Machine Learning: Supervised Learning

Vikas Thammanna Gowda

02/13/2025

1 Supervised Learning

Supervised Machine Learning is a type of machine learning where the algorithm is trained on a labeled dataset, meaning each training example is paired with the correct output. The goal of the algorithm is to learn the mapping between inputs (features) and outputs (labels), so it can predict the output for unseen data.

1.1 Key Components of Supervised Learning

- **Labeled Dataset:** The dataset used for training includes both input features (independent variables) and corresponding output labels (dependent variables).
  - *Example:* A dataset of housing prices with features such as square footage, number of bedrooms, and location, and the output label as the price of the house.
- **Training and Testing:** The dataset is typically split into:
  - **Training Set:** Used to train the model.
  - **Testing Set:** Used to evaluate the model's performance.
- **Objective:** To minimize the error between the predicted and actual labels by adjusting the model's parameters.

1.2 Types of Supervised Learning

- **Regression:** Predicts continuous values.
  - *Example:* Predicting house prices based on features like area, number of rooms, and location.
- **Classification:** Predicts discrete labels.
  - *Example:* Determining whether an email is "spam" or "not spam."

1.3 Basic Working of Supervised Learning

1. Data Collection

Gather a labeled dataset containing input features (independent variables) and their corresponding output labels (dependent variables).

- *Example:* A dataset containing images of fruits with labels specifying their type (e.g., "apple," "banana," etc.).

2. Data Preprocessing

Prepare the data to ensure it is clean and suitable for training. This is a critical step and involves multiple sub-tasks:

- **Feature Selection:** Choose the most relevant features from the dataset that are strongly correlated with the output label. Irrelevant or redundant features can be removed to improve performance.
  - *Example:* For predicting house prices, features like "square footage" and "location" may be selected, while "owner name" is removed.
- **Handling Missing Values:** Address incomplete data by:
  - Filling missing values using techniques like mean, median, or mode.
  - Removing rows or columns with excessive missing data.
- **Outlier Treatment:** Detect and handle extreme values that can skew the results using techniques such as:
  - Removing outliers based on statistical thresholds (e.g., Z-score, IQR).
  - Applying robust scaling methods.
- **Feature Scaling:** Standardize or normalize the features to bring them to a similar scale, especially for algorithms sensitive to scale (e.g., k-NN, SVM). Techniques include:
  - Min-Max Scaling (range $[0, 1]$).
  - Standardization (mean $= 0$, standard deviation $= 1$).
- **Feature Extraction (Optional):** Create new features or reduce dimensionality using techniques such as:
  - Principal Component Analysis (PCA).
  - Text embeddings for Natural Language Processing (NLP).

3. Model Selection

Choose an appropriate algorithm for the problem:

- **Regression Algorithms:** For continuous outputs (e.g., predicting house prices).
- **Classification Algorithms:** For discrete outputs (e.g., spam vs. not spam).

4. Training

Train the chosen model on the preprocessed training data. The model learns patterns by:

- Calculating the loss function, which measures the difference between predicted and actual outputs.
- Optimizing parameters to minimize the loss using algorithms like Gradient Descent.

5. Testing

Evaluate the trained model on the testing set to measure its performance using appropriate metrics:

- **Regression Metrics:** Mean Absolute Error (MAE), Mean Squared Error (MSE), or R^2.
- **Classification Metrics:** Accuracy, Precision, Recall, F1-score, or ROC-AUC.

6. Prediction

Use the trained model to predict outcomes for new, unseen data.

2 Regression Analysis

Regression analysis is a statistical technique for modeling the relationship between a dependent variable (the target or outcome) and one or more independent variables (predictors or features). It helps to understand how changes in the independent variables influence the dependent variable and is widely used for prediction and forecasting.

2.1 Purpose of Regression Analysis

- **Understand Relationships:** Explore how variables are related, e.g., does advertising budget affect sales?
- **Prediction:** Make predictions based on the model, e.g., predict the salary of an employee based on experience.
- **Feature Importance:** Determine the significance of variables in influencing the outcome.
- **Optimization:** Use the model for decision-making in business, economics, healthcare, etc.

2.2 Applications

- Predicting sales, stock prices, or customer demand.
- Estimating the impact of education on income levels.
- Forecasting weather patterns.
- Modeling the effect of drug dosage on patient outcomes.

2.3 Simple Linear Regression

Simple Linear Regression models the relationship between one dependent variable and one independent variable using a straight line. This is the simplest form of regression.

2.3.1 Mathematical Form

$$y = \beta_0 + \beta_1 x + \epsilon$$

Where:

- y: Dependent variable (e.g., house price).
- x: Independent variable (e.g., square footage).
- β_0: Intercept, representing the baseline value of y when $x = 0$.
- β_1: Slope, indicating how much y changes for a unit change in x.
- ϵ: Error term accounting for variability not explained by the model.

2.3.2 Example

A simple model for predicting a house price based on square footage:

$$\text{Price} = 10,000 + (\text{Square Footage} \times 126.6) + \epsilon$$

- $10,000$: Represents a base price that you pay regardless of the house size (e.g., land value, minimum building cost).
- 126.6: Indicates that for every additional square foot, the house price increases by \$126.6.
- ϵ: Captures random factors affecting price, such as market conditions or location.

How This Model Works:

- During training, the model uses data to compute β_0 (intercept), β_1 (slope), and estimates ϵ (error).
- For a house with 1,000 square feet:

$$\text{Price} = 10,000 + (1000 \times 126.6) = 136,600 \quad \text{(ignoring error for simplicity)}.$$

2.4 Multiple Linear Regression

Multiple Linear Regression extends simple linear regression to include two or more independent variables. It can capture the combined effect of multiple predictors on the dependent variable.

2.4.1 Mathematical Form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

Where:

- y: Dependent variable (e.g., house price).
- $x_1, x_2, \ldots, x_n$: Independent variables (e.g., square footage, number of bedrooms, location rating).
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$: Coefficients determined during training.
- ϵ: Error term capturing variability not explained by the predictors.

2.4.2 Example

A model for predicting house prices based on square footage, number of bedrooms, and location score:

$$\text{Price} = 15,000 + (\text{Square Footage} \times 120) + (\text{Bedrooms} \times 8,000) + (\text{Location Score} \times 25,000) + \epsilon$$

- $15,000$: Base price, considering factors like land cost or minimum value of the property.
- 120: Indicates how much the price increases for each additional square foot.
- $8,000$: Represents the price increment for each additional bedroom.
- $25,000$: Captures the value added based on location score (e.g., proximity to schools or amenities).
- ϵ: Accounts for unmeasured factors such as market fluctuations or neighborhood conditions.

How This Model Works:

- If a house has 1,200 square feet, 3 bedrooms, and a location score of 4, the predicted price is:

$$\text{Price} = 15,000 + (1200 \times 120) + (3 \times 8,000) + (4 \times 25,000)$$

$$\text{Price} = 15,000 + 144,000 + 24,000 + 100,000 = 283,000 \quad \text{(ignoring error for simplicity)}.$$

2.5 Non-Linear Regression

Non-linear regression models relationships that cannot be represented with a straight line. Instead, it uses more complex equations, such as polynomials, exponential functions, or logarithmic relationships.

2.5.1 Mathematical Form

A non-linear regression equation could take many forms, such as:

- Polynomial:
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n + \epsilon$$
- Exponential:
$$y = \beta_0 e^{\beta_1 x} + \epsilon$$
- Logarithmic:
$$y = \beta_0 + \beta_1 \ln(x) + \epsilon$$

2.5.2 Example

A model for predicting sales based on advertising budget where the returns diminish after a certain point:

$$\text{Sales} = 1,000 + 500 \times \ln(\text{Advertising Budget}) + \epsilon$$

- $1,000$: Base sales without any advertising.
- $500 \times \ln(\text{Budget})$: Reflects diminishing returns—spending more on advertising increases sales, but at a slower rate over time.

Another example: Population growth (exponential model):

$$\text{Population} = 10,000 \times e^{0.03 \times \text{Years}} + \epsilon$$

- $10,000$: Initial population.
- $e^{0.03 \times \text{Years}}$: Captures exponential growth, where the growth rate is 3% per year.

How This Model Works:

- Non-linear regression is used when the relationship between variables shows patterns like diminishing returns, exponential growth, or seasonal variations.
- For example, if the advertising budget is \$1,000:

$$\text{Sales} = 1,000 + 500 \times \ln(1000) = 1,000 + 500 \times 6.91 = 4,455 \quad \text{(ignoring error for simplicity)}.$$

3 Evaluation Metrics

3.1 R-Squared (Coefficient of Determination)

R-squared (R^2) is a statistical measure used to evaluate the goodness-of-fit of a regression model. It tells us how well the independent variables (predictors) explain the variance in the dependent variable (target).

Variance:

- Variance measures how much the dependent variable (y) varies from its mean ($\bar{y}$).
- If a regression model explains most of this variance, it is considered a good model.

R-squared:

- R^2 represents the proportion of the variance in the dependent variable that is explained by the independent variables.
- Ranges between 0 and 1:
  - 0: The model explains none of the variance in y (poor model).
  - 1: The model explains all of the variance in y (perfect model).

3.1.1 Mathematical Formula

The formula for R^2 is:

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}}$$

Where:

- SS_{res} (Residual Sum of Squares):

$$\text{SS}_{\text{res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

  Captures the error or unexplained variance in the model.

- SS_{tot} (Total Sum of Squares):

$$\text{SS}_{\text{tot}} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

  Measures the total variance in the dependent variable.

- SS_{reg} (Regression Sum of Squares):

$$\text{SS}_{\text{reg}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Relationship:

$$\text{SS}_{\text{tot}} = \text{SS}_{\text{res}} + \text{SS}_{\text{reg}}$$

3.1.2 Interpretation of R^2

- $R^2 = 0$: The model explains none of the variance in the data. Predictions are no better than the mean of the dependent variable.
- $R^2 = 1$: The model perfectly explains the variance in the data. All data points lie exactly on the regression line.
- $0 < R^2 < 1$: The model explains some of the variance but not all.

Example: If $R^2 = 0.75$, it means 75% of the variance in the dependent variable is explained by the independent variables, and 25% is unexplained.

3.1.3 Example Use Case

Suppose we are building a regression model to predict house prices based on features like square footage and the number of bedrooms.
Model 1 (Using only square footage):

$$R^2 = 0.70$$

This means 70% of the variance in house prices is explained by square footage.
Model 2 (Using square footage and number of bedrooms):

$$R^2 = 0.85$$

This means 85% of the variance in house prices is explained by both square footage and the number of bedrooms.
Conclusion: Model 2 is better than Model 1 in terms of explaining the variance in house prices.

3.2 Mean Squared Error (MSE)

MSE calculates the average squared difference between actual values (y_i) and predicted values ($\hat{y}_i$). It penalizes larger errors more because the errors are squared.

3.2.1 Formula

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where:

- y_i is the actual value.
- $\hat{y}_i$ is the predicted value.
- n is the number of data points.

3.2.2 Interpretation

- The closer MSE is to zero, the better the model.
- Since errors are squared, large deviations have a greater impact on MSE.
- The unit of MSE is not the same as the dependent variable due to squaring.

3.2.3 Example Calculation

Given: Actual house prices: $[200, 250, 300]$ Model predictions: $[210, 245, 290]$
Squared errors:

$$[(200-210)^2, (250-245)^2, (300-290)^2] = [100, 25, 100]$$

$$MSE = \frac{100 + 25 + 100}{3} = \frac{225}{3} = 75$$

3.2.4 Advantages and Disadvantages

Advantages:

- Penalizes large errors more, making it useful when large deviations are costly.
- Differentiable, making it useful in optimization algorithms like gradient descent.

Disadvantages:

- The error is squared, making it difficult to interpret in terms of the original dependent variable.
- Very sensitive to outliers, which can distort the results.

3.3 Root Mean Squared Error (RMSE)

RMSE is the square root of MSE. It restores the error metric to the same unit as the target variable, making it easier to interpret.

3.3.1 Formula

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Where:

- y_i is the actual value.
- $\hat{y}_i$ is the predicted value.
- n is the number of data points.

3.3.2 Interpretation

- RMSE represents the average prediction error in the same unit as y.
- A lower RMSE means a better model.

3.3.3 Example Calculation

Given: Actual house prices: $[200, 250, 300]$ Model predictions: $[210, 245, 290]$
Squared errors:

$$[(200-210)^2, (250-245)^2, (300-290)^2] = [100, 25, 100]$$

$$RMSE = \sqrt{\frac{100+25+100}{3}} = \sqrt{\frac{225}{3}} = \sqrt{75} = 8.66$$

This means, on average, the model's prediction error is 8.66 units.

3.3.4 Advantages and Disadvantages

Advantages:

- Easier to interpret than MSE because the error is in the same unit as y.
- Still penalizes large errors, but in a way that maintains interpretability.

Disadvantages:

- Sensitive to outliers, just like MSE.
- More computationally expensive due to the square root operation.

3.4 Mean Absolute Error (MAE)

MAE measures the average absolute difference between actual and predicted values. Unlike MSE, it does not square the errors, making it less sensitive to outliers.

3.4.1 Formula

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Where:

- y_i is the actual value.
- $\hat{y}_i$ is the predicted value.
- n is the number of data points.

3.4.2 Interpretation

- MAE represents the average absolute prediction error in the same unit as y.
- Lower MAE means better accuracy.

3.4.3 Example Calculation

Given: Actual house prices: $[200, 250, 300]$ Model predictions: $[210, 245, 290]$ Absolute errors:

$$[|200 - 210|, |250 - 245|, |300 - 290|] = [10, 5, 10]$$

$$MAE = \frac{10 + 5 + 10}{3} = \frac{25}{3} = 8.33$$

3.4.4 Advantages and Disadvantages

Advantages:

- More robust to outliers because it does not square the errors.
- The error is in the same unit as the dependent variable, making it easy to interpret.

Disadvantages:

- Does not penalize large errors as much as MSE or RMSE.
- Not differentiable at zero, which can be problematic for certain optimization algorithms.

3.4.5 When to Use Each Metric

- **MSE:** Use when you want to penalize large errors more (e.g., in financial modeling).
- **RMSE:** Use when you need an interpretable metric that still considers large errors.
- **MAE:** Use when you want a robust metric that is not overly influenced by outliers.

4 Regularization in Machine Learning

Regularization is a technique used in machine learning to prevent overfitting by adding a penalty term to the model's loss function. Overfitting occurs when a model learns the training data too well, including noise, leading to poor generalization to new data. Regularization restricts the complexity of the model by controlling the magnitude of the model parameters (coefficients).

Why Regularization?

In linear regression, we fit a model of the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

Without regularization, the model minimizes the Sum of Squared Errors (SSE):

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

However, if there are too many features (high-dimensional data), the model may overfit, meaning it captures noise rather than meaningful patterns. Regularization adds a penalty term to the cost function, discouraging large coefficients and leading to a simpler, more generalizable model.

4.1 Ridge Regression (L2 Regularization)

Ridge Regression adds the sum of squared coefficients as a penalty term to the cost function. This prevents coefficients from becoming too large but does not shrink them to zero.

4.1.1 Mathematical Formula

$$\text{Loss} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{m} \beta_j^2$$

Where:

- The first term is the ordinary least squares (OLS) loss function.
- λ (regularization parameter) controls the strength of the penalty.
- The second term $\sum \beta_j^2$ shrinks coefficients but does not eliminate them.

4.1.2 Effect of Ridge

- Small $\lambda \rightarrow$ The model behaves like ordinary linear regression.
- Large $\lambda \rightarrow$ The model forces coefficients to be smaller.

Use Case: Used when all features are relevant, but some should be regularized to prevent overfitting.
Example: Predicting house prices where square footage, number of bedrooms, and location all contribute, but some features may have excessive influence.

4.2 Lasso Regression (L1 Regularization)

Lasso (Least Absolute Shrinkage and Selection Operator) adds the sum of absolute values of coefficients as a penalty. Unlike Ridge, Lasso can shrink some coefficients to exactly zero, effectively performing feature selection.

4.2.1 Mathematical Formula

$$\text{Loss} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{m} |\beta_j|$$

Where:

- The second term $\sum |\beta_j|$ shrinks some coefficients to exactly zero, removing unimportant features.

4.2.2 Effect of Lasso

- Small $\lambda \rightarrow$ The model behaves like ordinary linear regression.
- Large $\lambda \rightarrow$ Some coefficients are forced to zero, effectively removing irrelevant features.

Use Case: Used when some features are irrelevant and should be removed automatically.
Example: Predicting stock prices, where only a few economic indicators actually impact the outcome.

4.3 Elastic Net (Combination of L1 and L2 Regularization)

Elastic Net combines Ridge (L2) and Lasso (L1) regularization to create a balanced model. It helps when:

- Lasso may remove too many features.
- Ridge may keep too many irrelevant features.

4.3.1 Mathematical Formula

$$\text{Loss} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{m} |\beta_j| + \lambda_2 \sum_{j=1}^{m} \beta_j^2$$

Where:

- λ_1 controls the L1 (Lasso) penalty.
- λ_2 controls the L2 (Ridge) penalty.

4.3.2 Effect of Elastic Net

- Reduces large coefficients (like Ridge) but eliminates irrelevant features (like Lasso).
- Helps when data has many correlated features (which Lasso struggles with).

Use Case: Used when there are many features, some of which should be removed, but we don't want to remove too many.
Example: Medical diagnosis models, where some features may be strongly correlated and should be regularized instead of removed completely.

4.4 Choosing the Right Regularization Model

Use Ridge when:

- All features are relevant, but some should have reduced impact.
- There is multicollinearity (correlation between features).
- **Example:** Predicting sales with multiple factors like seasonality, promotions, and pricing.

Use Lasso when:

- You suspect that some features are completely irrelevant and should be removed.
- **Example:** Genetic studies, where only a few genes affect disease outcomes.

Use Elastic Net when:

- You have many correlated features, and Lasso alone might remove too many.
- Ridge alone is keeping too many irrelevant features.
- **Example:** Financial modeling, where multiple economic indicators may be correlated.

4.5 Final Takeaways

- Regularization prevents overfitting by penalizing large coefficients.
- Ridge (L2) shrinks coefficients but does not remove them.
- Lasso (L1) shrinks and eliminates features, making it useful for feature selection.
- Elastic Net balances both Ridge and Lasso, ideal for high-dimensional and correlated data.

5 Bias-Variance Trade-off

The bias-variance trade-off is a key concept in machine learning that describes the balance between underfitting and overfitting. It explains how different types of errors affect a model's ability to generalize to new data. In supervised learning, we aim to build a model that generalizes well to unseen data. However, two types of errors can prevent this:

1. Bias (Systematic Error)

- Bias measures how much a model's predictions differ from the true values on average.
- A high-bias model makes strong assumptions about the data and may oversimplify the relationship between features and the target variable.
- Such models fail to capture complex patterns, leading to underfitting.

Example of High Bias (Underfitting):

- A linear regression model trying to fit a dataset with a non-linear relationship.
- The model makes simple assumptions and fails to capture the actual pattern.

2. Variance (Model Sensitivity)

- Variance measures how much a model's predictions change when trained on different subsets of the data.
- A high-variance model is too flexible, learning noise and random fluctuations rather than the true relationship.
- This results in overfitting, where the model performs well on training data but poorly on new data.

Example of High Variance (Overfitting):

- A complex deep learning model applied to a small dataset.
- The model memorizes the training data, including noise, leading to poor generalization.

5.1 The Trade-off

- If a model has high bias, it is too simple and cannot capture patterns in the data (underfitting).
- If a model has high variance, it is too complex and captures noise as if it were a real pattern (overfitting).
- A good model balances bias and variance to achieve low total error.

5.2 Total Error Decomposition

The total error (expected test error) can be broken down as:

$$\text{Total Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

Where:

- Bias^2: Error due to incorrect model assumptions.
- Variance: Error due to excessive sensitivity to training data.
- Irreducible Error: The error inherent in the data (e.g., noise in real-world measurements).

Goal: Minimize both bias and variance to achieve the lowest possible error.

5.3 Examples

High Bias Models (Underfitting)

- Linear Regression on a non-linear dataset.
- Decision Stumps (shallow decision trees).
- Naïve Bayes (makes strong independence assumptions).

High Variance Models (Overfitting)

- Deep Neural Networks with insufficient data.
- Decision Trees with too many splits.
- k-Nearest Neighbors (k-NN) with very small k.

Balanced Models

- Ridge Regression (regularized linear regression).
- Pruned Decision Trees (restrict excessive branching).
- k-NN with a moderate k value.

5.4 Techniques to Handle Bias-Variance Trade-off

Reducing Bias (Addressing Underfitting)

- Use a more complex model (e.g., switch from linear regression to polynomial regression).
- Add more relevant features to improve prediction.
- Reduce regularization (e.g., decrease alpha in Ridge/Lasso).

Reducing Variance (Addressing Overfitting)

- Collect more training data to improve generalization.
- Apply regularization (e.g., Ridge, Lasso, Elastic Net).
- Use techniques like Dropout (for deep learning) to prevent excessive complexity.
- Perform feature selection to remove unnecessary predictors.

6 Exercise

1. What is the primary objective of supervised learning?
   (a) To find hidden patterns in unlabeled data
   (b) To learn the mapping between inputs and outputs
   (c) To cluster data into different groups
   (d) To generate synthetic data
2. Which of the following is an example of a classification problem?
   (a) Predicting house prices
   (b) Forecasting stock prices
   (c) Determining if an email is spam or not
   (d) Estimating the effect of temperature on ice cream sales
3. What is a labeled dataset?
   (a) A dataset with only input features
   (b) A dataset where each input is paired with a correct output
   (c) A dataset that has no missing values
   (d) A dataset that requires clustering
4. In supervised learning, the dataset is typically split into what two subsets?
   (a) Training and testing sets
   (b) Input and output sets
   (c) Regression and classification sets
   (d) Feature and label sets
5. Which of the following is not a supervised learning algorithm?
   (a) Decision Tree
   (b) Linear Regression
   (c) Naïve Bayes
   (d) k-Means Clustering
6. What is the primary goal of data preprocessing?
   (a) Increase the size of the dataset
   (b) Improve the accuracy of the model
   (c) Ensure data is in text format
   (d) Reduce computational cost
7. Which technique is used for handling missing values?
   (a) Removing all missing data
   (b) Filling with mean, median, or mode
   (c) Replacing missing values with zeros
   (d) Ignoring missing values

8. Why is feature scaling important?

   (a) It ensures all features have equal weight in distance-based algorithms
   (b) It increases the dimensionality of the dataset
   (c) It eliminates the need for a testing dataset
   (d) It converts categorical features to numerical ones

9. What is the primary difference between simple and multiple linear regression?

   (a) Simple regression uses one independent variable, multiple regression uses many
   (b) Multiple regression is only used for categorical data
   (c) Simple regression does not require labeled data
   (d) Multiple regression cannot be used in real-world scenarios

10. What is the purpose of a loss function in supervised learning?

    (a) To determine the model's complexity
    (b) To calculate the difference between predicted and actual values
    (c) To increase computational efficiency
    (d) To eliminate bias in the dataset

11. Which of the following is NOT an evaluation metric for regression models?

    (a) Mean Squared Error (MSE)
    (b) Root Mean Squared Error (RMSE)
    (c) Precision
    (d) R-Squared (R^2)

12. What does R^2 measure in regression models?

    (a) The proportion of variance explained by the model
    (b) The accuracy of the model
    (c) The complexity of the model
    (d) The number of features in the dataset

13. Which regression type can model relationships that look like a parabola when plotted?

    (a) Simple Linear Regression
    (b) Multiple Linear Regression
    (c) Non-Linear Regression
    (d) Logistic Regression

14. What is the purpose of regularization in machine learning?

    (a) To make the model more complex
    (b) To reduce overfitting by adding a penalty term
    (c) To increase the variance of the model
    (d) To remove the need for training data

15. What type of regularization does Ridge Regression use?

    (a) L1 Regularization
    (b) L2 Regularization
    (c) Both L1 and L2 Regularization
    (d) No regularization

16. Which regularization method can shrink some feature coefficients to exactly zero?

    (a) Ridge Regression
    (b) Lasso Regression
    (c) Elastic Net
    (d) Linear Regression

17. What is the main advantage of Elastic Net over Ridge and Lasso regression?

    (a) It applies both L1 and L2 regularization
    (b) It eliminates all features
    (c) It prevents feature selection
    (d) It is only used for classification problems

18. What does high bias in a model indicate?

    (a) The model is overfitting
    (b) The model is underfitting
    (c) The model has too many features
    (d) The model performs well on unseen data

19. What does high variance in a model indicate?

    (a) The model is too simple
    (b) The model is underfitting
    (c) The model is too complex and overfitting
    (d) The model generalizes well

20. Which of the following is an example of underfitting?

    (a) A deep neural network with high accuracy on training data but low on test data
    (b) A decision tree with too many branches
    (c) A linear regression model applied to a non-linear dataset
    (d) A k-NN model with very small k

21. Which of the following helps reduce variance in a model?

    (a) Increasing model complexity
    (b) Adding more training data
    (c) Using a simpler model
    (d) Removing regularization

22. What is the total error formula in the bias-variance trade-off?

    (a) Bias + Variance
    (b) Bias^2 + Variance + Irreducible Error
    (c) Variance - Bias^2
    (d) Bias^2 + Variance - Error

23. What is the main purpose of feature selection?

    (a) To increase the number of features in a dataset
    (b) To improve model performance by using the most relevant features
    (c) To ensure the dataset has no missing values
    (d) To reduce training time without affecting accuracy

24. Which of the following is a classification evaluation metric?

    (a) Mean Absolute Error (MAE)
    (b) Root Mean Squared Error (RMSE)
    (c) Precision
    (d) R-Squared (R^2)

25. Which of the following techniques is used to prevent overfitting?

    (a) Increasing the dataset size
    (b) Regularization
    (c) Reducing the number of features
    (d) All of the above

26. Which of the following models is prone to high variance?

    (a) Linear Regression
    (b) Decision Stump
    (c) Deep Neural Network trained on a small dataset
    (d) Ridge Regression

27. What is the key characteristic of a well-generalized model?

    (a) High training accuracy but low test accuracy
    (b) Low bias and high variance
    (c) A balance between bias and variance
    (d) High complexity with a large number of parameters

28. Which of the following techniques can be used to handle outliers?

    (a) Removing outliers using IQR
    (b) Normalizing features
    (c) Filling missing values
    (d) Reducing the dataset size

29. Why is regularization important in high-dimensional datasets?

    (a) To reduce computational cost
    (b) To improve training speed
    (c) To prevent overfitting by reducing feature importance
    (d) To eliminate missing values

30. What is the key difference between Ridge and Lasso regression?

    (a) Ridge reduces all coefficients but does not eliminate them, while Lasso can shrink some coefficients to zero
    (b) Ridge eliminates features, while Lasso only reduces their impact
    (c) Lasso increases bias, while Ridge increases variance
    (d) Ridge and Lasso work only for linear regression