

* Q: What is the difference between Random Forest and XGBoost?

A: Random Forest uses bagging and builds trees in parallel. XGBoost uses boosting and builds trees sequentially.

* Q: How do gradient descent and stochastic gradient descent differ?

A: Gradient Descent uses the whole dataset to compute gradients. SGD uses one or few samples per iteration, which is faster and more scalable.

* Q: Explain logistic regression and where it’s used.

A: Logistic regression models the probability of a binary outcome using a logistic function. Used in classification tasks.

* Q: What are ensemble models?

A: Ensemble models combine predictions from multiple models to improve performance (e.g., bagging, boosting, stacking).

* Q: What is the difference between L1 and L2 regularization?

A: L1 (Lasso) adds absolute value of coefficients. L2 (Ridge) adds squared values. L1 can zero out coefficients, useful for feature selection.

# IV. DEEP LEARNING & NEURAL NETWORKS

* Q: What is a perceptron?

A: A perceptron is the simplest neural network unit with weights, bias, and an activation function for binary classification.

* Q: How do activation functions like ReLU, Sigmoid, and Tanh work?

A: ReLU returns max(0,x), Sigmoid squashes input to [0,1], Tanh squashes to [-1,1]. Each helps in introducing non-linearity.

* Q: What are epochs, batch size, and learning rate?

A: Epochs: full passes over data. Batch size: number of samples per update. Learning rate: step size during optimization.

* Q: What is the vanishing gradient problem?

A: In deep networks, gradients become very small during backpropagation, making training difficult.

* Q: What is the difference between CNN and RNN?

A: CNNs process grid-like data (images), RNNs handle sequential data (text, time series).

* Q: How does an LSTM work and where is it used?

A: LSTM is a type of RNN with gates to manage long-term dependencies. Used in NLP, time-series prediction.

* Q: What are convolutional layers in CNN?

A: They apply filters over input to detect spatial features in images.

* Q: What is transfer learning?

A: Using a pre-trained model on a new task with fine-tuning, saving time and resources.

* Q: What is dropout and why is it used?

A: Dropout randomly disables neurons during training to prevent overfitting.

* Q: How do you prevent overfitting in deep learning models?

A: Techniques include dropout, regularization, data augmentation, and early stopping.

# V. NATURAL LANGUAGE PROCESSING

* Q: What is tokenization in NLP?

A: Tokenization splits text into words, subwords, or sentences, helping convert text to a numerical format.

* Q: How do word embeddings like Word2Vec or GloVe work?

A: They map words to dense vectors capturing semantic relationships, learned from large corpora.

* Q: What is the difference between stemming and lemmatization?

A: Stemming crudely removes suffixes. Lemmatization uses vocabulary and grammar to get base form.

* Q: What is TF-IDF and why is it used?

A: TF-IDF reflects term importance in a document relative to a corpus, useful for feature extraction.

* Q: What are Transformers in NLP?

A: Transformers use self-attention to process sequences in parallel. Foundation of BERT, GPT, etc.

# VI. MODEL DEPLOYMENT & MLOps

* Q: How do you save and load a trained model in Python?

A: Using `joblib` or `pickle` to serialize and deserialize the model object.

* Q: What is model drift and how do you monitor it?

A: Model drift is performance degradation due to data changes. Monitor using performance metrics over time.

* Q: How do you deploy a machine learning model as an API?

A: Use frameworks like Flask or FastAPI to create endpoints serving predictions.

* Q: What are common tools for deploying ML models (Flask, Docker, Streamlit)?

A: Flask for APIs, Docker for containerization, Streamlit for web apps, all help in deployment.

* Q: Explain the CI/CD pipeline in MLOps

A: CI/CD in MLOps automates code testing, model validation, deployment, and monitoring for reproducibility and scalability.