Assignment

Q1. What is Min-Max scaling, and how is it used in data preprocessing? Provide an example to illustrate its
application.

Ans:Min-Max scaling is a data preprocessing technique used to normalize the features of a dataset by scaling them to a specific range, typically between 0 and 1. The goal of Min-Max scaling is to transform the values of the features while preserving their relative relationships and ensuring that they fall within a consistent numerical range. This scaling is particularly useful when features have different ranges and magnitudes, as it can help prevent certain features from dominating others during the training of machine learning models.

The formula for Min-Max scaling is given by:

�scaled=�−�min�max−�min

X

scaled

$=$

X

max

$-X$

min

$X-X$

min

where:

- �

- X is the original value of a feature.
- �min
- X
- min
-
- is the minimum value of that feature in the dataset.
- �max
- X
- max
-
- is the maximum value of that feature in the dataset.
- �scaled
- X
- scaled
-
- is the scaled value of the feature.

Here's an example to illustrate the application of Min-Max scaling:

Suppose you have a dataset with a feature "Age" and another feature "Income." The "Age" values range from 20 to 60, while the "Income" values range from 30,000 to 120,000.

Original data:

- Age: 20, 30, 40, 50, 60
- Income: 30,000, 50,000, 80,000, 100,000, 120,000

To apply Min-Max scaling:

For the "Age" feature:

- �min=20
- X
- min
-
- =20 (minimum age)
- �max=60
- X

- max
-
- =60 (maximum age)
- �scaled=�−�min�max−�min
- X
- scaled
-
- =
- X
- max
-
- $-X$
- min
-
- $X-X$
- min
-
  -
-

So, for the ages 20, 30, 40, 50, and 60, the scaled values would be:

- �scaled=20−2060−20=0
- X
- scaled
-
- =
  - 60−20
  - 20−20
  -
- =0
- �scaled=30−2060−20=0.2
- X
- scaled
-
- =
  - 60−20
  - 30−20
  -
- =0.2
- �scaled=40−2060−20=0.4

- X
- scaled
-
- =
    - 60−20
    - 40−20
        -
- =0.4
- �scaled=50−2060−20=0.6
- X
- scaled
-
- =
    - 60−20
    - 50−20
        -
- =0.6
- �scaled=60−2060−20=1
- X
- scaled
-
- =
    - 60−20
    - 60−20
        -
- =1

For the "Income" feature:

- �min=30,000
- X
- min
-
- =30,000 (minimum income)
- �max=120,000
- X
- max
-
- =120,000 (maximum income)
- �scaled=�−�min�max−�min

- X
- scaled
-
- =
- X
- max
-
- $-X$
- min
-
- $X-X$
- min
-
-
-

So, for the incomes 30,000, 50,000, 80,000, 100,000, and 120,000, the scaled values would be:

- �scaled=30,000−30,000120,000−30,000=0
- X
- scaled
-
- =
  - 120,000−30,000
  - 30,000−30,000
  -
- =0
- �scaled=50,000−30,000120,000−30,000=0.2
- X
- scaled
-
- =
  - 120,000−30,000
  - 50,000−30,000
  -
- =0.2
- �scaled=80,000−30,000120,000−30,000=0.6
- X
- scaled
-
- =

- 120,000−30,000
- 80,000−30,000
-

- =0.6
- �scaled=100,000−30,000120,000−30,000=0.8
- X
- scaled
-
- =
  - 120,000−30,000
  - 100,000−30,000
  -
- =0.8
- �scaled=120,000−30,000120,000−30,000=1
- X
- scaled
-
- =
  - 120,000−30,000
  - 120,000−30,000
  -
- =1

After Min-Max scaling, both "Age" and "Income" features will now have values in the range [0, 1], making them comparable and preventing one feature from dominating the other due to differences in scale. This is especially beneficial for algorithms sensitive to feature magnitudes, such as distance-based algorithms or those using gradient descent optimization.

Q2. What is the Unit Vector technique in feature scaling, and how does it differ from Min-Max scaling?
Provide an example to illustrate its application.
Ans:The Unit Vector technique in feature scaling is also known as vector normalization or L2 normalization. It involves scaling individual data points to have a length of 1, effectively transforming them into unit vectors. This normalization method is particularly useful when the direction of the data points is more important than their magnitudes. The technique is applied to each data point separately, ensuring that the resulting vectors have a Euclidean norm (L2 norm) of 1.

The formula for Unit Vector scaling is given by:

Unit Vector=�‖�‖2

Unit Vector=

$\| X \|$

2

X

where:

- �
- X is the original vector.
- $\|$ �‖2
- $\| X \|$
- 2
-
- is the Euclidean norm (L2 norm) of the vector
- �
- X.

The primary difference between Unit Vector scaling and Min-Max scaling lies in their objectives. While Min-Max scaling aims to normalize the range of values for each feature, Unit Vector scaling focuses on the direction of the data points, making them comparable in terms of their orientation rather than their magnitudes.

Here's an example to illustrate the application of Unit Vector scaling:

Suppose you have a dataset with two features, "X" and "Y," representing the coordinates of data points in a 2-dimensional space.

Original data points:

- Data Point 1: (3, 4)
- Data Point 2: (1, 2)

To apply Unit Vector scaling:

Calculate the Euclidean norm for each data point:

- For Data Point 1:
- // �‖2=32+42=5
- // X //
- 2
-
- =
- 3
- 2
- +4
- 2
-
- =5
- For Data Point 2:
- // �‖2=12+22=5
- // X //
- 2
-
- =
- 1
- 2
- +2
- 2
-
- =
- 5
-
-

Scale each data point to have a unit vector:

For Data Point 1:

-

- Unit Vector for Data Point 1=(35,45)
- Unit Vector for Data Point 1=(
  - 5
  - 3
  -
- ,
  - 5
  - 4
  -
-)

For Data Point 2:

-
- Unit Vector for Data Point 2=(15,25)
- Unit Vector for Data Point 2=(
- 5
-
  - 1
  -
- ,
- 5
-
  - 2
  -
-)

After Unit Vector scaling, the data points are now represented as unit vectors, emphasizing their directions while making them comparable in terms of orientation. Note that the magnitudes of the vectors are now 1.

Comparison with Min-Max Scaling:

- Min-Max scaling would have focused on scaling the individual features to a specific range, such as [0, 1], without considering the overall vector length or direction.

In summary, Unit Vector scaling is particularly useful when the orientation or direction of data points is more relevant than their magnitudes. It is commonly used in machine learning applications where the relative angles or relationships between vectors are crucial, such as in clustering or classification tasks.

Q3. What is PCA (Principle Component Analysis), and how is it used in dimensionality reduction? Provide an
example to illustrate its application.
Ans:Principal Component Analysis (PCA) is a dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional representation while retaining as much of the original variance as possible. The fundamental idea behind PCA is to identify the principal components, which are linear combinations of the original features, that capture the maximum variance in the data.

The steps involved in PCA are as follows:

Standardize the Data:
- If the features of the dataset have different scales, it is common practice to standardize the data (subtract the mean and divide by the standard deviation) to ensure that all features contribute equally to the PCA.

Compute the Covariance Matrix:
- Calculate the covariance matrix for the standardized data. The covariance matrix provides information about the relationships between different features.

Compute Eigenvectors and Eigenvalues:
- Find the eigenvectors and eigenvalues of the covariance matrix. Eigenvectors represent the directions or components of maximum variance, while eigenvalues quantify the amount of variance along each eigenvector.

Sort Eigenvectors by Eigenvalues:
- Sort the eigenvectors based on their corresponding eigenvalues in descending order. The eigenvectors with higher eigenvalues capture more variance and are considered the principal components.

Select Principal Components:
- Choose the top
- �
- k eigenvectors to form the new feature space. The value of
- �
- k is determined by the desired dimensionality of the reduced dataset.

Projection:
- Project the original data onto the selected principal components to obtain the lower-dimensional representation.

PCA is widely used in various fields, including image processing, pattern recognition, and data compression. It is particularly useful when dealing with datasets with a large number of correlated features.

Here's a simplified example to illustrate PCA's application:

Suppose you have a dataset with two features, "X" and "Y," representing points in a 2-dimensional space.

Original data points:

- Data Point 1: (2, 3)
- Data Point 2: (3, 4)
- Data Point 3: (5, 6)

  Standardize the Data:
  - If needed, standardize the data by subtracting the mean and dividing by the standard deviation.

  Compute Covariance Matrix:
  - Calculate the covariance matrix for the standardized data.

  Compute Eigenvectors and Eigenvalues:
  - Find the eigenvectors and eigenvalues of the covariance matrix.

  Sort Eigenvectors by Eigenvalues:
  - Assume the sorted eigenvectors are
  - �1
  - **v**
  - 1
  -
  - and
  - �2
  - **v**
  - 2
  -
  - , and their corresponding eigenvalues are
  - �1
  - λ
  - 1
  -

- and
- �2
- λ
- 2
-
- , respectively.

Select Principal Components:

- If you choose to reduce the dimensionality to 1 (
- �=1
- k=1), you would select the top eigenvector
- �1
- **v**
- 1
-
- .

Projection:

- Project the original data onto the selected principal component to obtain the lower-dimensional representation.

The result might be a reduced dataset with a single feature based on the chosen principal component, capturing the most significant variation in the original data. The reduced dataset can be used for analysis, visualization, or feeding into machine learning models with fewer dimensions.

Q4. What is the relationship between PCA and Feature Extraction, and how can PCA be used for Feature
Extraction? Provide an example to illustrate this concept.
Ans:PCA (Principal Component Analysis) can be considered a form of feature extraction, as it transforms the original features of a dataset into a new set of uncorrelated variables called principal components. Feature extraction aims to capture the most important information from the original features in a more compact representation, reducing dimensionality while retaining as much relevant information as possible. PCA achieves this by identifying the directions (principal components) in which the data varies the most.

The relationship between PCA and feature extraction can be summarized as follows:

Dimensionality Reduction:

- PCA is primarily used for dimensionality reduction. It projects the data onto a lower-dimensional subspace defined by the principal components. This reduction in dimensionality is a form of feature extraction, as the new features (principal components) capture the most significant information from the original features.

Decorrelation of Features:

- PCA aims to decorrelate the features by identifying the directions of maximum variance in the data. In the new feature space (defined by principal components), the features are uncorrelated. This decorrelation is beneficial for downstream machine learning tasks, as it simplifies the relationships between features.

Here's an example to illustrate how PCA can be used for feature extraction:

Suppose you have a dataset with three features: "X1," "X2," and "X3," representing measurements of different physical quantities.

Original data points:

- Data Point 1: (2, 3, 4)
- Data Point 2: (1, 5, 2)
- Data Point 3: (4, 4, 6)

Standardize the Data:

- If needed, standardize the data by subtracting the mean and dividing by the standard deviation.

Apply PCA:

- Compute the covariance matrix, find the eigenvectors and eigenvalues, and sort them in descending order.

Select Principal Components:

- Decide on the number of principal components to retain based on the desired level of dimensionality reduction. Let's say you choose to retain the top two principal components.

Projection:

- Project the original data onto the selected two principal components to obtain a new set of features.

The result might look like this:

- Transformed Data Point 1: (1.5, -0.5) *(Projection onto the first two principal components)*
- Transformed Data Point 2: (-2.2, 0.7)
- Transformed Data Point 3: (2.2, -0.2)

In this transformed feature space, the new features (principal components) are uncorrelated and capture the most significant information from the original features. You have effectively extracted a reduced set of features that retains as much variance as possible while reducing the dimensionality of the dataset.

This extracted feature representation can be used for various purposes, such as visualization, clustering, or feeding into machine learning models with reduced input dimensions.

Q5. You are working on a project to build a recommendation system for a food delivery service. The dataset
contains features such as price, rating, and delivery time. Explain how you would use Min-Max scaling to
preprocess the data.
Ans:Min-Max scaling is a data preprocessing technique used to transform the features of a dataset so that they fall within a specific range, typically between 0 and 1. This technique is useful when features have different scales, and you want to normalize them to a common range, preventing certain features from dominating others due to differences in magnitudes. Here's how you can use Min-Max scaling to preprocess the data for building a recommendation system for a food delivery service:

Assuming you have features like "price," "rating," and "delivery time" in your dataset:

Understand the Data:

- Gain a thorough understanding of the dataset, including the nature of the features, their distributions, and their scales.

Identify Features for Scaling:

- Determine which features need to be scaled. In the case of a recommendation system for a food delivery service, features like "price," "rating," and "delivery time" may have different units and scales.

Choose a Scaling Range:

- Decide on the scaling range. Min-Max scaling typically scales features to a range between 0 and 1, but you can choose a different range based on the specific requirements of your recommendation system.

Apply Min-Max Scaling:

- For each feature
- �

X, apply the Min-Max scaling formula:

- �scaled=�−�min�max−�min
- X
- scaled
- $=$
- X
- max
- $-X$
- min
- $X-X$
- min
- Where
- �min
- X
- min
- is the minimum value of the feature,
- �max
- X
- max
- is the maximum value of the feature, and
- �scaled
- X
- scaled
- is the scaled value.

For example, let's say you have a "price" feature with values ranging from $5 to $20, a "rating" feature with values from 2 to 5, and a "delivery time" feature with values in minutes ranging from 15 to 45.

Price:

- Pricescaled=Price−520−5
- Price
- scaled
-
- =
  - 20−5
  - Price−5
  -
-

Rating:

-
- Ratingscaled=Rating−25−2
- Rating
- scaled
-
- =
  - 5−2
  - Rating−2
  -
-

Delivery Time:

-
- Delivery Timescaled=Delivery Time−1545−15
- Delivery Time
- scaled
-
- =
  - 45−15
  - Delivery Time−15
  -
-

This scaling ensures that each feature is transformed to a common range (e.g., [0, 1]).

Updated Dataset:

- Replace the original values of the features with their scaled counterparts in the dataset.

Normalization Parameters:

- Keep track of the scaling parameters (e.g.,
- �min
- X
- min
-
- and
- �max
- X
- max
-
-) for each feature. These parameters may be needed later when making predictions or recommendations.

Min-Max scaling helps create a standardized representation of the features, making it easier for the recommendation system to consider and compare different aspects, such as price, rating, and delivery time, without being influenced by differences in their original scales. This normalization step is beneficial for many machine learning algorithms, ensuring that each feature contributes equally to the model's training and improving the system's overall performance.

Q6. You are working on a project to build a model to predict stock prices. The dataset contains many
features, such as company financial data and market trends. Explain how you would use PCA to reduce the
dimensionality of the dataset.
Ans:When dealing with a dataset containing numerous features, such as financial data and market trends for predicting stock prices, PCA (Principal Component Analysis) can be employed for dimensionality reduction. PCA helps in transforming the high-dimensional data into a lower-dimensional representation while retaining as much variance as possible. Here's how you can use PCA for dimensionality reduction in the context of predicting stock prices:

Standardize the Data:

- If the features in your dataset have different scales, it's a good practice to standardize the data by subtracting the mean and dividing by the standard deviation. This ensures that all features contribute equally to the PCA.

Compute the Covariance Matrix:

- Calculate the covariance matrix for the standardized data. The covariance matrix provides information about the relationships between different features.

Compute Eigenvectors and Eigenvalues:

- Find the eigenvectors and eigenvalues of the covariance matrix. Eigenvectors represent the directions or components of maximum variance, while eigenvalues quantify the amount of variance along each eigenvector.

Sort Eigenvectors by Eigenvalues:

- Sort the eigenvectors based on their corresponding eigenvalues in descending order. The eigenvectors with higher eigenvalues capture more variance and are considered the principal components.

Select Principal Components:

- Choose the top
- �
- k eigenvectors to form the new feature space. The value of
- �
- k is determined by the desired level of dimensionality reduction. You may choose a value based on the explained variance or based on the specific requirements of your project.

Projection:

- Project the original data onto the selected
- �
- k principal components to obtain the lower-dimensional representation.

The reduced dataset will have

�

k features, where

�

k is less than the original number of features. This new set of features, the principal components, captures the most significant information from the original features.

Here are the steps in a more detailed manner:

Example:

Suppose you have a dataset with financial features such as revenue, expenses, profit margin, and market trend features such as trading volume, volatility, and overall market performance.

Standardize the Data:

- Subtract the mean and divide by the standard deviation for each feature.

Compute the Covariance Matrix:

- Calculate the covariance matrix using the standardized data.

Compute Eigenvectors and Eigenvalues:

- Find the eigenvectors and eigenvalues of the covariance matrix.

Sort Eigenvectors by Eigenvalues:

- Sort the eigenvectors in descending order based on their corresponding eigenvalues.

Select Principal Components:

- Decide on the number of principal components (
- �
- k) to retain. This decision can be based on the explained variance or specific project requirements.

Projection:

- Project the original data onto the selected
- �
- k principal components to obtain the lower-dimensional representation.

By using PCA, you have effectively reduced the dimensionality of the dataset while retaining the most critical information. The reduced set of features can be used to train a predictive model for stock price prediction, potentially improving computational efficiency and mitigating the risk of overfitting. Keep in mind that the choice of

�

k may involve trade-offs between dimensionality reduction and preserving information, and it

may require experimentation to find the optimal value for your specific project.

Q7. For a dataset containing the following values: [1, 5, 10, 15, 20], perform Min-Max scaling to transform the
values to a range of -1 to 1.
Ans:To perform Min-Max scaling on a dataset and transform the values to a range of -1 to 1, you can use the following formula:

�scaled=�−�min�max−�min

X

scaled

=

X

max

$-X$

min

$X-X$

min

where:

- �
- X is the original value in the dataset,
- �min
- X
- min
-
- is the minimum value in the dataset,
- �max
- X
- max
-
- is the maximum value in the dataset, and
- �scaled
- X
- scaled
-
- is the scaled value.

For the given dataset [1, 5, 10, 15, 20], let's calculate the Min-Max scaling with a target range of -1 to 1:

Find
�min
X
min

and
�max
X
max

:

- �min=1
- X
- min
-
- $=1$ (minimum value in the dataset)
- �max=20
- X
- max
-
- $=20$ (maximum value in the dataset)

Apply Min-Max Scaling:

- For each value
- �
- X in the dataset, calculate
- �scaled
- X
- scaled
-
- using the formula.

�scaled=�−�min�max−�min
X
scaled

$=$

X

max

$-X$

min

$X-X$

min

- Let's calculate for each value:
  - For
  - �=1

$X=1$:

  -
  - �scaled=1−120−1=0
  - X
  - scaled
  -
  - $=$
    - 20−1
    - 1−1
    -
  - =0
  - For
  - �=5

$X=5$:

  -
  - �scaled=5−120−1=419
  - X
  - scaled
  -
  - $=$
    - 20−1

- 5−1
-
- =
- 19
- 4
-
- For
- �=10

X=10:

-
- �scaled=10−120−1=919
- X
- scaled
-
- =
- 20−1
- 10−1
-
- =
- 19
- 9
-
- For
- �=15

X=15:

-
- �scaled=15−120−1=1419
- X
- scaled
-
- =
- 20−1
- 15−1
-
- =

- 19
- 14
-

- For
- �=20

X=20:

-
- �scaled=20−120−1=1
- X
- scaled
-
- =
  - 20−1
  - 20−1
  -
- =1

Scale to the Range of -1 to 1:

Now that the values are between 0 and 1, scale them to the desired range of -1 to 1.

- X_{\text{scaled}}_{\text{final}} = 2 \times (X_{\text{scaled}} - 0.5)
- For each value, calculate the final scaled value:
  - For
  - �=1

X=1:

- X_{\text{scaled}}_{\text{final}} = 2 \times (0 - 0.5) = -1
- For
- �=5

X=5:

- X_{\text{scaled}}_{\text{final}} = 2 \times \left(\frac{4}{19} - 0.5\right)
- For
- �=10

X=10:

- X_{\text{scaled}}_{\text{final}} = 2 \times \left(\frac{9}{19} - 0.5\right)
- For
- �=15

X=15:

- X_{\text{scaled}}_{\text{final}} = 2 \times \left(\frac{14}{19} - 0.5\right)
- For
- �=20

X=20:

- X_{\text{scaled}}_{\text{final}} = 2 \times (1 - 0.5) = 1

So, the Min-Max scaled values for the given dataset [1, 5, 10, 15, 20] in the range of -1 to 1 would be approximately [-1, -0.5789, 0.2632, 0.9474, 1].

Q8. For a dataset containing the following features: [height, weight, age, gender, blood pressure], perform
Feature Extraction using PCA. How many principal components would you choose to retain, and why?
Ans:The decision of how many principal components to retain in PCA is crucial and depends on the explained variance, which indicates the proportion of the total variance in the original dataset that is captured by each principal component. A common approach is to retain enough principal components to explain a high percentage of the total variance, typically a threshold like 95% or 99%. Here's how you can perform feature extraction using PCA for the given dataset:

Standardize the Data:
- If the features have different scales, standardize the data by subtracting the mean and dividing by the standard deviation for each feature.

Compute Covariance Matrix:
- Calculate the covariance matrix for the standardized data.

Compute Eigenvectors and Eigenvalues:
- Find the eigenvectors and eigenvalues of the covariance matrix.

Sort Eigenvectors by Eigenvalues:
- Sort the eigenvectors in descending order based on their corresponding eigenvalues.

Calculate Explained Variance:
- Calculate the explained variance for each principal component. The explained variance of a principal component is given by the ratio of its eigenvalue to the sum of all eigenvalues.

Determine the Number of Principal Components to Retain:

- Decide on the number of principal components to retain based on the desired explained variance. A common threshold is to retain enough components to explain at least 95% or 99% of the total variance.

Projection:

- Project the original data onto the selected principal components to obtain the lower-dimensional representation.

Here's a general guideline for determining the number of principal components to retain:

- If the explained variance of the first few principal components is very high (e.g., 95% or more), you may choose to retain those components.
- If the explained variance sharply decreases after a certain point, you might consider retaining fewer components up to that point.

It's common to plot a cumulative explained variance curve to visualize how much variance is explained by each additional principal component. The point where adding more components doesn't significantly increase the explained variance is often chosen as the cutoff.

In the context of the given dataset with features [height, weight, age, gender, blood pressure], the choice of the number of principal components to retain would depend on the specific dataset and the amount of variance you want to preserve. You may perform the analysis, plot the cumulative explained variance curve, and then make an informed decision based on the characteristics of your data and the desired level of dimensionality reduction.