

ML UNIT – 3 (Supervised Learning: Regression) – END-SEM PYQ Answers➤ **MAY / JUN 2023****Q1) a) Explain the following terms with suitable examples. [6]****i) Bias:**

- **Definition:** Bias refers to the error introduced by approximating a real-world problem, which may be complex, by a simplified model. High bias means the model makes strong assumptions about the data, often leading to **underfitting**.
- **Example:** If we use a **linear model** to predict house prices based only on square footage, ignoring other features like location or number of rooms, the model may consistently make errors because it is too simple.

ii) Variance:

- **Definition:** Variance refers to the model's sensitivity to small fluctuations in the training data. High variance means the model captures noise along with patterns, often leading to **overfitting**.
- **Example:** Using a **high-degree polynomial regression** to predict stock prices may fit the training data very closely but fail to predict new data accurately.

iii) Underfitting and Overfitting:

- **Underfitting:** Occurs when a model is too simple to capture the underlying pattern in the data. It performs poorly on both training and test data.
 - *Example:* Using a **linear model** for a clearly nonlinear relationship.
- **Overfitting:** Occurs when a model is too complex and captures noise along with the actual pattern. It performs well on training data but poorly on unseen test data.
 - *Example:* Using a **10th-degree polynomial regression** on a small dataset; it fits training data perfectly but fails on test data.

Extra Note:

- The **goal** is to find a balance between bias and variance to achieve **good generalization** on unseen data.

b) Differentiate between Lasso Regression and Ridge Regression . [6]

| Characteristic | Ridge Regression | Lasso Regression |
|----------------------------|--|--|
| Regularization Type | Applies L2 regularization , adding a penalty term proportional to the square of the coefficients | Applies L1 regularization , adding a penalty term proportional to the absolute value of the coefficients . |

| Characteristic | Ridge Regression | Lasso Regression |
|-----------------------------|---|---|
| Feature Selection | Does not perform feature selection . All predictors are retained, although their coefficients are reduced in size to minimize overfitting | Performs automatic feature selection . Less important predictors are completely excluded by setting their coefficients to zero. |
| When to use | Best suited for situations where all predictors are potentially relevant , and the goal is to reduce overfitting rather than eliminate features | Ideal when you suspect that only a subset of predictors is important, and the model should focus on those while ignoring the irrelevant ones. |
| Output model | Produces a model that includes all features , but their coefficients are smaller in magnitude to prevent overfitting | Produces a model that is simpler , retaining only the most significant features and ignoring the rest by setting their coefficients to zero. |
| Impact on Prediction | Reduces the magnitude of coefficients, shrinking them towards zero, but does not set any coefficients exactly to zero. All predictors remain in the model | Shrinks some coefficients to exactly zero , effectively removing their influence from the model. This leads to a simpler model with fewer features |
| Computation | Generally faster as it doesn't involve feature selection | May be slower due to the feature selection process |
| Example Use Case | Use when you have many predictors, all contributing to the outcome (e.g., predicting house prices where all features like size, location, etc., matter) | Use when you believe only some predictors are truly important (e.g., genetic studies where only a few genes out of thousands are relevant). |

c) Explain gradient descent algorithm with example. [6]

i) Gradient Descent Algorithm: Gradient Descent is an optimization algorithm used to **minimize a cost (loss) function** by iteratively moving in the direction of its **negative gradient**.

It is widely used in **machine learning**, especially in training models like **linear regression**, **logistic regression**, and **neural networks**.

The main idea:

- Compute the slope (gradient) of the cost function
- Move opposite to the slope (downhill)
- Update parameters until the minimum is reached.

ii) Algorithm Steps

1. **Initialize parameters** (e.g., weights) with random values.

2. **Choose a learning rate (α)** — it decides the step size while moving towards the minimum.
3. **Compute the gradient** of the cost function with respect to each parameter.
4. **Update the parameters** using the rule: $\theta := \theta - \alpha \cdot \frac{\partial \theta}{\partial J}$
5. **Repeat steps 3–4** until the algorithm converges (i.e., cost function stops changing significantly).

iii) Example: Gradient Descent for Linear Regression

1. Consider a simple linear regression model: $y = mx + c$
2. Cost function: $J(m, c) = \frac{1}{2n} \sum_{i=1}^n (y_i - (mx_i + c))^2$
3. Gradients:

$$\frac{\partial J}{\partial m} = -\frac{1}{n} \sum_i x_i (y_i - (mx_i + c)) \quad \frac{\partial J}{\partial c} = -\frac{1}{n} \sum_i (y_i - (mx_i + c))$$

Parameter Update

Update slope m : $m := m - \alpha \cdot \frac{\partial J}{\partial m}$

Update intercept c : $c := c - \alpha \cdot \frac{\partial J}{\partial c}$

Intuitive Example (Simple Numeric):

Suppose the initial guess for slope is $m = 0$, and the true relation is $y = 2x$.

- Compute the gradient \rightarrow positive value
- Move in the negative direction $\rightarrow m$ increases toward 2
- On each iteration, m becomes closer to 2
- Finally, cost is minimized and model converges.

Advantages

- Simple to understand and implement.
- Works well for large datasets and high-dimensional problems.

Conclusion: Gradient descent finds the optimal model parameters by **iteratively reducing the cost function** using gradient information. It is a **core algorithm in machine learning and optimization**.

Q2) a) What do you mean by regression? Explain with suitable example. [6]

Regression is a statistical and machine-learning technique used to **find the relationship between a dependent variable (output) and one or more independent variables (inputs)**. It is mainly used to **predict continuous numerical values** such as price, temperature, sales, or system load.

Key Points:

1. Regression helps in understanding **how the output changes** when input factors change.
2. It fits a **mathematical equation (line/curve)** to represent the pattern in the given data.
3. Used widely for **forecasting, trend analysis, and prediction** in various domains.
4. It **minimizes prediction error** by reducing the difference between actual and predicted values. *(extra explanatory line)*
5. Regression also helps in **performance analysis and anomaly detection**, especially in cybersecurity. *(extra explanatory line)*
6. It improves decision-making by providing **data-driven insights** about future outcomes. *(extra explanatory line)*

Example: Suppose a cybersecurity analyst wants to **predict the number of cyber-attacks per day** based on factors like:

- Internet traffic volume
- Number of active users
- System vulnerabilities

Using **Linear Regression**, the model may learn a relationship such as:

$$\text{Predicted Attacks} = 2.5 \times (\text{Traffic Volume}) + 1.8 \times (\text{Active Users}) + 5$$

If tomorrow's traffic and active users are known, the analyst can **predict the expected attack count**, helping in proactive defense planning. This example shows how regression turns real-world data into actionable insights, supporting early detection and preparation. It also demonstrates how continuous patterns can be modeled to improve security operations.

b) Write a short note on: [6]**i) MAE****ii) RMSE****iii) R²****i) MAE (Mean Absolute Error)**

Definition: MAE is a regression evaluation metric that measures the **average absolute difference** between actual values and predicted values.

Points:

1. It shows **how much error** the model makes on average.
2. It treats all errors **equally**, without giving more weight to bigger errors.
3. Formula: $MAE = \frac{1}{n} \sum |y_{\text{actual}} - y_{\text{predicted}}|$
4. Lower MAE means the model's predictions are **closer to the real values**.
5. It is simple to understand and commonly used in **forecasting and anomaly detection**.
6. MAE is useful when we want a **stable and balanced measure** of prediction error.

ii) RMSE (Root Mean Squared Error)

Definition: RMSE is a regression metric that calculates the **square root of the average of squared errors**, giving **more weight to larger errors**.

Points:

1. It highlights **large deviations** more strongly than smaller ones.
2. Formula: $RMSE = \sqrt{\frac{1}{n} \sum (y_{\text{actual}} - y_{\text{predicted}})^2}$
3. Lower RMSE means the model is performing accurately.
4. RMSE is useful when **large errors are critical** and must be minimized.
5. It is widely used in real-time systems like **cyberattack prediction and load forecasting**.
6. It reflects overall model performance by considering **both variance and bias**.

iii) R² (Coefficient of Determination)

Definition: R² is a metric that explains **how much of the variation in the output** is explained by the regression model. It represents the **goodness of fit**.

Points:

1. R² value ranges from **0 to 1**, where 1 means a perfect fit.
2. A higher R² means the model explains more of the data's variation.
3. Formula: $R^2 = 1 - \frac{\text{Sum of Squared Errors}}{\text{Total Sum of Squares}}$
4. R² helps in comparing **multiple regression models** for selection.
5. It is commonly used to measure how well the model fits **historical cyber data**.
6. A low R² means the model fails to capture important patterns in the data.

c) What is Gradient Descent? Compare Batch Gradient Descent and Stochastic Gradient Descent.**Gradient Descent (Definition)**

- Gradient Descent is an optimization algorithm used to **minimize a loss function** by repeatedly moving in the direction of the **negative gradient**.
- It helps machine-learning models adjust their parameters (weights) so the prediction error becomes as small as possible.
- The algorithm updates parameters step-by-step until it reaches the point where the error is minimum.

Key Points:

1. It calculates the gradient (slope) of the loss function with respect to model parameters.
2. Parameters are updated using a learning rate which controls the step size.

3. It is widely used in training regression models, neural networks, and deep learning.
4. Gradient Descent ensures continuous improvement in model accuracy during training.
5. It is an iterative method and stops when further updates give no significant change.
6. Helps find optimal parameters that give the lowest prediction error.

| Aspect | Batch Gradient Descent | Stochastic Gradient Descent (SGD) |
|--|--|---|
| Data Processing | Uses the whole training dataset to compute the gradient. | Uses a single training sample to compute the gradient. |
| Convergence Speed | Slower, takes longer to converge. | Faster, converges quicker due to frequent updates. |
| Convergence Accuracy | More accurate, gives precise gradient estimates. | Less accurate due to noisy gradient estimates. |
| Computational and Memory Requirements | Requires significant computation and memory. | Requires less computation and memory. |
| Optimization of Non-Convex Functions | Can get stuck in local minima. | Can escape local minima and find the global minimum. |
| Suitability for Large Datasets | Not ideal for very large datasets due to slow computation. | Can handle large datasets effectively. |
| Nature | Deterministic: Same result for the same initial conditions. | Stochastic: Results can vary with different initial conditions. |
| Learning Rate | Fixed learning rate. | Learning rate can be adjusted dynamically. |
| Shuffling of Data | No need for shuffling. | Requires shuffling of data before each epoch. |
| Overfitting | Can overfit if the model is too complex. | Can reduce overfitting due to more frequent updates. |
| Escape Local Minima | Cannot escape shallow local minima. | Can escape shallow local minima more easily. |
| Computational Cost | High due to processing the entire dataset at once. | Low due to processing one sample at a time. |
| Final Solution | Tends to converge to the global minimum for convex loss functions. | May converge to a local minimum or saddle point. |

➤ **MAY / JUN 2024**

Q1) a) Define different regression models. [6]

- Regression models are statistical and machine-learning techniques used to **predict a continuous output** by finding relationships between input variables (independent variables) and the target value (dependent variable).
- Different regression models use different mathematical approaches to fit data and minimize prediction error.

Different Regression Models (Any 6 Models Explained):

1) Linear Regression

- Models the relationship using a **straight line**.
- Works when the relationship between input and output is linear.
- Simple and widely used for prediction problems.

2) Multiple Linear Regression

- Extension of linear regression with **more than one input variable**.
- Helps analyze the effect of multiple factors on the output.
- Useful in real-world forecasting and analysis.

3) Polynomial Regression

- Fits a **curved line** by adding polynomial terms (x^2 , x^3 ...).
- Used when data shows nonlinear patterns.
- More flexible than simple linear regression.

4) Logistic Regression

- Used to predict **binary outcomes** (Yes/No, 0/1).
- Applies a sigmoid function to convert outputs to probability.
- Common in classification problems like spam detection.

5) Ridge Regression

- A regularized regression model that **adds penalty** to reduce overfitting.
- Useful when data has multicollinearity.
- Helps maintain stable coefficient values.

6) Lasso Regression

- Similar to ridge but uses **L1 penalty**.
- Can shrink some coefficients to zero → feature selection.
- Useful when we want a simpler, more interpretable model.

b) What are different techniques to reduce under fitting? [6]

Underfitting occurs when a model is **too simple** to learn the underlying pattern of the data, resulting in **high training error and poor performance**.

Techniques to Reduce Underfitting (Any 6 Techniques):**1) Increase Model Complexity**

- Use more complex algorithms (e.g., decision trees, neural networks).
- Allows the model to capture deeper patterns.

2) Increase Training Time / Epochs

- Train the model for more iterations.
- Helps the model learn better representations.

3) Add More Features

- Introduce new relevant variables.
- Makes the model more informative and reduces simplicity.

4) Reduce Regularization

- Lower L1 or L2 penalty values.
- Prevents the model from being overly restricted.

5) Use Polynomial Features

- Convert linear models into nonlinear models.
- Helps capture curves and complex relationships.

6) Improve Data Quality

- Remove noise, fill missing values, correct errors.
- Cleaner data improves model learning efficiency.

c) Using regression model, predict expenditure of 6th month. [6]

| | | | | | |
|-----------------|----|----|----|----|----|
| Month (x) | 1 | 2 | 3 | 4 | 5 |
| Expenditure (y) | 12 | 19 | 29 | 37 | 45 |

Step 1: Calculate required values

Step 1: Calculate required values

$$n = 5$$

$$\sum x = 1 + 2 + 3 + 4 + 5 = 15$$

$$\sum y = 12 + 19 + 29 + 37 + 45 = 142$$

$$\sum xy = (1)(12) + (2)(19) + (3)(29) + (4)(37) + (5)(45) = 12 + 38 + 87 + 148 + 225 = 510$$

$$\sum x^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 = 55$$

Step 2: Find slope (b)

$$b = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

Substitute values:

$$b = \frac{5(510) - (15)(142)}{5(55) - 15^2}$$

$$b = \frac{2550 - 2130}{275 - 225}$$

$$b = \frac{420}{50} = 8.4$$

Step 3: Find intercept (a)

$$a = \frac{\sum y - b \sum x}{n}$$

$$a = \frac{142 - (8.4)(15)}{5}$$

$$a = \frac{142 - 126}{5} = \frac{16}{5} = 3.2$$

Step 4: Regression Equation

$$y = a + bx$$

$$y = 3.2 + 8.4x$$

Step 5: Predict expenditure for 6th month

$$y_6 = 3.2 + 8.4(6)$$

$$y_6 = 3.2 + 50.4 = 53.6$$

Predicted Expenditure for 6th Months = 53.6 units

Q2) a) What is R^2 measure of evaluation? [6]

R^2 (Coefficient of Determination) is an evaluation metric used in regression to measure how well the model explains the variation in the dependent variable. It represents the goodness of fit of a regression model.

Key Points:**1) Measures Explained Variance**

- R^2 tells what **percentage of total variation** in the output is explained by the input variables.

2) Value Ranges from 0 to 1

- $R^2 = 0 \rightarrow$ Model explains nothing.
- $R^2 = 1 \rightarrow$ Model perfectly explains the data.

3) Formula

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Where:

- SS_{res} = Sum of squared errors
- SS_{tot} = Total variation in the actual values

4) Higher R^2 Means Better Fit: A higher value shows that the model fits historical data more accurately.

5) Helps Compare Models: Used to compare multiple regression models and select the best performing one.

6) Indicates Pattern Learning Ability: Low R^2 means the model is not capturing important patterns, while high R^2 means it learns meaningful relationships.

b) What do you mean by least square method? Explain least square method in the context of linear regression. [6]

- The Least Square Method is a mathematical technique used to find the best-fit line for a given set of data points by minimizing the sum of the squared differences between actual values and predicted values.
- It ensures that the regression line reduces overall error as much as possible.

Explanation in Context of Linear Regression**1) Objective**

- The goal is to find a straight line: $y = a + bx$ **that best represents the data points.**

2) Error Calculation

- For each data point, error = actual value – predicted value.

- These errors are squared to avoid negative cancellation.

3) Minimizing Total Error

- Least squares minimizes the total squared error: $\sum (y - \hat{y})^2$
- A smaller value means a better-fitting line.

4) Finding Parameters (a and b)

- The least square formulas are used:

$$b = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

$$a = \frac{\sum y - b \sum x}{n}$$

5) Produces Optimal Best-Fit Line: The resulting line is mathematically proven to be the most accurate line among all possible lines.

6) Used in Prediction: Once the best-fit line is found, it is used to predict future values, especially in trend analysis and forecasting.

c) Write a short note on Stochastic Gradient Descent (SGD) algorithms. [6]

Stochastic Gradient Descent (SGD) is an optimization algorithm used to **minimize the loss function** in machine learning models by updating parameters using **one training sample at a time**. It is a faster and more efficient version of gradient descent, especially suitable for large datasets.

Key Points:

1) Works on Single Sample per Update: Instead of using the whole dataset, SGD picks **one random data point** to compute the gradient.

2) Faster Computation: Since only one sample is processed, parameter updates happen quickly, making SGD suitable for big data.

3) Noisy Updates: Frequent updates cause fluctuations, giving a **zig-zag path** toward the minimum, but still reaching a good solution.

4) Helps Escape Local Minima: The randomness in updates helps the algorithm jump out of **local minima**, improving training of complex models.

5) Used in Deep Learning: Most neural network training uses SGD because it handles high-dimensional data efficiently.

6) Supports Online / Real-Time Learning: New data can be fed continuously, making it ideal for systems that require **real-time model updates**.

➤ MAY / JUN 2025

Q1) a) Explain Linear Regression in brief. [6]

- Linear Regression is a statistical and machine-learning method used to **model the relationship** between a dependent variable (output) and one or more independent variables (inputs) using a **straight-line equation**.
- It is mainly used to **predict continuous values** based on past data.

Explanation (Point-wise):

1) Straight-Line Relationship: Linear Regression assumes the output changes **linearly** with respect to the input.

2) Regression Equation: The general equation is: $y = a + bx$

where

$a = \text{intercept}$, $b = \text{slope}$.

3) Slope Meaning: The slope (b) shows how much **y changes** when x increases by 1 unit.

4) Error Minimization: Linear Regression finds the best-fit line by **minimizing the prediction error** using the Least Squares Method.

5) Used for Prediction: After the line is created, it is used to **predict future values** such as sales, temperature, or system usage.

6) Simple and Interpretable: It is easy to understand, easy to compute, and widely used for **trend analysis and forecasting**.

b) Differentiate Overfitting and Underfitting with Example. [6]

Overfitting:: Overfitting happens when a model **learns too much from the training data**, including noise and unnecessary patterns. It performs **very well on training data** but **poorly on new/unseen data**.

Underfitting: Underfitting happens when a model is **too simple** and fails to learn important patterns from the data. It performs **poorly on both training and test data**.

Difference between Overfitting and Underfitting

| Feature | Overfitting | Underfitting |
|-------------------------------|---------------------------------------|---|
| Model Behavior | Model learns too many details & noise | Model learns too little & misses patterns |
| Training Accuracy | Very high | Low |
| Testing Accuracy | Low | Low |
| Model Complexity | Too complex | Too simple |
| Cause | Too many features or long training | Too few features or insufficient training |
| Generalization Ability | Poor | Poor |

Example**Overfitting Example:**

A regression model using **polynomial of degree 10** for simple linear data:

- Fits training data perfectly
- But gives wrong predictions on new data because it learned noise

Underfitting Example:

A linear model trying to fit **curved data**:

- Cannot capture the curve
- Produces large errors even on training data

c) Write a short note on Training Error and Generalization Error in Machine Learning. [6]

- 1) **Training Error:** Training error is the **error the model makes on the training dataset** — the same data that was used to train the model.

Key Points:

1. It measures how well the model has learned the patterns in the training data.
2. A **low training error** indicates good learning of the given data.
3. A **very low training error** may also indicate **overfitting**, where the model memorizes data instead of learning patterns.
4. It helps to check if the model needs more training or more complexity.
5. Training error reduces as the number of epochs/iterations increases.
6. It does **not** guarantee good performance on unseen data.

- 2) **Generalization Error:** Generalization error is the **error the model makes on new, unseen test data**. It shows how well the model performs in real-world situations.

Key Points:

1. It indicates the model's **ability to generalize** beyond the training data.
2. A **low generalization error** shows that the model predictions are accurate on new data.
3. A **high generalization error** usually means the model is **overfitting**.
4. It is measured using test data, validation data, or cross-validation techniques.
5. Generalization error helps determine if the model is suitable for real deployment.
6. A good model always keeps generalization error **low and consistent**.

Q2) a) Explain in brief the importance of Evaluation Metrics. [6]

Evaluation metrics are **quantitative measures** used to assess how well a machine-learning model performs.

They help determine the **accuracy, reliability, and usefulness** of a model before deploying it in real-world applications.

Importance of Evaluation Metrics

1) Measure Model Performance: Metrics show how accurately the model predicts outcomes on training and test data.

2) Detect Overfitting or Underfitting: By comparing training and testing scores, metrics help identify if a model is too simple or too complex.

3) Model Comparison: Different models can be compared using the same metric to select the **best performing model**.

4) Helps in Hyperparameter Tuning: Metrics guide how to adjust model settings (like learning rate, depth, parameters) to improve accuracy.

5) Ensures Real-World Reliability: Good evaluation metrics make sure the model performs well on unseen or future data.

6) Supports Better Decision-Making: Metrics help in understanding whether the model is suitable for deployment, improvement, or replacement.

b) Write a short note on Lasso and Ridge Regression. [6]

1) Lasso Regression: Lasso (Least Absolute Shrinkage and Selection Operator) Regression is a **regularization technique** that adds an **L1 penalty** to the loss function to prevent overfitting.

Key Points:

1. Uses **L1 regularization**, which can shrink some coefficients exactly to **zero**.
2. Helps in **feature selection**, keeping only the most important variables.
3. Useful when there are many features and we want a **simpler model**.
4. Reduces model complexity and improves generalization on new data.
5. Prevents overfitting by controlling the size of coefficient values.
6. Works well when some features are irrelevant or weakly related.

2) Ridge Regression: Ridge Regression is a regularization technique that adds an **L2 penalty** to the loss function to reduce overfitting and stabilize the model.

Key Points:

1. Uses **L2 regularization**, which **shrinks coefficients** but does **not** make them zero.
2. Useful in datasets with **multicollinearity** (highly correlated features).
3. Keeps all features but reduces their impact by lowering coefficient values.
4. Improves model stability and reduces variance.
5. Produces smoother, more generalized predictions on test data.
6. Good for situations where all features may contribute but with different strengths.

c) Memory (Capacity) and cost of RAM as shown in below table. [6]

| | | | | |
|------------------------------|----|----|----|----|
| X (Memory Capacity) in GB | 2 | 4 | 8 | 16 |
| Y (Cost in \$) | 12 | 16 | 28 | 62 |

- Find Regression line $Y = aX + b$ using least square method.
- Estimate the cost of 32 GB RAM using line as equation.

Given Data

| X (GB) | 2 | 4 | 8 | 16 |
|--------|----|----|----|----|
| Y (\$) | 12 | 16 | 28 | 62 |

Let: $n = 4$

Step 1: Calculate required sums

$$\sum X = 2 + 4 + 8 + 16 = 30$$

$$\sum Y = 12 + 16 + 28 + 62 = 118$$

$$\begin{aligned}\sum XY &= (2)(12) + (4)(16) + (8)(28) + (16)(62) \\ &= 24 + 64 + 224 + 992 = 1304\end{aligned}$$

$$\begin{aligned}\sum X^2 &= 2^2 + 4^2 + 8^2 + 16^2 \\ &= 4 + 16 + 64 + 256 = 340\end{aligned}$$

Step 2: Find slope (a)

Formula:

$$a = \frac{n \sum XY - (\sum X)(\sum Y)}{n \sum X^2 - (\sum X)^2}$$

Substitute:

$$a = \frac{4(1304) - 30(118)}{4(340) - 30^2}$$

$$a = \frac{5216 - 3540}{1360 - 900}$$

$$a = \frac{1676}{460} = 3.643$$

Step 3: Find intercept (b)

$$b = \frac{\sum Y - a \sum X}{n} = \frac{118 - (3.643)(30)}{4} = \frac{118 - 109.29}{4} = \frac{8.71}{4} = 2.1775$$

(i) **Regression Line:** $Y = 3.643X + 2.178$

(ii) **Estimate the cost of 32 GB RAM**

$$Y = 3.643(32) + 2.178 = 116.576 + 2.178 = 118.754$$

Estimated Cost of 32 GB RAM = \$118.75 (approx.)

➤ NOV / DEC 2023

Q1) a) Differentiate between overfitting and underfitting. [6]

→ DONE

- b) The table below shows the number of grams of carbohydrates, X and the number of Calories, Y of six different foods. Find linear regression equation for this dataset. [8]

| | | | | | | |
|-------------------|-----|-----|-----|----|-----|----|
| Carbohydrates (X) | 8 | 9.5 | 10 | 6 | 7 | 4 |
| Calories (Y) | 112 | 138 | 147 | 88 | 108 | 62 |

Also find the value of Y for X = 12

Step 1: Calculate required sums**Sum of X**

$$\sum X = 8 + 9.5 + 10 + 6 + 7 + 4 = 44.5$$

Sum of Y

$$\sum Y = 112 + 138 + 147 + 88 + 108 + 62 = 655$$

Sum of XY

$$\begin{aligned} XY &= (8)(112) + (9.5)(138) + (10)(147) + (6)(88) + (7)(108) + (4)(62) \\ &= 896 + 1311 + 1470 + 528 + 756 + 248 = 5209 \end{aligned}$$

Sum of X²

$$X^2 = 8^2 + 9.5^2 + 10^2 + 6^2 + 7^2 + 4^2 = 64 + 90.25 + 100 + 36 + 49 + 16 = 355.25$$

Step 2: Calculate slope (b)

$$b = \frac{n \sum XY - (\sum X)(\sum Y)}{n \sum X^2 - (\sum X)^2}$$

Substitute values:

$$b = \frac{6(5209) - (44.5)(655)}{6(355.25) - (44.5)^2} = \frac{31254 - 29147.5}{2131.5 - 1980.25} = \frac{2106.5}{151.25} = 13.93 \text{ (approx)}$$

Step 3: Calculate intercept (a)

$$a = \frac{\sum Y - b \sum X}{n} = \frac{655 - 13.93(44.5)}{6} = \frac{655 - 619.885}{6} = \frac{35.115}{6} = 5.8525$$

Regression Line: $Y = 5.85 + 13.93 X$ **Find Y for X = 12**

$$Y = 5.85 + 13.93(12) = 5.85 + 167.16 = 173.01$$

Predicted Calories for X = 12 g carbs = 173.01 calories (approx.)

c) Explain Bias–Variance Trade-off. [4]**Bias–Variance Trade-off – Explanation:****1) Bias**

- Bias is the **error due to overly simple assumptions** made by the model.
- High bias → model is too simple → leads to **underfitting**.

2) Variance

- Variance is the **error due to too much sensitivity** to the training data.
- High variance → model is too complex → leads to **overfitting**.

3) Trade-off

- Increasing model complexity decreases bias but **increases variance**.
- Decreasing model complexity increases bias but **reduces variance**.
- So, we must find a **balance** where both bias and variance are low.

4) Goal of Machine Learning

- The aim is to build a model that does **not underfit** (high bias) or **overfit** (high variance).
- A good model achieves the **optimal trade-off** and performs well on unseen data.

Q2) a) What is Linear Regression? Explain the concept of Ridge Regression. [9]

1) Linear Regression: Linear Regression is a statistical and machine-learning method used to model the **relationship between an dependent variable (Y)** and **one or more independent variables (X)** by fitting a **straight-line equation**.

Key Points:

1. The general equation of linear regression is:

$$Y = a + bX$$

where **a = intercept**, **b = slope**.

2. It predicts **continuous numerical values** such as price, temperature, sales, etc.
3. Linear regression finds the best-fit line using the **Least Squares Method**, which minimizes the sum of squared errors.
4. It helps identify **how much Y changes** when X increases by 1 unit.
5. It is simple, interpretable, and widely used for **trend analysis and forecasting**.

2) Ridge Regression: Ridge Regression is a type of **regularized linear regression** that uses **L2 regularization** to prevent **overfitting** by adding a penalty term to the loss function.

Key Points:

1. Ridge adds a penalty equal to the **sum of the squares** of the coefficients:

$$\text{Loss} = \sum (Y - \hat{Y})^2 + \lambda \sum b^2$$

where λ (**lambda**) controls the strength of regularization.

2. It **shrinks the coefficient values**, but does **not make them zero**.
3. Ridge is useful when dataset has **multicollinearity** (highly correlated inputs).
4. It reduces model complexity and improves **generalization performance** on test data.
5. Larger $\lambda \rightarrow$ stronger penalty \rightarrow smaller coefficients \rightarrow less variance.
6. Helps avoid overfitting by controlling how much the model depends on training data.
7. Ridge ensures more **stable predictions**, especially when features are many or noisy.
8. It keeps all features in the model but reduces their influence.
9. Commonly used in modern ML to prevent models from becoming too complex.

b) Explain the following Evaluation Metrics :

i) MAE \rightarrow DONE

ii) RMSE \rightarrow DONE

iii) $R^2 \rightarrow$ DONE

➤ NOV / DEC 2024

Q1) a) Explain Lasso Regression. Explain how Lasso Regression is used for Feature Selection. [6]

1) Lasso Regression: Lasso Regression (Least Absolute Shrinkage and Selection Operator) is a **regularization technique** that improves linear regression by adding an **L1 penalty** to the loss function.

Key Points:

1. Lasso adds a penalty equal to the **absolute values of coefficients**:
- $$\text{Loss} = \sum (Y - \hat{Y})^2 + \lambda \sum |b|$$
2. The parameter λ (**lambda**) controls the strength of regularization.
 3. It helps prevent **overfitting** by reducing the magnitude of coefficients.
 4. Lasso is useful when the dataset contains **many input variables**, and we want a simpler model.
 5. Due to L1 penalty, Lasso can completely **shrink some coefficients to zero**.
 6. It produces simpler, more interpretable models compared to standard linear regression.

2) How Lasso Regression is Used for Feature Selection

1. The L1 regularization term forces some coefficients to become **exactly zero** during training.
2. Features with coefficients reduced to zero are considered **unimportant** and removed from the model.
3. Only the most useful and relevant features keep non-zero coefficients.

4. This automatic elimination of unnecessary variables acts as **built-in feature selection**.
5. Lasso reduces model complexity while keeping only meaningful predictors.
6. This makes the model faster, more interpretable, and less likely to overfit.

b) Define different regression models. [6]

→ DONE

c) Describe the bias–variance trade-off and its relationship to underfitting and overfitting. [6]

1) Bias–Variance Trade-off: The bias–variance trade-off describes how the **complexity of a machine learning model** affects its prediction error. It explains that total error = **bias error + variance error**, and balancing both leads to the best model performance.

2) Bias

1. Bias is the error caused by **oversimplifying** the model.
2. High bias means the model cannot learn important patterns.
3. High bias models usually make systematic mistakes.

3) Variance

1. Variance is the error caused by the model being **too sensitive** to training data.
2. High variance means the model learns noise instead of pattern.
3. Such models perform poorly on new, unseen data.

4) Relationship to Underfitting and Overfitting

- **Underfitting:**
 1. Happens when **bias is high** and variance is low.
 2. Model is too simple → cannot capture data patterns.
 3. Leads to high training error and high test error.
- **Overfitting:**
 1. Happens when **variance is high** and bias is low.
 2. Model becomes too complex → memorizes training data.
 3. Low training error but high test error.

5) Trade-off Goal

1. The aim is to find a model with **low bias and low variance**.
2. Balanced complexity gives best performance on unseen data.
3. Achieving this balance improves generalization.

Q2) a) Explain the advantages of RMSE over MSE as an evaluation metric. [6]**1) RMSE is in the Same Units as the Target Variable**

- RMSE takes the square root of MSE, so the final value is in the **original units of Y**.
- This makes interpretation easier compared to MSE, which is in **squared units**.

2) More Intuitive and Understandable

- Since RMSE is in the same units as the output, it is easier to understand the magnitude of error.
- For example, "RMSE = 5 calories" is more meaningful than "MSE = 25 calories²".

3) Strong Penalty for Large Errors

- RMSE still retains the **squaring property**, meaning large errors have a big impact.
- This helps detect models that make occasional large mistakes.

4) Better for Sensitive Prediction Tasks

- RMSE is preferred when large prediction errors are costly (e.g., medical, finance, security).
- It highlights how severe the worst-case errors are.

5) Suitable for Continuous Data with Normal Error Distribution

- RMSE works very well when errors follow a normal distribution.
- Models can be compared effectively using RMSE in such cases.

6) Helps Optimize Models More Accurately

- Since it is sensitive to large deviations, RMSE encourages the model to **reduce high-impact errors**, improving overall performance.

b) What do you mean by Least Square Method? Explain least square method in the context of linear regression. [6]

1) Least Square Method: The Least Square Method is a mathematical technique used to find the **best-fit line** by **minimizing the sum of squared errors** between actual values and predicted values. It ensures that the fitted line represents the data as accurately as possible.

2) Explanation in Context of Linear Regression**1) Purpose of Least Squares**

- In linear regression, the goal is to find a line: $Y = a + bX$ that best fits the data points.

2) Error Calculation

- For each data point, error = $(Y_{\text{actual}} - Y_{\text{predicted}})$.
- Errors are **squared** to avoid negative cancellation and to give importance to bigger errors.

3) Minimizing Total Error

- The least square method minimizes: $\sum (Y - \hat{Y})^2$
- The line that gives the **minimum total squared error** is chosen as the regression line.

4) Finding Slope (b)

- The slope is calculated using: $b = \frac{n \sum XY - (\sum X)(\sum Y)}{n \sum X^2 - (\sum X)^2}$

5) Finding Intercept (a)

- Once slope is known, $a = \frac{\sum Y - b \sum X}{n}$

6) Produces the Best-Fit Line:

- The least squares method provides the line that gives the **least total error**, making it ideal for prediction and forecasting.

c) Write a short note on Stochastic gradient descent algorithms. [6]

→ DONE

➤ **Additional MAY / JUNE 2022 Question:****Q1) b) Find the Equation of linear Regression line using following data: [6]**Equation of the regression line of Y on X is: $Y = 1.5 + 1.1 X$ **Brief steps**

- Use simple linear regression form: $Y = a + bX$
- Compute required sums from data:
 $\sum X = 1 + 2 + 3 + 4 = 10$, $\sum Y = 3 + 4 + 5 + 7 = 19$
 $\sum X^2 = 1^2 + 2^2 + 3^2 + 4^2 = 30$, $\sum XY = 1 \cdot 3 + 2 \cdot 4 + 3 \cdot 5 + 4 \cdot 7 = 60$
- Number of pairs $n = 4$
- Slope: $b = \frac{n \sum XY - (\sum X)(\sum Y)}{n \sum X^2 - (\sum X)^2} = \frac{4 \cdot 60 - 10 \cdot 19}{4 \cdot 30 - 10^2} = \frac{44}{40} = 1.1$
- Intercept: $a = \bar{Y} - b \bar{X} = \frac{19}{4} - 1.1 \cdot \frac{10}{4} = 4.75 - 2.75 = 2.0$

(or directly from standard intercept formula).

So the fitted linear regression line is $Y = 2.0 + 1.1 X$

| X | Y |
|---|---|
| 1 | 3 |
| 2 | 4 |
| 3 | 5 |
| 4 | 7 |

NOTE: Please verify all answers before referring.