Data Science & Artificial Intelligence

Warehousing

Data Warehouse Modelling

ONESHOT

Astha Singh Ma'am

Recap of Previous Lecture

- OLAP Technology
- Data Transformation

{ — types
operation

Topics to be covered

- Normalization (Standardization)
- Aggregation
- Discretization
- DataWarehouseModelling

→ Part-2

Data Normalization

Sorting large data set & identify

Data normalization is a technique used in data mining to transform the values of a dataset into a common scale. This is important because many machine learning algorithms are sensitive to the scale of the input features and can produce better results when the data is normalized.

Types of Data Normalization

- Min-Max Normalization
  - *Range should be {0, 1}
- Z Score Normalization
  - {Mean = 0, Standard = 1}
- Decimal Scaling Normalization
  - {it should be ≤ 1}

Types of Data Normalization

1. **Min-Max Normalization**

   Standardizes data within a specific range, typically between 0 and 1.

2. **Z-Score Normalization**

   Transforms data to have a mean of 0 and a standard deviation of 1, enabling easy comparison across variables.

3. **Decimal Scaling Normalization**

   Shifts the decimal point of the values, making the largest value less than or equal to 1.

Min-Max Normalization

Given data

$$V = \frac{X_i - Min}{Max - Min}$$

$x = \{3, 4, 5, 6, 7, 12\}$

Min = 3 (Min)

Max = 12 (Max)

Min = 0

$i = 1, 2, \ldots$ 10 times

Example

x

Data (v)	ND
200 Min	0
300	0.125
400	0.25
600	? 0.5
1000 Max	? 1

Normalized data

$$V = \frac{x - \text{Min}}{\text{Max} - \text{Min}}$$

Min = 200 | Max = 1000

$$V_1 = \frac{x - \text{Min}}{\text{Max} - \text{Min}} = \frac{200 - 200}{1000 - 200} = 0$$

$$V_2 = \frac{x - \text{Min}}{\text{Max} - \text{Min}} \Rightarrow \frac{300 - 200}{1000 - 200} = 0.125$$

$$V_3 = \frac{x - \text{Min}}{\text{Max} - \text{Min}} \Rightarrow \frac{400 - 200}{1000 - 200} = \frac{200}{800} = 0.25$$

$x = 600, 1000$

$$V_4 = \frac{x - \text{Min}}{\text{Max} - \text{Min}} = \frac{600 - 200}{1000 - 200} = \frac{400}{800} = 0.5$$

$$V_5 = \frac{x - \text{Min}}{\text{Max} - \text{Min}} = \frac{1000 - 200}{1000 - 200} = 1$$

Z-Score Normalization

Standardized

Absolute Deviation

Give data in table

* Mean = 0
  Stand = 1

$$Z = \frac{x_i - \mu}{\sigma}$$

μ → Mean

σ → Standard deviation

$$\text{Mean} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

$V = \{x_1, x_2, x_3, x_4, x_5\}$

Standard deviation = 0

Standard deviation $= \sigma$

$$\sigma = \sqrt{\frac{(x_i - \mu)}{n}}$$

μ : Mean

x_i : Given data

Example

$$Z = \frac{x - \mu}{\sigma}$$

$\mu = ?$
$\sigma = ?$

Data (v) (x)	ND — Normalized data
200	−1.06
300	−0.707
400	−0.354
600	0.354
1000	1.88

$$\text{Mean}(\mu) = \frac{200 + 300 + 400 + 600 + 1000}{5}$$

$$\mu = 500$$

$$\text{Stand}(\sigma) = \sqrt{\frac{(x_i - \mu)^2}{n}}$$

$$\Rightarrow \sqrt{\frac{(200-500)^2 + (300-500)^2 + (400-500)^2 + (600-500)^2 + (1000-500)^2}{5}}$$

$$\Rightarrow 282.8$$

$\mu = 500$

$\sigma = 282.8$

$$Z_i = \frac{x_i - \mu}{\sigma}$$

$$Z_1 = \frac{200 - 500}{282.8} = -1.06$$

$$Z_2 = \frac{x_i - \mu}{\sigma} = \frac{300 - 500}{282.8} = -0.707$$

$$\frac{400 - 500}{282.8} = -0.354$$

$Z_3 \Rightarrow$

$Z_4 \Rightarrow$

$$\frac{600 - 500}{282.8} = 0.354$$

$$\zeta = \frac{1000 - 500}{282.8}$$

$$\Rightarrow 1.77$$

Method = 2

* Mean Absolute deviation = A

* $Z = \frac{x_i - M}{A}$

x_i → Mean

A → Absolute deviation

Z, Score $Z = \frac{x - \mu}{\sigma}$ — Mean

Absolute Deviation

$\mu = 500$

Data	ND
200	-1.25
300	0.833
400	-0.417
600	0.417
1800	2.08

$$A = \frac{|x_i - \mu|}{n}$$

$$\frac{|200-500| + |300-500| + |400-500| + |600-500| + |1000-500|}{5}$$

$$A \Rightarrow = 240$$

$$z_i = \frac{x_i - \mu}{A}$$

$x = 200$, $\mu = 500$ $A = 240$

$$z_1 \Rightarrow \frac{200 - 500}{240} = -1.25$$

$z_2 \Rightarrow ? \quad \frac{300 - 500}{240} \Rightarrow -0.833$

$z_3 \Rightarrow ? - \frac{400 - 500}{240} \Rightarrow -0.417$

$z_4 = ? = \frac{600 - 500}{240} \Rightarrow +0.41$

$z_5 = ? - \frac{1000 - 500}{240} \Rightarrow 240 \frac{500}{240} = 2.08$

Decimal Scaling Normalization

find the Value of j

The Smallest integer j such that the Max pt

$\neq \left(\frac{|v_i|}{10^j}\right) \leq 1$

Given data

Example

Data (v)	ND
200	0.2
300	0.3
400	0.4
600	0.6
1000	1

$$\frac{V_i}{10^j} \le 1$$

$$\frac{200}{10^1} \le 1 \times$$

$$\frac{200}{10^2} = \frac{200}{100} \le 1 \checkmark$$

$$\frac{200}{10^3} = \le 1 \checkmark \quad \frac{200}{1000} = 0.2$$

$$\frac{200}{10^5} = \frac{200}{100000} \quad 0.0052$$

$$\frac{V_i}{10^3}$$

$$V_1 = \frac{200}{10^3} = 0.2$$

$$V_2 = \frac{300}{10^3} = 0.3$$

$$X_2 = \frac{400}{10^3} = 0.4$$

$$V_4 = \frac{600}{10^3} = 0.6$$

$$V_5 = \frac{1000}{1000} = 1$$

* Min-Max Normalization

$$V = \frac{x_i - Min}{Max - Min}$$

Given data in the Table

* Z-Score Normalization

$$V = \frac{x_i - \mu}{\sigma}$$

μ — Mean

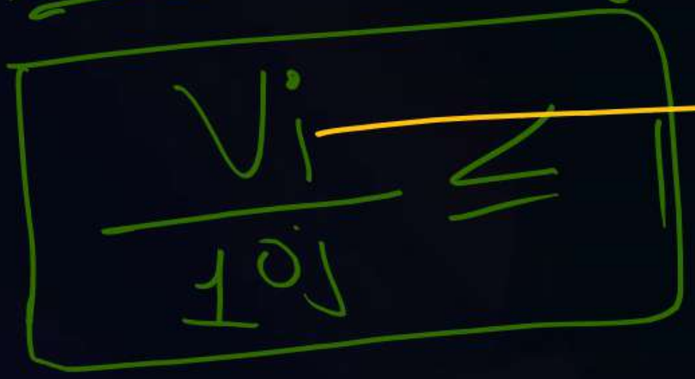
Standardization Method

$$V = \frac{x_i - \mu}{A}$$

μ — Mean

A — Absolute deviation

* Decimal Scaling Normalization



→ Given data in the stock

* first check it should be less than 1 (≤ 1)

Comparing Normalization Techniques

Min-Max Normalization

Simple and intuitive method

Z-Score Normalization

Preserves more information and handles outliers

Decimal Scaling Normalization

Useful in situations where the range of values is unknown

Benefits of Data Normalization in Data Science

1. **Improved Data Accuracy**

   Eliminates discrepancies caused by varying scales, allowing meaningful comparisons.

2. **Enhanced Model Performance**

   Enables models to learn effectively by reducing the impact of outliers and data skew.

3. **Facilitates Data Analysis**

   Enables efficient data exploration and pattern recognition, leading to valuable insights.

Challenges in Data Normalization

1. **Determining Appropriate Normalization Method**

   Selecting the most suitable normalization technique for the specific dataset and context.

2. **Dealing with Outliers during Normalization**

   Addressing extreme values that may affect the normalization process and subsequent analysis.

MCQ

Sample question

Which normalization technique is less sensitive to outliers compared to Min-Max Scaling?

a. Z-score Normalization
b. Robust Scaling
c. Log Transformation
d. Exponential Smoothing

Which of the following normalization techniques scales the data to a specific range, typically [0, 1]?

1. A) Z-Score Normalization
2. B) Min-Max Scaling
3. C) Log Transformation
4. D) Standardization

→ Mean 0 & 1
{0, 1}
1

Which normalization technique involves subtracting the mean and dividing by the standard deviation?

a. Min-Max Scaling
b. Robust Scaling
c. Z-score Normalization
d. Log Transformation

Ans — ?

Z-Score normalization transforms the data to have a mean of: — 0

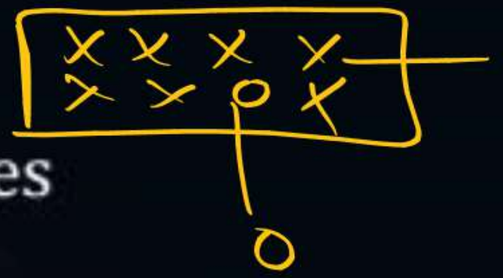
A) 0 — Mean
B) 1 — St
C) Any value
D) The median of the data

Which normalization technique is suitable for data that follows a power-law distribution?

A) Min-Max Scaling
B) Log Transformation
C) Z-Score Normalization
D) Robust Scaling

What is a potential drawback of Min-Max scaling?

A) It is sensitive to outliers
B) It cannot handle missing values
C) It increases data complexity
D) It leads to overfitting in models



Significantly different from rest of the data

Topic-Aggregation

Aggregation in data science refers to the process of combining, summarizing, and analyzing large volumes of data to reveal meaningful patterns and insights. It involves applying mathematical functions to data points to generate aggregated results.

The Purpose and Importance of Aggregation

Aggregation plays a vital role in data science by transforming raw data into valuable insights. It helps to identify trends, anomalies, and correlations, enabling informed decision making, efficient resource allocation, and improved business performance

Methods and Techniques Used in Aggregation

Roll-up

Summarizing data from low-level details to higher-level categories or hierarchies.

Drill-down

Exploring aggregated data at different levels of granularity for detailed analysis.

Grouping

Categorizing data based on specific attributes or criteria for meaningful analysis.

Pivoting

Restructuring data by transforming rows into columns or vice versa to gain different perspectives.

Benefits and Limitations of Aggregation

Benefits

Provides a holistic view of complex data, simplifies decision-making, and uncovers hidden patterns.

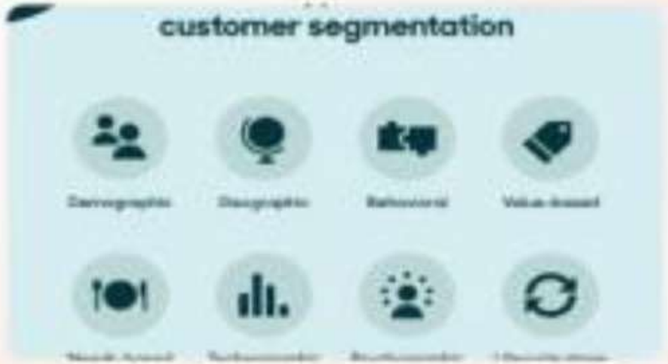
Limitations

May lead to information loss, oversimplification, and incorrect conclusions if not used properly.

Real-World Examples of Aggregation in Data Science

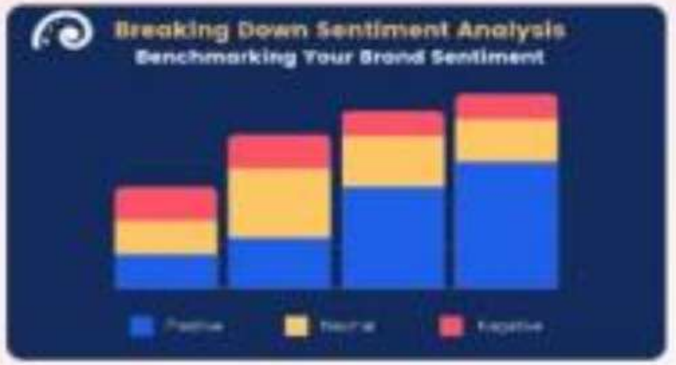
Financial Analysis

Using aggregation to analyze stock market trends, portfolio performance, and risk assessment.



Customer Segmentation

Applying aggregation to group customers based on demographics, behaviors, and preferences.



Social Media Sentiment Sentiment Analysis

Employing aggregation to analyze public opinion and sentiment towards brands, products, or events.

Challenges and Considerations in Implementing Aggregation

1. **Data Quality**

   Inaccurate or incomplete data can impact the validity and reliability of aggregated results.

2. **Scalability**

   Handling large volumes of data efficiently to ensure timely and accurate aggregation.

3. **Privacy and Security**

   Safeguarding sensitive data during the aggregation process to protect individual privacy rights.

4. **Data Bias**

   Awareness of biases that can arise during the aggregation process and mitigating their impact.

MCQ

1. What is the primary purpose of aggregation in data science?
   A) Increasing data complexity
   B) Simplifying data for analysis
   C) Introducing noise to the data
   D) Ignoring missing values

2. Which SQL clause is used for grouping data in aggregation operations?
   1. A) WHERE
   2. B) GROUP BY
   3. C) HAVING
   4. D) ORDER BY

What does the COUNT function in SQL aggregation do?

A) Calculates the average
B) Counts the number of rows
C) Finds the minimum value
D) Concatenates text data

In time series analysis, what does aggregation over a monthly period typically involve?

A) Calculating moving averages
B) Summarizing data for each month
C) Finding the maximum value
D) Ignoring data outliers

Discretization

Data Generalization is the process of summarizing data by replacing relatively low level values with higher level concepts.

Method of Discretization

1. **Equal-width Binning**

   Grouping data into fixed-width intervals to create equally sized bins.

2. **Equal-frequency Binning**

   Dividing data into bins of equal frequency to ensure similar instances are grouped together.

3. **Clustering-based Discretization**

   Utilizing clustering algorithms to group similar data points into discrete values.

4. **Entropy-based Discretization**

   Using information theory to find the optimal cut-off points for data discretization.

Advantages of Discretization

1. **Reduction of Data Complexity**

   Simplifying complex continuous data into discrete categories for easier analysis.

2. **Improved Performance of Analysis Algorithms**

   Enhancing the efficiency and accuracy of certain algorithms designed for discrete data.

3. **Enhanced Interpretability of Results**

   Allowing for clearer insights and easier communication of findings to diverse audiences.

Challenges and Considerations in Discretization

1. Determining Appropriate Number of Bins

Strategically selecting the number of bins to ensure meaningful data granularity.

Handling Outliers and Missing Values

Dealing with extreme values or missing data points during the discretization process.

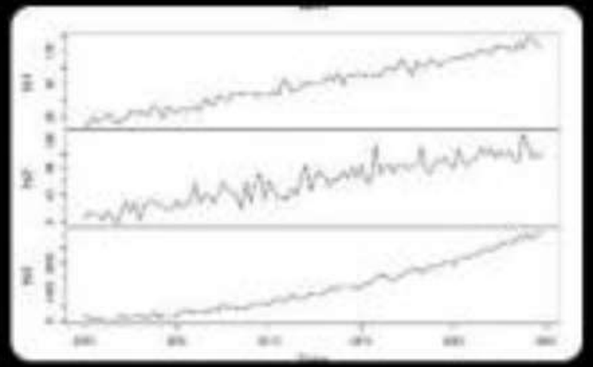
Selecting Suitable Discretization Method

Choosing the most suitable technique based on the characteristics of the dataset and the specific analysis requirements.

Examples of Discretization in Practice

Feature Engineering for Machine Learning

How discretization contributes to effective feature engineering for better predictive models.



Discretization in Time Series Analysis

Unveiling hidden patterns and trends in time-dependent data through discretization techniques.

Discretization in Text Mining and Sentiment Analysis

Unifying semantic information and unraveling sentiments through intelligent text discretization.

Numerical

Calculate the bin width. If you decide to have four group and the age range is 18 to 70 the bin width would be Original data [21,35,28,42,50,18,70,30,40,60]

* Original data = {21, 35, 28, 64, 50, 18, 70, 30, 40, 60}

Min = 21, Min = 64, Max = 70

* Group of bin (n) = 4 (Given)

* $\boxed{\text{Width} = \frac{\text{Max} - \text{Min}}{n}} = \frac{70 - 18}{4} = 13$

Width = 13, Min = 18, Max = 70

B_1 =  ⇒

$Min_2(min + width - 1)$

$Min_1(min + width - 1)$

$18_{,1}(18 + 13 - 1) \Rightarrow 18_{,}(30) = (18-30)$

Data Warehouse Modelling

Dimensional modeling represents data with a cube operation, making more suitable logical data representation with OLAP data management. The perception of Dimensional Modeling was developed by Ralph Kimball and is consist of "fact" and "dimension" tables

Objectives of Dimensional Modeling

The purposes of dimensional modeling are:

To produce database architecture that is easy for end-clients to understand and write queries.

To maximize the efficiency of queries. It achieves these goals by minimizing the number of tables and relationships between them.

Fact

It is a collection of associated data items, consisting of measures and context data. It typically represents business items or business transactions.

Dimensions

It is a collection of data which describe one business dimension. Dimensions decide the contextual background for the facts, and they are the framework over which OLAP is performed.

Measure

It is a numeric attribute of a fact, representing the performance or behavior of the business relative to the dimensions.

Fact Table

Fact tables are used to data facts or measures in the business. Facts are the numeric data elements that are of interest to the company.

Characteristics of the Fact table

The fact table includes numerical values of what we measure. For example, a fact value of 20 might means that 20 widgets have been sold.

Each fact table includes the keys to associated dimension tables. These are known as foreign keys in the fact table.

Dimension Table

Dimension tables establish the context of the facts. Dimensional tables store fields that describe the facts.

Characteristics of the Dimension table

Dimension tables contain the details about the facts. That, as an example, enables the business analysts to understand the data and their reports better.

2 Min Summary

THANKYOU