Unit-4

Data Pre-processing

Outline

- Why to preprocess data?
- Mean, median, mode & range
- Attribute types
- Data preprocessing tasks
  - Data cleaning
  - Data integration
  - Data transformation
  - Data reduction
- Data mining task primitives

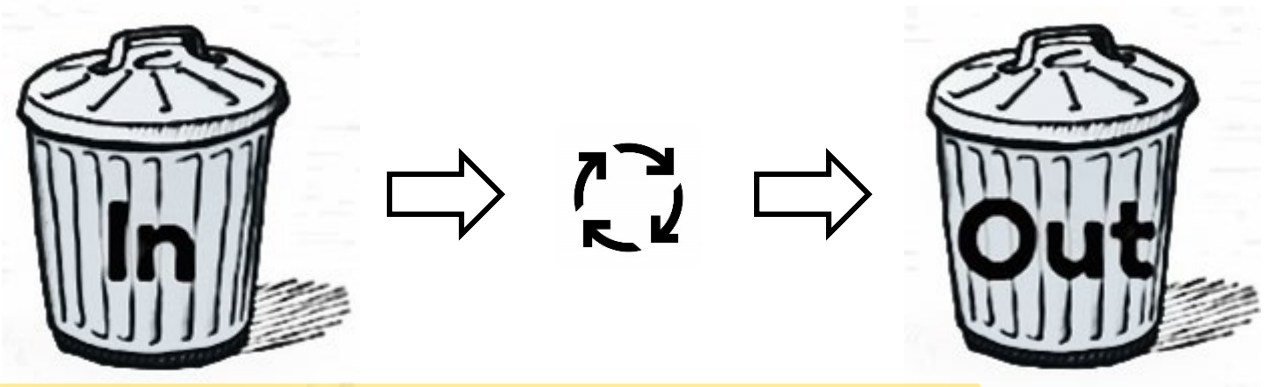
Why to preprocess data?

- Real world data are generally "**dirty**"
  - **Incomplete**: Missing attribute values, lack of certain attributes of interest, or containing only aggregate data.
    - **E.g.** Occupation=" "
  - **Noisy**: Containing errors or outliers.
    - **E.g.** Salary="abcxy"
  - **Inconsistent**: Containing similarity in codes or names.
    - **E.g.** "Gujarat" & "Gujrat" (Common mistakes like **spelling, grammar, articles**)

Why data preprocessing is important?

"No quality data, No quality results"

- It looks like **Garbage In Garbage Out (GIGO)**.



- Quality decisions must be based on **quality data**.
- Duplicate or missing data may cause incorrect or even misleading statistics.
- **Data preparation, cleaning and transformation are the majority task in data mining. (could be as high as 90%).**
- Data preprocessing **prepares** raw data for **further processing**.

Mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x$$

- Mean is the **average** of a dataset.
- To find the mean, calculate the sum of all the data and then divide by the total number of data.
- Example
  - ✓ Find out mean for **12, 15, 11, 11, 7, 13**

First, find the **sum of the data**.

12 + 15 +11 + 11 + 7 + 13 = **69**

Then **divide by the total number of data**.

69 / 6 = **11.5** ⇦ **Mean**

Median

- Median is the **middle number** in a dataset when the data is arranged in numerical order (Sorted Order).

If count is **Odd** then **middle number** is **Median**

If count is **Even** then take **average of middle two numbers** that is **Median**

Median - Odd (Cont..)

- Example
  - ✓ Find out Median for 12, 15, 11, 11, 7, 13, 15

In above example, count of data is **7**. (Odd)

First, arrange the **data** in **ascending order**.

7, 11, 11, 12, 13, 15, 15

Partitioning data into equal halfs

7, 11, 11, **12**, 13, 15, 15

12 ⇦ Median

Median - Even (Cont..)

- Example
  - ✓ Find out median for 12, 15, 11, 11, 7, 13

In above example, count of data is **6**. (Even)

First, arrange the **data** in **ascending order**.

7, 11, 11, 12, 13, 15

Calculate an **average** of the **two numbers** in the **middle**.

7, 11, 11, 12, 13, 15

(11 + 12)/2 = **11.5** ⇦ **Median**

Mode

- The mode is the **number that occurs most often** within a set of numbers.
- Example

1

Find mode.

12, 15, 11, 11, 7, 13

11 ⇦ **Mode** (Unimodal)

2

Find mode.

12, 15, 11, 11, 7, 12, 13

11, 12 ⇦ **Mode** (Bimodal)

Mode (Cont..)

- Example

Find mode.

12, 12 15, 11, 11, 7, 13, 7

7, 11, 12 ⇦ Mode (Trimodal)

Find mode.

12, 15, 11, 10, 7, 14, 13

No Mode

Range

- The range of a set of data is the **difference** between the **largest and the smallest number in the set**.
- Example
  - ✓ Find range for given data 40, 30, 43, 48, 26, 50, 55, 40, 34, 42, 47, 50

First, arrange the **data** in **ascending order**.

26, 30, 34, 40, 40, 42, 43, 47, 48, 50, 50, 55

- In our example **largest number is 55**, and subtract the **smallest number is 26**.

55 – 26 = 29⇦Range

Standard deviation

- The Standard Deviation is a measure of **how spread out any data are**.
- Its symbol is **σ** (the Greek letter sigma).
- $Sample\ variance: (s)^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x - mean)^2$
- Standard Deviation is **Square root of sample variance**.

Standard deviation (Cont..)

- The **Variance** is defined as:

The average of the **squared** differences from the Mean.

To calculate the variance follow these steps:

1. Calculate the mean**, x**.
2. Write a table that subtracts the mean from each observed value.
3. Square each of the differences, add this column.
4. Divide by **n -1** where **n** is the number of items in the **sample**, this is the **variance** (In actual case take n).
5. To get the **standard deviation** we take the **square root** of the variance.

Standard deviation - example

- The owner of the Indian restaurant is interested in how much people spend at the restaurant.
- He examines **10** randomly selected receipts for parties and writes down the following data.

44, 50, 38, 96, 42, 47, 40, 39, 46, 50

1. Find out Mean (1st step)
   - ✓ Mean is **49.2**
2. Write a table that subtracts the mean from each observed value. (2nd step)

Standard deviation – example (Cont..)

Step : 3

X	X – Mean	(X – Mean)^2
44	-5.2	27.04
50	0.8	0.64
38	11.2	125.44
96	46.8	2190.24
42	-7.2	51.84
47	-2.2	4.84
40	-9.2	84.64
39	-10.2	104.04
46	-3.2	10.24
50	0.8	0.64
	Total	**2600.4**

Step : 4

$$= \frac{2600.4}{10 - 1}$$

$\mathbf{S^2 = 288.7 \sim 289}$

Step : 5

$$S = \sqrt{289}$$

$\mathbf{S = 17}$

Standard deviation – example (Cont..)

- Standard deviation can be thought of measuring **how far the data values lie from the mean**, we take the mean and move on standard deviation in either direction.
- The **mean** for this example is **49.2** and the **standard deviation** is **17**.
- Now, 49.2 - 17 = 32.2 and 49.2 + 17 = 66.2
- This means that most of the data probably spend between **32.2** and **66.2**.
- If all data are same then variance & standard deviation is 0 (zero).

Example (Try it)

- Calculate Mean, Median, Mode, Range, Variance & Standard deviation .

  13, 18, 13, 14, 13, 16, 14, 21, 13

- Mean is **15**.
- Median is **14**.
- Mode is **13 & 14 (Bimodal)**.
- Range is **8**.
- Variance is **289.**
- Standard deviation is **17.**

Attribute Types

- An attribute is a **property of the object.**
- It also represents different **features of the object**.
  - **E.g.** Person → Name, Age, Qualification etc.
- Attribute types can be divided into four categories.
  1. Nominal
  2. Ordinal
  3. Interval
  4. Ratio

1) Nominal Attribute

Attribute Types

- Nominal attributes are **named** attributes which can be **separated into discrete (individual) categories** which do not overlap.
- Nominal attributes values also called as **distinct values**.
- Example

What is your gender?
Male
Female
Other

What is your hair color?
Black
Brown
Gray
Blonde
Other

2) Ordinal Attribute

Attribute Types

- Ordinal attribute is the **order of the values,** that's important and significant, but the differences between each one is not really known.
- Example
  - **Rankings** → 1st, 2nd, 3rd
  - **Ratings** → ★ ★ ★ , ★ ★ ★ ★ ★
- We know that a 5 star is better than a 2 star or 3 star, but we don't know and cannot quantify–how much better it is?

3) Interval Attribute

Attribute Types

- Interval attribute comes in the form of a **numerical value** where the **difference** between points is **meaningful**.
- Example
  - **Temperature** → 10°-20°, 30°-50°, 35°-45°
  - **Calendar Dates** → 15^{th} – 22^{nd}, 10^{th} – 30^{th}
- We can not find true zero (absolute) value with interval attributes.

4) Ratio Attribute

Attribute Types

- Ratio attribute is looks **like interval attribute**, but it **must have** a **true zero (absolute)** value.
- It tells us about the order and the exact value between units or data.
- Example
  - **Age Group** → 10-20, 30-50, 35-45 (In years)
  - **Mass** → 20-30 kg, 10-15 kg
- It does have a true zero (absolute) so, it is possible to compute ratios.

Data Preprocessing

- Data have quality if they satisfy the requirements of the intended use.
- There are many factors comprising data quality, including accuracy, completeness, consistency, timeliness, believability, and interpretability.
- The data you wish to analyze by data mining techniques are
  - ✓ incomplete (lacking attribute values or certain attributes of interest, or containing only aggregate data);
  - ✓ inaccurate or noisy (containing errors, or values that deviate from the expected); and
  - ✓ inconsistent (e.g., containing discrepancies in the department codes used to categorize items)

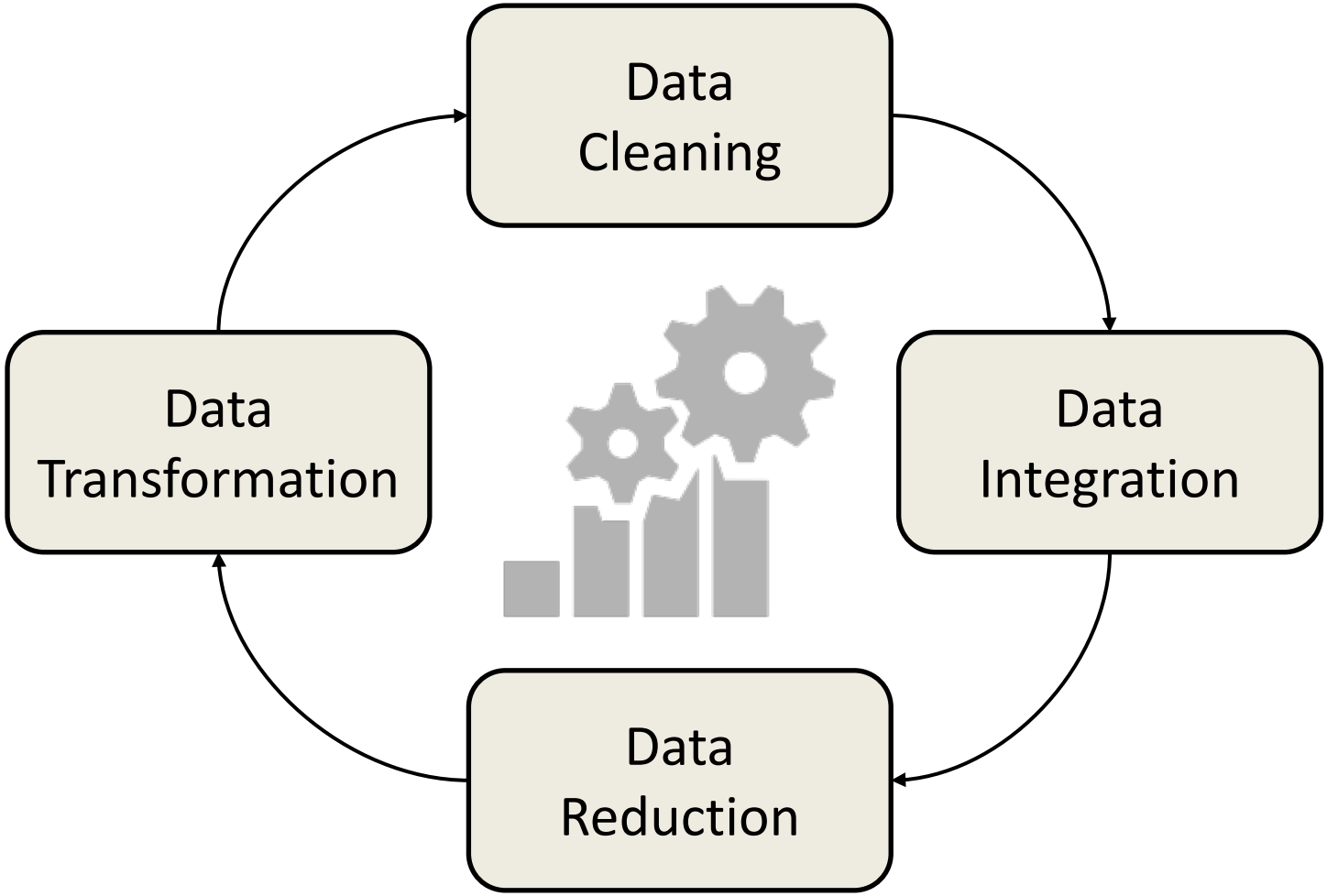
Data Preprocessing

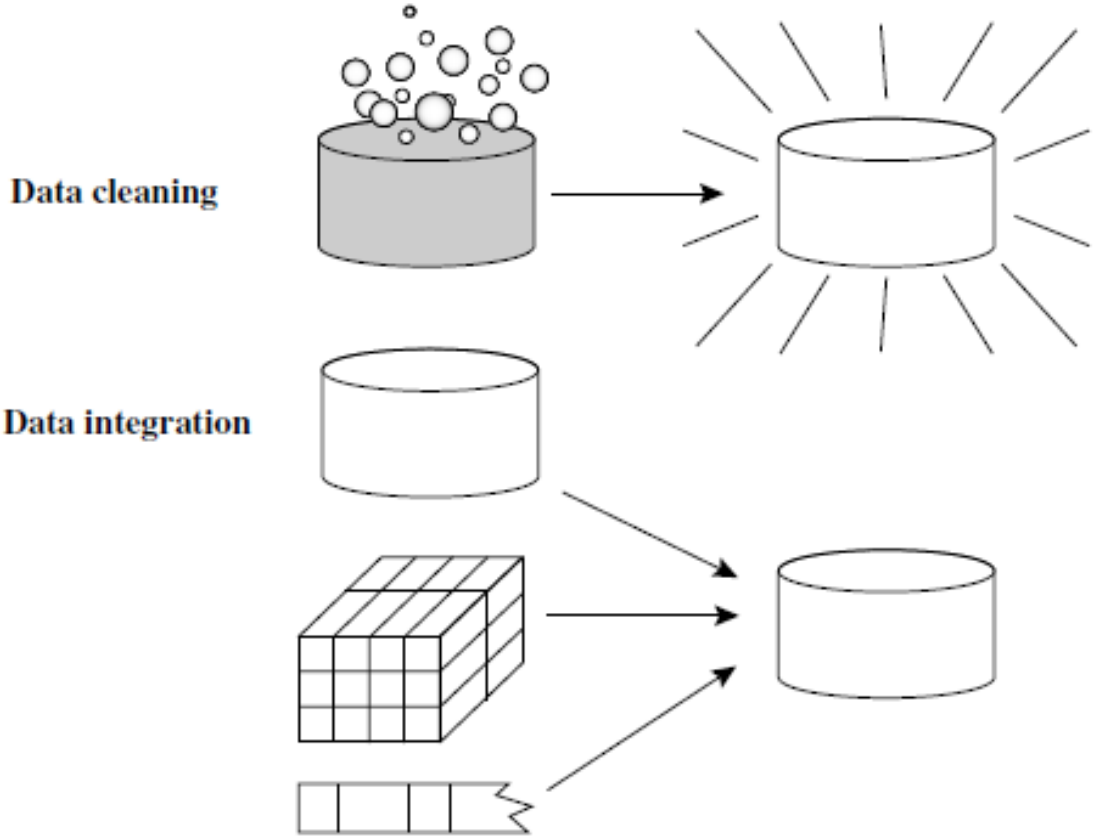
- The elements defining data quality:
  - ✓ accuracy,
  - ✓ completeness,
  - ✓ consistency.
- Inaccurate, incomplete, and inconsistent data are commonplace properties of large real-world databases and data warehouses.

- Reasons for inaccurate data (i.e., having incorrect attribute values):
  - ✓ The data collection instruments used may be faulty.
  - ✓ There may have been human or computer errors occurring at data entry.
  - ✓ Users may purposely submit incorrect data values for mandatory fields when they do not wish to submit personal information (e.g., by choosing the default value "January 1" displayed for birthday). This is known as disguised missing data.
  - ✓ Errors in data transmission can also occur.
  - ✓ There may be technology limitations such as limited buffer size for coordinating synchronized data transfer and consumption.
  - ✓ Incorrect data may also result from inconsistencies in naming conventions or data codes, or inconsistent formats for input fields (e.g., date).
  - ✓ Duplicate tuples also require data cleaning.

- Incomplete data can occur for a number of reasons.
  - ✓ Attributes of interest may not always be available
  - ✓ Data may not be included simply because they were not considered important at the time of entry.
  - ✓ Relevant data may not be recorded due to a misunderstanding or because of equipment malfunctions.

Data Preprocessing Tasks




Data cleaning
Data integration
Data reduction
Attributes
A1 A2 A3 ... A126
Transactions
T1
T2
T3
T4
...
T2000
Attributes
A1 A3 ... A115
Transactions
T1
T4
...
T1456
Data transformation
−2, 32, 100, 59, 48
−0.02, 0.32, 1.00, 0.59, 0.48

1) Data Cleaning

- Real-world data tend to be incomplete, noisy, and inconsistent.
- Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

1) Data Cleaning

1. **Fill in missing values**
    1. Ignore the tuple
    2. Fill missing value manually
    3. Fill in the missing value automatically
    4. Use a global constant to fill in the missing value
2. **Identify outliers and smooth out noisy data**
    1. Binning Method
    2. Clustering
3. **Correct inconsistent data**
4. **Resolve redundancy caused by data integration**

1) Fill missing values

Data Cleaning

- **Ignore the tuple (record/row):**
  - Usually done when **class label is missing**.
  - This method is not very effective, unless the tuple contains several attributes with missing values.
  - It is especially poor when the percentage of missing values per attribute varies considerably
- **Fill missing value manually:**
  - This approach is time consuming and may not be feasible given a large data set with many missing values.
- **Use a global constant to fill in the missing value:**
  - Replace all missing attribute values by the same constant such as a label like “Unknown” or -∞.

1) Fill missing values (Cont..)

Data Cleaning

- **Use a measure of central tendency for the attribute (e.g., the mean or median) to fill in the missing value**:
  - Use mean or median to fill missing values.
- **Use the attribute mean or median for all samples belonging to the same class as the given tuple**
- **Use the most probable value to fill in the missing value**:
  - This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction.

2) Noisy Data

Data Cleaning

- Noise is a random error or variance in a measured variable.

1. **Binning method**
2. **Regression**
3. **Outlier analysis**

1) Binning method

- Binning methods smooth a sorted data value by consulting its "neighborhood," that is, the values around it.
- The sorted values are distributed into a number of "buckets," or bins.
- Because binning methods consult the neighborhood of values, they perform local smoothing

Binning methods for data smoothing

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

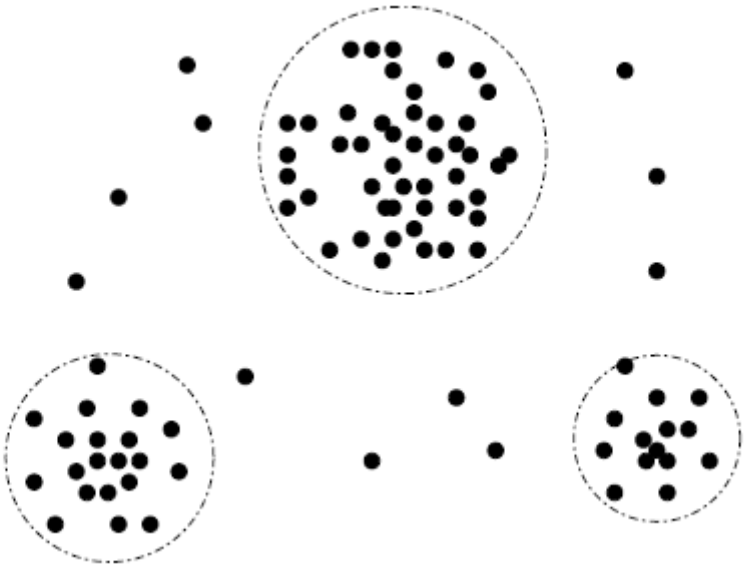
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

2) Regression:

- **Data smoothing can also be done by regression, a technique that conforms data values to a function.**
- **Linear regression involves finding the "best" line to fit two attributes (or variables) so that one attribute can be used to predict the other.**
- **Multiple linear regression is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.**

2) Outlier analysis:

- Outliers may be detected by clustering, for example, where similar values are organized into groups, or “clusters.”
- Values that fall outside of the set of clusters may be considered outliers.



Data Integration

- Data mining often requires data integration—the merging of data from multiple data stores.
- Careful integration can help reduce and avoid redundancies and inconsistencies in the resulting data set.
- This can help improve the accuracy and speed of the subsequent data mining process.

Entity Identification Problem

- Data analysis task will involve data integration, which combines data from multiple sources into a coherent data store, as in data warehousing.
- These sources may include multiple databases, data cubes, or flat files.
- There are a number of issues to consider during data integration.
- **Schema integration and object matching** can be tricky.
- How can equivalent real-world entities from multiple sources be matched up? This is referred to as the **entity identification problem.**

Redundancy and Correlation Analysis

- Redundancy is another important issue in data integration.
- An attribute (such as annual revenue, for instance) may be redundant if it can be "derived" from another attribute or set of attributes.
- Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set.
- Some redundancies can be detected by correlation analysis.
- Given two attributes, such analysis can measure how strongly one attribute implies the other, based on the available data.
- For nominal data, we use the X^2 (chi-square) test.
- For numeric attributes, use the correlation coefficient and covariance, both of which access how one attribute's values vary from those of another.

X^2 Correlation Test for Nominal Data

- For nominal data, a correlation relationship between two attributes, A and B, can be discovered by a X2 (chi-square) test.
- Suppose A has c distinct values, namely a1,a2,….,ac. B has r distinct values, namely b1,b2,…,br.
- The data tuples described by A and B can be shown as a contingency table, with the c values of A making up the columns and the r values of B making up the rows.
- Let (Ai ,Bj) denote the joint event that attribute A takes on value ai and attribute B takes on value bj.

- The X2 value is computed as

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}, \tag{3.1}$$

where o_{ij} is the *observed frequency* (i.e., actual count) of the joint event (A_i, B_j) and e_{ij} is the *expected frequency* of (A_i, B_j), which can be computed as

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n}, \tag{3.2}$$

- where n is the number of data tuples

- Suppose that a group of 1500 people was surveyed.
- The gender of each person was noted. Each person was polled as to whether his or her preferred type of reading material was fiction or nonfiction.
- Thus, we have two attributes, gender and preferred reading. The observed frequency (or count) of each possible joint event is summarized in the contingency table where the numbers in parentheses are the expected frequencies.
- For example, the expected frequency for the cell (male, fiction) is

$$e_{11} = \frac{count(male) \times count(fiction)}{n} = \frac{300 \times 450}{1500} = 90$$

Table 3.1 Example 2.1's 2 × 2 Contingency Table Data

	male	*female*	*Total*
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

Note: Are *gender* and *preferred_reading* correlated?

Using Eq. (3.1) for χ^2 computation, we get

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$
$$= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.$$

- For this 2X2 table, the degrees of freedom are

  (r-1)(c-1) = (2-1)(2-1).

- For 1 degree of freedom, the X2 value needed to reject the hypothesis at the 0.001 significance level is 10.828.
- Since our computed value is above this, we can reject the hypothesis that gender and preferred reading are independent and conclude that the two attributes are (strongly) correlated for the given group of people.

chi square distribution table

	P										
DF	0.995	0.975	0.2	0.1	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	.0004	.00016	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.55	10.828
2	0.01	0.0506	3.219	4.605	5.991	7.378	7.824	9.21	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.86	16.924	18.467
5	0.412	0.831	7.289	9.236	11.07	12.833	13.388	15.086	16.75	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.69	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.18	11.03	13.362	15.507	17.535	18.168	20.09	21.955	24.352	26.124
9	1.735	2.7	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.92	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.3	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697
16	5.142	6.908	20.465	23.542	26.296	28.845	29.633	32	34.267	37.146	39.252
17	5.697	7.564	21.615	24.769	27.587	30.191	30.995	33.409	35.718	38.648	40.79
18	6.265	8.231	22.76	25.989	28.869	31.526	32.346	34.805	37.156	40.136	42.312
19	6.844	8.907	23.9	27.204	30.144	32.852	33.687	36.191	38.582	41.61	43.82
20	7.434	9.591	25.038	28.412	31.41	34.17	35.02	37.566	39.997	43.072	45.315

Covariance of Numeric Data

In probability theory and statistics, correlation and covariance are two similar measures for assessing how much two attributes change together. Consider two numeric attributes A and B, and a set of n observations $\{(a_1, b_1), \ldots, (a_n, b_n)\}$. The mean values of A and B, respectively, are also known as the **expected values** on A and B, that is,

$$E(A) = \bar{A} = \frac{\sum_{i=1}^{n} a_i}{n}$$

and

$$E(B) = \bar{B} = \frac{\sum_{i=1}^{n} b_i}{n}.$$

The **covariance** between A and B is defined as

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{n}. \tag{3.4}$$

If we compare Eq. (3.3) for $r_{A,B}$ (correlation coefficient) with Eq. (3.4) for covariance, we see that

$$r_{A,B} = \frac{Cov(A, B)}{\sigma_A \sigma_B}, \tag{3.5}$$

where σ_A and σ_B are the standard deviations of A and B, respectively. It can also be shown that

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}. \tag{3.6}$$

This equation may simplify calculations.

For two attributes A and B that tend to change together, if A is larger than $\bar{A}$ (the expected value of A), then B is likely to be larger than $\bar{B}$ (the expected value of B). Therefore, the covariance between A and B is *positive*. On the other hand, if one of the attributes tends to be above its expected value when the other attribute is below its expected value, then the covariance of A and B is *negative*.

If A and B are *independent* (i.e., they do not have correlation), then $E(A \cdot B) = E(A) \cdot E(B)$. Therefore, the covariance is $Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B} = E(A) \cdot E(B) - \bar{A}\bar{B} = 0$. However, the converse is not true. Some pairs of random variables (attributes) may have a covariance of 0 but are not independent. Only under some additional assumptions

Table 3.2 Stock Prices for *AllElectronics* and *HighTech*

Time point	AllElectronics	HighTech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

(e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence.

Example 3.2 Covariance analysis of numeric attributes. Consider Table 3.2, which presents a simplified example of stock prices observed at five time points for *AllElectronics* and *HighTech*, a high-tech company. If the stocks are affected by the same industry trends, will their prices rise or fall together?

$$E(AllElectronics) = \frac{6+5+4+3+2}{5} = \frac{20}{5} = \$4$$

and

$$E(HighTech) = \frac{20+10+14+5+5}{5} = \frac{54}{5} = \$10.80.$$

Thus, using Eq. (3.4), we compute

$$Cov(AllElectroncis, HighTech) = \frac{6\times 20+5\times 10+4\times 14+3\times 5+2\times 5}{5} - 4\times 10.80$$
$$= 50.2 - 43.2 = 7.$$

Therefore, given the positive covariance we can say that stock prices for both companies rise together. ■

Variance is a special case of covariance, where the two attributes are identical (i.e., the covariance of an attribute with itself). Variance was discussed in Chapter 2.

Tuple Duplication

- To detect redundancies between attributes, duplication should also be detected at the tuple level (e.g., where there are two or more identical tuples for a given unique data entry case)
- For example, if a purchase order database contains attributes for the purchaser's name and address instead of a key to this information in a purchaser database, discrepancies can occur, such as the same purchaser's name appearing with different addresses within the purchase order database.

Data Value Conflict Detection and Resolution

- Data integration also involves the detection and resolution of data value conflicts.
- For example, for the same real-world entity, attribute values from different sources may differ.
- This may be due to differences in representation, scaling, or encoding.

Data Reduction

- Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data.
- That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results.

Overview of Data Reduction Strategies

- Data reduction strategies include
  - ✓ Dimensionality reduction,
  - ✓ Numerosity reduction, and
  - ✓ Data compression.

Dimensionality reduction

- Dimensionality reduction is the process of reducing the number of random variables or attributes under consideration.
- Dimensionality reduction methods include wavelet transforms and principal components analysis, which transform or project the original data onto a smaller space.
- Attribute subset selection is a method of dimensionality reduction in which irrelevant, weakly relevant, or redundant attributes or dimensions are detected and removed.

Numerosity reduction

- **Numerosity reduction** techniques replace the original data volume by alternative, smaller forms of data representation.
- These techniques may be **parametric or nonparametric**.
- For **parametric methods**, a model is used to estimate the data, so that typically only the data parameters need to be stored, instead of the actual data.
- Regression and log-linear models are examples.
- **Nonparametric methods** for storing reduced representations of the data include histograms, clustering, sampling, and data cube aggregation.

Data Compression

- In data compression, transformations are applied so as to obtain a reduced or “compressed” representation of the original data.
- If the original data can be reconstructed from the compressed data without any information loss, the data reduction is called lossless.
- If, instead, we can reconstruct only an approximation of the original data, then the data reduction is called lossy.

Data Transformation

- Data transformation is the process of **converting data from one form to another form**.
- Data often resides in different locations across the storage and also differs in format.
- Data transformation is necessary to ensure that data from one application or database is understandable to other applications and databases also.

Data Transformation (Cont..)

- Data transformation strategies includes the following:
  1. **Smoothing**
  2. **Attribute construction**
  3. **Aggregation**
  4. **Normalization**
  5. **Discretization**
  6. **Concept hierarchy generation for nominal data**

Data Transformation (Cont..)

1. **Smoothing**
   - It works to **remove noise from the data**.
   - It is a form of data cleaning where users specify transformations to correct data inconsistencies.
   - Such techniques include **binning, regression and clustering**.
2. **Attribute construction**
   - It is referred as **new attributes are constructed** and added from the given set of attributes to help the mining process.
3. **Aggregation**
   - In this, **summary or aggregation operations** are applied to the data.
   - **E.g.** Daily sales data are aggregated at individual source so sales manager can compute monthly and annually total amounts.

Data Transformation (Cont..)

4. Normalization

- Normalization is **scaling technique** or a **mapping technique**.
- With normalization, we can find **new range from an existing range**.
- There are three techniques for normalization.

1. Min-Max Normalization

- This is a simple normalization technique in which we fit given data in a pre-defined boundary, or a pre-defined interval [0,1].

2. Decimal scaling

- In this technique we move the decimal point of values of the attribute.

1) Min-max normalization

- Min max is a technique that helps to **normalizing the data**.
- It will **scale the data between 0 and 1**.
- Example

Age
16
20
30
40

1) Min-max normalization (Cont..)

- Min : Minimum value = 16
- Max : Maximum value = 40
- V = Respective value of attributes. In our example V_1= 16, V_2=20, V_3=30 & V_4=40.
- NewMax = 1
- NewMin = 0

Formula : $V' = \frac{v - Min_A}{Max_A - MinA} (NewMax_A - NewMin_A) + NewMin_A$

1) Min-max normalization (Cont..)

Formula : $V' = \frac{v - Min_A}{Max_A - MinA}(NewMax_A - NewMin_A) + NewMin_A$

For Age 16 :

MinMax (v') = (16 – 16)/(40-16) * (1 – 0) + 0
= 0 / 24 * 1
= **0**

For Age 20 :

MinMax (v') = (20 – 16)/(40-16) * (1 – 0) + 0
= 4 / 24 * 1
= **0.16**

1) Min-max normalization (Cont..)

For Age 30 :

MinMax (v') = (30 – 16)/(40-16) * (1 – 0) + 0
= 14 / 24 * 1
= **0.58**

For Age 40 :

MinMax (v') = (40 – 16)/(40-16) * (1 – 0) + 0
= 24 / 24 * 1
= **1**

Age	After Min-max normalization
16	0
20	0.16
30	0.58
40	1

In **z-score normalization** (or *zero-mean normalization*), the values for an attribute, A, are normalized based on the mean (i.e., average) and standard deviation of A. A value, v_i, of A is normalized to v'_i by computing

$$v'_i = \frac{v_i - \bar{A}}{\sigma_A}, \tag{3.9}$$

where $\bar{A}$ and σ_A are the mean and standard deviation, respectively, of attribute A. The mean and standard deviation were discussed in Section 2.2, where $\bar{A} = \frac{1}{n}(v_1 + v_2 + \cdots + v_n)$ and σ_A is computed as the square root of the variance of A (see Eq. (2.6)). This method of normalization is useful when the actual minimum and maximum of attribute A are unknown, or when there are outliers that dominate the min-max normalization.

Example 3.5 **z-score normalization.** Suppose that the mean and standard deviation of the values for the attribute *income* are \$54,000 and \$16,000, respectively. With z-score normalization, a value of \$73,600 for *income* is transformed to $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$. ■

A variation of this z-score normalization replaces the standard deviation of Eq. (3.9) by the *mean absolute deviation* of A. The *mean absolute deviation* of A, denoted s_A, is

$$s_A = \frac{1}{n}(|v_1 - \bar{A}| + |v_2 - \bar{A}| + \cdots + |v_n - \bar{A}|). \tag{3.10}$$

2) Decimal scaling

- In this technique we move the decimal point of values of the attribute.
- This movement of decimal points totally depends on the **maximum value among all values** in the attribute.
- Value V of attribute A can be normalized by the following formula

Normalized value of attribute = $\mathbf{(v^i / 10^j)}$

Decimal scaling - Example

CGPA	Formula	After Decimal Scaling
2	2 / 10	**0.2**
3	3 / 10	**0.3**

- We will check maximum value among our attribute CGPA.
- Maximum value is 3 so, we can convert it into decimal by dividing with 10. why 10?
- We will count total digits in our maximum value and then put 1.
- After 1 we can put zeros equal to the length of maximum value.
- Here 3 is maximum value and total digits in this value is only 1 so, we will put one zero after 1.

Decimal scaling (Try it!)

Bonus	Formula	After Decimal Scaling
400	400/1000	**0.4**
310	310/1000	**0.31**

Salary	Formula	After Decimal Scaling
40,000	40000/100000	**0.4**
31,000	31000/100000	**0.31**

Data Transformation (Cont..)

5. Discretization

- Discretization techniques can be categorized based on **how the separation is performed**, such as whether it uses class information or which direction it proceeds (top-down or bottom-up).
- The raw values of a numeric attribute (e.g. age) are replaced by interval labels (e.g. 0-10, 11-20 etc.) or conceptual labels (e.g. youth, adult, senior).

6. Concept hierarchy generation for nominal data

- In this, attributes such as address can be **generalized to higher-level concepts**, like street or city or state or country.
- Many hierarchies for nominal attributes are implicit within the database schema.
- **E.g.** city, country or state table in RDBMS.

	Feature Selection	Feature Extraction
1.	Selects a subset of relevant features from the original set of features.	Extracts a new set of features that are more informative and compact.
2.	Reduces the dimensionality of the feature space and simplifies the model.	Captures the essential information from the original features and represents it in a lower-dimensional feature space.
3.	Can be categorized into filter, wrapper, and embedded methods.	Can be categorized into linear and nonlinear methods.
4.	Requires domain knowledge and feature engineering.	Can be applied to raw data without feature engineering.
5.	Can improve the model's interpretability and reduce overfitting.	Can improve the model performance and handle nonlinear relationships.
6.	May lose some information and introduce bias if the wrong features are selected.	May introduce some noise and redundancy if the extracted features are not informative.

Data mining task primitives (Cont..)

- The data mining task primitives includes the following:
  - Task-relevant data
  - Kind of knowledge to be mined
  - Background knowledge
  - Interestingness measurement
  - Presentation for visualizing the discovered patterns