19 March Assignment

May 26, 2023

[ ]:
```
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 that scales the data to a␣
↪fixed range of [0,1]. It is used to normalize the data
so that all features are on the same scale. This technique is useful when the␣
↪data has different scales and ranges. Min-Max scaling
is done using the following formula:

x_scaled = (x - x_min) / (x_max - x_min)

where x is the feature value, x_min is the minimum value of that feature, and␣
↪x_max is the maximum value of that feature. The scaled
value of x will be between 0 and 1.

For example, suppose we have a dataset with two features: age and income. The␣
↪age feature has a range of [0,100], while the income
feature has a range of [0,100000]. If we apply Min-Max scaling to this dataset,␣
↪both features will be scaled to the range [0,1].
This will ensure that both features are on the same scale and have equal␣
↪importance in the analysis.
```

[ ]:
```

```

[ ]:
```
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 is a feature scaling method that scales the values of␣
↪a feature to a range between -1 and 1. It is also
known as L2 normalization. This technique is useful when we have sparse data.␣
↪Sparse data is when we have a lot of zeros in our data.
```

```
For example, if we have a dataset with 1000 features and only 10 of them have␣
↪non-zero values, then we can use the Unit Vector
technique to scale the values of these 10 features between -1 and 1 1.

On the other hand, Min-Max scaling scales the values of a feature to a range␣
↪between 0 and 1. This technique is useful when we have
features with hard boundaries. For example, when dealing with any image file,␣
↪the colors can range from only 0 to 255 2.

Here's an example to illustrate the application of these techniques:

Suppose we have a dataset with two features: age and income. The age ranges␣
↪from 0 to 100 and income ranges from 0 to 100000. We want
to scale these features using both techniques.

Using Min-Max scaling:

age: (age - min(age)) / (max(age) - min(age)) = (age - 0) / (100 - 0) = age /␣
↪100
income: (income - min(income)) / (max(income) - min(income)) = (income - 0) /␣
↪(100000 - 0) = income / 100000
Using Unit Vector technique:

age: age / sqrt(age^2 + income^2)
income: income / sqrt(age^2 + income^2)
```

[ ]:

```

```

[ ]:

```
Q3. What is PCA (Principle Component Analysis), and how is it used in␣
↪dimensionality reduction? Provide an
example to illustrate its application.
```

ANS -

[ ]:

```
PCA (Principal Component Analysis) is a technique used for dimensionality␣
↪reduction. It is a statistical method that reduces the
number of variables in a dataset while retaining most of the information. PCA␣
↪is used to transform the data into a new coordinate
system such that the first axis has the largest possible variance, and each␣
↪succeeding axis has the highest variance possible under
the constraint that it is orthogonal to the preceding axes.

An intuitive example of dimensionality reduction can be discussed through a␣
↪simple e-mail classification problem, where we need to
classify whether the e-mail is spam or not. This can involve a large number of␣
↪features, such as whether or not the e-mail has a
```

```
generic title, the content of the e-mail, whether the e-mail uses a template,
etc.

Here's an example of how PCA can be used for dimensionality reduction: Suppose
we have a dataset with 1000 examples and 20 input
features. We can use PCA to reduce this dataset to 15 input features while
retaining most of the information.
```

[ ]:

```
[ ]:
```

[ ]:

```
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) is a linear feature extraction method that
can be used to obtain required variables
(important ones) from a large set of variables available in a data set¹. PCA is
used to decompose a multivariate dataset into a set
of successive orthogonal components that explain a maximum amount of the
variance.

PCA can be used for a variety of purposes, including data visualization,
feature selection, and data compression. PCA is basically a
method to obtain required variables (important ones) from a large set of
variables available in a data set¹. One of the examples of
linear feature extraction is PCA (Principal Component Analysis). A principal
component is a normalized linear combination of the
original features in a dataset.

For example, consider an image classification problem where we want to use the
red, green and blue components of each pixel in an
image to classify the image (e.g. detect dogs versus cats). Image sensors that
are most sensitive to red light also capture some
blue and green light. PCA can be used as a decorrelation method when features
are correlated.
```

[ ]:

```
[ ]:
```

[ ]:

```
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 normalization technique that enables us to scale data in a␣
  ↪dataset to a specific range using each feature's
minimum and maximum value. It shrinks the data within the given range, usually␣
  ↪of 0 to 1. It transforms data by scaling features to
a given range. It scales the values to a specific value range without changing␣
  ↪the shape of the original distribution.

In this case, we can use Min-Max scaling to preprocess the data by scaling the␣
  ↪features such as price, rating, and delivery time to
a specific value range without changing their shape. This will help us compare␣
  ↪these features on the same scale and avoid any bias
that may arise due to differences in their scales.

To perform Min-Max scaling on the dataset, we can use the following formula:

x_std = (x – x.min (axis=0)) / (x.max (axis=0) – x.min (axis=0))

x_scaled = x_std * (max – min) + min

Where,

min, max = feature_range

x.min (axis=0) : Minimum feature value

x.max (axis=0): Maximum feature value

This transformation is often used as an alternative to zero mean, unit variance␣
  ↪scaling.
```

[ ]:
```
```

[ ]:
```
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 -

[ ]:
```
 PCA (Principal Component Analysis) is a technique used to reduce the␣
  ↪dimensionality of your dataset by transforming it into a new
coordinate system. The reduced features are called principal components or␣
  ↪latent features. These principal components are simply a
linear combination of the original features in your dataset. The components␣
  ↪have two major properties: they are orthogonal
(perpendicular) and they capture the maximum amount of variance in the data.
```

```
To use PCA to reduce the dimensionality of your dataset, you would first
 ↪standardize your data so that each feature has a mean of
zero and a standard deviation of one. Then you would compute the covariance
 ↪matrix of the standardized data. Next, you would compute
the eigenvectors and eigenvalues of the covariance matrix. The eigenvectors
 ↪represent the principal components, and the eigenvalues
represent the amount of variance captured by each principal component. Finally,
 ↪you would select the top k eigenvectors that capture
most of the variance in your data.
```

[ ]:

[ ]:
```
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 the dataset containing the following values: [1,
 ↪5, 10, 15, 20], we can use the following formula:

m = (x -xmin) / (xmax -xmin)

where x is the value in the dataset, xmin is the minimum value in the dataset
 ↪and xmax is the maximum value in the dataset.

So for this dataset, we have:

xmin = 1 xmax = 20

m = (x - 1) / (20 - 1)

To transform the values to a range of -1 to 1, we can use the following formula:

scaled_value = (2 * m) - 1

So for this dataset, we have:

scaled_value(1) = (2 * ((1 - 1) / (20 - 1))) - 1 = -1 scaled_value(5) = (2 *
 ↪((5 - 1) / (20 - 1))) - 1 = -0.6 scaled_value(10)
= (2 * ((10 - 1) / (20 - 1))) - 1 = -0.2 scaled_value(15) = (2 * ((15 - 1) /
 ↪(20 - 1))) - 1 = 0.2
scaled_value(20) = (2 * ((20 - 1) / (20 - 1))) - 1 = 0.6

Therefore, after performing Min-Max scaling on this dataset, we get [-1, -0.6,
 ↪-0.2, 0.2, 0.6].
```

[ ]:
```

```

[ ]:
```
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 -

[ ]:
```
PCA is a technique used to reduce the dimensionality of a dataset. It works by␣
 ↪identifying the principal components of the data,
which are the directions in which the data varies the most. These principal␣
 ↪components are then used to create a new set of variables
that capture most of the variation in the original data.

The number of principal components to retain depends on the amount of variance␣
 ↪that needs to be explained. A common rule of thumb is
to retain enough principal components to explain at least 80% of the variance␣
 ↪in the data. However, this rule is not set in stone and
can vary depending on the specific dataset and application.

In your case, you have a dataset with five features: height, weight, age,␣
 ↪gender, and blood pressure. The number of principal
components you would choose to retain would depend on how much variance you␣
 ↪want to explain. If you want to explain at least 80% of
the variance in the data, you would need to retain enough principal components␣
 ↪to achieve this goal.

To determine how many principal components you need to retain, you can perform␣
 ↪a PCA analysis on your dataset and look at the scree
plot. The scree plot shows the amount of variance explained by each principal␣
 ↪component. You can then choose the number of principal
components that explain at least 80% of the variance in the data.
```