ML Interview Preparation Guide

Q1. Bias vs Variance Tradeoff

Bias refers to the error introduced by approximating a complex problem with a simpler model. High bias leads to underfitting.

Variance refers to how much a model changes when trained on different subsets of data. High variance leads to overfitting.

To balance them:

- Use regularization (L1/L2) for linear models.
- Use ensemble methods like Random Forest or Boosting.
- Apply cross-validation.
- Reduce model complexity or gather more data.

Q2. Gradient Descent and its Types

Gradient Descent is an optimization algorithm used to minimize a model's loss function by updating model parameters in the opposite direction of the gradient.

Types:

- Batch Gradient Descent: Uses the entire dataset for each update.
- Stochastic Gradient Descent (SGD): Uses one sample per update.
- Mini-Batch Gradient Descent: Uses small batches of data. It is most commonly used in deep learning.

Q3. Regularization (L1 vs L2)

Regularization adds a penalty term to the loss function to reduce overfitting.

- L1 Regularization (Lasso): Adds the absolute values of coefficients. Promotes sparsity and feature selection.
- L2 Regularization (Ridge): Adds squared values of coefficients. Shrinks coefficients but keeps all features.

Use L1 when feature selection is needed; L2 when features are informative and multicollinearity is present.

Q4. Ensemble Methods: Bagging vs Boosting

Bagging trains models in parallel on different subsets of data. It reduces variance. Example: Random Forest.

Boosting trains models sequentially, each new model correcting the errors of the previous. It reduces bias and variance. Examples: AdaBoost, Gradient Boosting.

Bagging is preferred for high-variance models; Boosting for more accuracy and imbalanced data.

Q5. Decision Tree Split Criteria

Decision Trees use:

- Gini Index: 1 - sum(p_i^2), where p_i is the probability of each class.
- Entropy: -sum(p_i * log2(p_i)).
- Information Gain: Reduction in entropy after a split.

For regression tasks, splits are based on variance reduction.

For classification, features with lowest impurity (Gini or Entropy) are selected for splits.

Q6. PCA vs LDA

PCA (Principal Component Analysis):

- Unsupervised method.
- Maximizes variance.
- Reduces dimensionality without using labels.

LDA (Linear Discriminant Analysis):

- Supervised method.
- Maximizes class separability.
- Projects data to maximize between-class variance and minimize within-class variance.

Use PCA for general compression; LDA for classification tasks with labeled data.

Q7. Cross-Validation and Stratified K-Fold

Cross-validation helps evaluate model performance more reliably by training and testing on different data splits.

- K-Fold CV: Divides data into k folds, each used once as validation.
- Stratified K-Fold: Ensures class distribution is maintained in each fold, important for imbalanced classification.

Q8. Random Forest vs Gradient Boosting

Random Forest:
- Bagging method using many decision trees in parallel.
- Reduces variance.

Gradient Boosting:
- Boosting method where trees are built sequentially.
- Each tree corrects errors of the previous one.
- Can handle imbalanced data better using class weights like 'scale_pos_weight'.

Use GBM when accuracy and class balance are critical.

Q9. ROC-AUC vs Precision-Recall Curve

ROC-AUC plots True Positive Rate vs False Positive Rate. It is suitable for balanced datasets.

Precision-Recall Curve plots Precision vs Recall. It is better for imbalanced datasets because it focuses on the performance of the positive class.

For imbalanced data, use Precision-Recall curves as they provide a clearer picture of model performance.

Q10. Handling Overfitting

Overfitting occurs when the model performs well on training data but poorly on unseen data.

Solutions:

- Use regularization (L1/L2).
- Use simpler models or prune trees.
- Use ensemble methods like Random Forest or Gradient Boosting.
- Apply cross-validation.
- Train on more data.
- Early stopping or dropout (in neural networks).