1. Core ML Concepts

Q1. Supervised vs Unsupervised Learning

- **Supervised:** labeled data (classification, regression)
- **Unsupervised:** no labels (clustering, dimensionality reduction)

Q2. Classification vs Regression

- **Classification:** predict categories (spam/ham)
- **Regression:** predict continuous values (price prediction)

Q3. Bias vs Variance

- **High bias → underfitting**
- **High variance → overfitting**
- Best model balances both.

---

2. Model Evaluation Metrics

Classification Metrics

- **Accuracy:** correct predictions / total

- **Precision:** TP / (TP + FP)
- **Recall:** TP / (TP + FN)
- **F1-score:** harmonic mean of precision and recall

Regression Metrics

- **MAE:** average absolute error
- **MSE/RMSE:** squared error, RMSE is interpretable
- **R^2 score:** variance explained by model

---

3. Overfitting Control

Ways to reduce overfitting:

- Cross-validation
- Regularization (L1, L2)
- Dropout (DL)
- Early stopping
-

More data / data augmentation

---

4. Common Algorithms (Interview)

Logistic Regression

- Used for binary classification
- Output is probability using sigmoid function

Decision Tree

- Rule-based model
- Can overfit if deep

Random Forest

- Ensemble of trees
- More stable, reduces overfitting

Gradient Boosting (XGBoost/LightGBM)

- Boosting improves weak learners
- Often gives best performance in tabular data

---

5. Deep Learning Basics

Activation Functions

- **ReLU:** most common, fast
- **Sigmoid:** probability output
- **Tanh:** centered at 0

Loss Functions

- **Cross-Entropy:** classification
- **MSE:** regression

Optimizers

- **SGD:** simple, stable
- **Adam:** fast convergence, most used

---

6. NLP & LLM Basics

Q: What is Tokenization?

Splitting text into smaller units (tokens) for model input.

Q: What is Embedding?

Vector representation of words/sentences capturing meaning.

Q: What is RAG?

Retrieval Augmented Generation = retrieve relevant info from docs and generate answer using LLM.

---

7. Important Interview Questions (Rapid Fire)

- What is confusion matrix?
- What is data leakage?
- What is feature scaling and why needed?
- Explain train-test split
- What is cross validation?
- What is class imbalance and how to handle it?
- What is ROC-AUC?
- What is hyperparameter tuning?

---

8. ML Deployment & Monitoring

Q: What is Data Drift?

When real-world input data changes over time.

Q: What is Concept Drift?

When the relationship between input and output changes.

Monitoring includes:

- accuracy drop
- drift detection
- inference latency
- data quality checks