

Note:

I am new to Machine learning, and I have knowledge in concepts like regression and NLP for now so to answer some of the questions in the task I had to do a lot of search and watched some videos so my answers may not be perfect or fully correct.

## **1. What is the rank of a matrix? Why is it important in machine learning?**

The rank of a matrix is a number of how many rows or columns in the matrix are independent. Independence means that no row or column can be made by combining other rows or columns.

Why is Rank Important in Machine Learning:

1-Understanding Data:

The rank tells us how many unique features are in the data. If the rank is small, some features might be redundant or useless.

2- Regularization:

when the data matrix is low rank, regularization techniques like ridge regression or L2 regularization are used to stabilize the solution.

3- Dimensionality Reduction:

The rank of a matrix shows the number of linearly independent features or dimensions in the data. Techniques like Principal Component Analysis (PCA) rely on the rank to reduce the dimensionality of data while preserving as much variance as possible.

## **2. Explain eigenvalues and eigenvectors. How are they used in PCA?**

An eigenvector is a special vector that doesn't change when a matrix transformation is applied to it. It only gets scaled (stretched or squished) by a factor and this factor is the Eigenvalue.

Principal Component Analysis (PCA) is a technique used to reduce the dimensionality of data while keeping as much information as possible.

Eigenvalues and eigenvectors are very important for doing PCA. And the PCA is done by:

1- Center the Data:

Subtract the mean from each feature so the data is centered around zero.

2- Compute the Covariance Matrix:

The covariance matrix shows how features in the data are related to each other.

### 3- Find Eigenvalues and Eigenvectors:

Calculate the eigenvalues and eigenvectors for the covariance matrix. Eigenvectors will represent the directions in which the data varies the most and Eigenvalues will show how much variance is in each of these directions.

### 4- Sort and Select:

Sort the eigenvalues in descending order. The largest eigenvalue shows the most important direction (principal component). Then we Select the top k eigenvectors (principal components) that capture most of the variance in the data.

### 5- Transform the Data and reduce its dimensions:

We transform the data using the z score and multiple it by the principal components

## **4. How would you compute the inverse of a matrix? What is its relevance in Machine Learning?**

To compute the inverse of matrix the steps are:

1- check if the matrix is square because only square matrices can have an inverse.

2-check if the matrix determinant is not 0 because if it is 0 the matrix is singular and do not have an inverse.

3- we find the adjoint matrix and divide it by the determinant of the matrix

After I did some search, I found out this method that we learned in the university course isn't the best method and is computationally expensive and instead methods like Gaussian elimination or LU decomposition are preferred for larger matrices.

Relevance of Matrix Inversion in Machine Learning:

1- used in solving Linear Equations for linear regression models for example.

2- used to remove transformation applied to a matrix.

## **5. What is the difference between correlation and causation?**

Correlation measures the relationship between 2 variables and shows how one of them changes when the other changes it can be Negative Correlation where one increases and the other decreases or positive where the increase together.

Causation means the that an event happening will make another event happen in math terms one variable causes change to the other.

Correlation and causation might sound similar and someone may say they are the same thing but now correlation doesn't mean causation for example we can say that the people how go to the hospitals have higher death rates and these 2 variables are correlated is this

because the hospital kills now but because people how are seriously ill will go to the hospital.

## **6. How does gradient descent work? Explain its importance in ML.**

Gradient Descent is an optimization algorithm used to minimize a function by iteratively moving toward the minimum value of the function. It works by minimizing the cost function. it works by finding the minimum value of the function by:

For a generic gradient descent workflow

1- we set initial random parameters for the cost function

2- we calculate the cost function

3- we find the gradient which is the partial derivatives of the loss function with respect to each parameter and it indicates the direction and the rate that the cost function changes by

4- we need to adjust the parameters using the gradients we computed

5- we repeat steps 2 to 4 until we converge or find the minimum value of the cost function which means the parameters we have no longer changes

Why is Gradient Descent Important in Machine Learning?

1- it helps to find parameter values and optimize a lot of ml models like linear regression, logistic regression, neural networks

2- it is very efficient and scalable and automated which make it easier to apply and maintain

## **7. What are the types of ML? Discuss each type and give examples on them.**

### **1- Supervised Learning**

In supervised learning, the model is trained on labeled data, where the input data is paired with the correct output. The goal is to learn to find the outputs from the inputs.

Key Characteristics:

1- Requires labeled data.

2- The model is trained to predict outputs for new, unseen inputs.

Examples of Supervised Learning:

## 1- Classification:

Which means Predicting discrete categories or labels.

Example: Email spam detection, input: email text, output: spam or not spam.

Algorithms: Logistic Regression, Decision Trees, Neural Networks.

## 2- Regression:

Which means Predicting continuous values.

Example: Predicting house prices, input: house features like size, location, output: price.

Algorithms: Linear Regression, Random Forests, Gradient Boosting.

## **2- Unsupervised Learning**

In unsupervised learning, the model is trained on unlabeled data, and the goal is to find hidden patterns or structures in the data.

Key Characteristics:

- No labeled data is required.

- The model learns to find patterns or relationships in the data.

Examples of Unsupervised Learning:

### 1- Clustering:

Which means Grouping similar data points together.

Example: Customer segmentation, input: customer purchase history, output: groups of similar customers.

Algorithms: K-Means, Hierarchical Clustering.

### 2- Dimensionality Reduction:

Which means Reducing the number of features while preserving important information.

Example: Visualizing high-dimensional data in 2D or 3D.

Algorithms: Principal Component Analysis (PCA).

### 3- Anomaly Detection:

Identifying unusual or rare data points.

Example: Detecting fraudulent transactions input: transaction data, output: normal or anomalous.

### 3. Reinforcement Learning

In reinforcement learning, an agent learns to make decisions by interacting with an environment. The agent receives feedback in the form of rewards or penalties and aims to maximize rewards over time.

Key Characteristics:

- No labeled data; learning happens through trial and error.
- The agent learns a strategy to take actions in an environment.

Examples of Reinforcement Learning:

1- Game Playing:

Example: Training an AI to play chess or Go input: game state, output: move to make.

2- Robotics:

Example: Training a robot to walk or perform tasks (input: sensor data, output: motor actions).

3- Recommendation Systems:

Example: Personalizing recommendations based on user interactions input: user behavior, output: recommended item.

### 8. Explain this code and provide a way to optimize it:

This SQL query finds the distinct productid from the sales table and find the total sales in it using a correlated sub query which is evaluated once for every product id and generally correlated sub queries have really bad performance and a good practice is to replace them an optimize code can use the GROUP BY clause which makes the query more efficient and readable.

```
SELECT ProductID, SUM(Amount) AS TotalSales
FROM Sales
GROUP BY ProductID;
```