Getting To Know Your Data
Road Map

1. Data Objects and Attribute Types
2. Descriptive Data Summarization
3. Measuring Data Similarity and Dissimilarity
Data Objects and Attribute Types

- Types of data sets
- Data objects
- Attributes and their types
Types of Data Sets

- **Record**
  - Relational records
  - Data matrix, e.g., numerical matrix, cross tabulations.
  - Document data: text documents: term-frequency vector
  - Transaction data

### Relational records

<table>
<thead>
<tr>
<th>Login</th>
<th>First name</th>
<th>Last name</th>
</tr>
</thead>
<tbody>
<tr>
<td>koala</td>
<td>John</td>
<td>Clemens</td>
</tr>
<tr>
<td>lion</td>
<td>Mary</td>
<td>Stevens</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Login</th>
<th>phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>koala</td>
<td>039689852639</td>
</tr>
</tbody>
</table>

### Transactional data

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bred, Cake, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Beer, Bred</td>
</tr>
</tbody>
</table>

### Cross tabulation

<table>
<thead>
<tr>
<th>Books</th>
<th>Multimedia devices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big spenders</td>
<td>30%</td>
</tr>
<tr>
<td>Budget spenders</td>
<td>60%</td>
</tr>
<tr>
<td>Very Tight spenders</td>
<td>10%</td>
</tr>
</tbody>
</table>
Types of Data Sets

- **Graph and Network**
  - World Wide Web
  - Social or information networks
  - Molecular structures networks
Types of Data Sets

- **Ordered**
  - Videos
  - Temporal data
  - Sequential data
  - Genetic sequence data

**Video:** sequence of images

**Temporal data:** Time-series

**Monthly Value of Building Approvals**

**Generic Sequence:** DNA-code

**Transactional sequence**

Computer -> Web cam -> USB key
Types of Data Sets

- **Spatial, image and multimedia**
  - Spatial data
  - Image data
  - Video data
  - Audio Data

- **Images**
- **Spatial data: maps**
- **Videos**
- **Audios**
Data Objects and Attributes

- Datasets are made up of data objects.
- A **data object** (or **sample**, **example**, **instance**, **data point**, **tuple**) represents an entity.

**Examples**
- **Sales database**: customers, store items, sales
- **Medical database**: patients, treatments
- **University database**: students, professors, courses

- Data objects are described by **attributes** (or **dimension**, **feature**, **variable**).
- Database rows -> data objects; columns -> attributes.

<table>
<thead>
<tr>
<th>Patient_ID</th>
<th>Age</th>
<th>Height</th>
<th>Weight</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>1569</td>
<td>30</td>
<td>1.76m</td>
<td>70 kg</td>
<td>male</td>
</tr>
<tr>
<td>2596</td>
<td>26</td>
<td>1.65m</td>
<td>58kg</td>
<td>female</td>
</tr>
</tbody>
</table>
Attribute Types

- **Nominal** categories, states, or "names of things"
  - Hair_color = \{black, brown, blond, red, grey, white\}
  - marital status, occupation, ID numbers, zip codes

- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., having cancer)

- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - Size = \{small, medium, large\}, grades, army rankings
Attributes Types

- **Numeric**: quantity (integer or real-valued)
  - **Interval-Scaled**
    - Measured on a scale of equal-sized units
    - Values have order
      - E.g., temperature in °C or °F, calendar dates
      - No true zero-point (we can add and subtract degrees -100° is 10° warmer than 90°-, we cannot multiply values or create ratios -100° is not twice as warm as 50°-).
  - **Ratio-Scaled**
    - Inherent zero-point
    - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K is twice as high as 5 K)
      - E.g., temperature in Kelvin, length, counts, monetary quantities
      - A 6-foot person is **20% taller** than a 5-foot person.
      - A baseball game lasting **3 hours** is **50% longer** than a game lasting 2 hours.
Discrete vs. Continuous Attributes

- **Discrete Attribute**
  - Has only a finite or countable infinite set of values
  - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes

- **Continuous Attribute**
  - Has real numbers as attribute values
  - E.g., temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables (float, double, long double)
What is the type of an attribute that describes the height of a person in centimeters?
- Nominal
- Ordinal
- Interval-scaled
- Ratio-scaled

In Olympic games, three types of medals are awarded: bronze, silver, or gold. To describe these medals, which type of attributes should be used?
- Nominal
- Ordinal
- Interval-scaled
- Ratio-scaled
Road Map

1. Data Objects and Attribute Types
2. Descriptive Data Summarization
3. Measuring Data Similarity and Dissimilarity
Descriptive Data Summarization

**Motivation**
- For data preprocessing, it is essential to have an overall picture of your data
- Data summarization techniques can be used to
  - Define the typical properties of the data
  - Highlight which data should be treated as noise or outliers

**Data properties**
- Centrality: use measures such as the median
- Variance: use measures such as the quantiles

**From the data mining point of view it is important to**
- Examine how these measures are computed efficiently
- Introduce the notions of distributive measure, algebraic measure and holistic measure
Measuring the Central Tendency

- **Mean (algebraic measure)**
  
  Note: $n$ is sample size
  
  - A **distributive** measure can be computed by partitioning the data into smaller subsets (e.g., **sum**, and **count**)
  
  - An **algebraic** measure can be computed by applying an algebraic function to one or more distributive measures (e.g., mean=sum/count)

- **Sometimes each value $x_i$ is weighted**
  
  - Weighted arithmetic mean

- **Problem**
  
  - The mean measure is sensitive to extreme (e.g., outlier) values
  
  - What to do?
  
  - Trimmed mean: chopping extreme values
Measuring the Central Tendency

- **Median (holistic measure)**
  - Middle value if odd number of values, or average of the middle two values otherwise
  - A holistic measure must be computed on the entire dataset
  - Holistic measures are much more expensive to compute than distributive measures
  - Can be estimated by interpolation (for grouped data):
Measuring the Central Tendency

- **Mode**
  - Value that occurs most frequently in the data
  - It is possible that several different values have the greatest frequency: Unimodal, bimodal, trimodal, multimodal
  - If each data value occurs only once then there is no mode
  - Empirical formula:
    \[ mean - mode = 3 \times (mean - median) \]

- **Midrange**
  - Can also be used to assess the central tendency
  - It is the average of the smallest and the largest value of the set
  - It is an algebraic measure that is easy to compute
Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data

Symmetric data

Positively skewed data

Negatively skewed data
Quiz

- Give an example of something having a positively skewed distribution

  - Income is a good example of a positively skewed variable -- there will be a few people with extremely high incomes, but most people will have incomes bunched together below the mean.

- Give an example of something having a bimodal distribution

  - Bimodal distribution has some kind of underlying binary variable that will result in a separate mean for each value of this variable. One example can be human weight – the gender is binary and is a statistically significant indicator of how heavy a person is.
Measuring the Dispersion of Data

- The degree in which data tend to spread is called the dispersion, or variance of the data.

- The most common measures for data dispersion are range, the five-number summary (based on quartiles), the inter-quartile range, and standard deviation.

Range
- The distance between the largest and the smallest values.

Kth percentile
- Value \( x_i \) having the property that \( k \% \) of the data lies at or below \( x_i \).
- The median is 50th percentile.
- The most popular percentiles other than the median are Quartiles \( Q1 \) (25th percentile), \( Q3 \) (75th percentile).
- Quartiles + median give some indication of the center, spread, and the shape of a distribution.
Measuring the Dispersion of Data

- **Inter-quartile range**
  - Distance between the first and the third quartiles $\text{IQR} = Q_3 - Q_1$
  - A simple measure of spread that gives the range covered by the middle half of the data
  - **Outlier**: usually, a value falling at least $1.5 \times \text{IQR}$ above the third quartile or below the first quartile

- **Five number summary**
  - Provide in addition information about the endpoints (e.g., tails)
  - $\text{min, } Q_1, \text{ median, } Q_3, \text{ max}$
  - E.g., min = $Q_1 - 1.5 \times \text{IQR}$, max = $Q_3 + 1.5 \times \text{IQR}$
  - Represented by a Boxplot
Measuring the Dispersion of Data

- **Variance and standard deviation**
  - **Variance**: (algebraic, scalable computation)
    \[ \sigma^2 = \frac{1}{N} \sum_{i=1}^{n} (x_i - \mu)^2 \]
  - **Standard deviation** \( \sigma \) is the square root of variance \( \sigma^2 \)

- **Basic properties of the standard deviation**
  - \( \sigma \) measures spread about the mean and should be used only when the mean is chosen as the measure of the center
  - \( \sigma = 0 \) only when there is no spread, that is, when all observations have the same value. Otherwise \( \sigma > 0 \)

- Variance and standard deviation are algebraic measures. Thus, their computation is scalable in large databases.
Graphic Displays

- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis represent frequencies
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane
Boxplot

- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum

- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - **Whiskers**: two lines outside the box extended to Minimum and Maximum
  - **Outliers**: points beyond a specified outlier threshold, plotted individually
Histogram Analysis

- **Histogram**: summarizes the distribution of a given attribute
- Partition the data distribution into disjoint subsets, or buckets
- If the attribute is **nominal** → bar chart
- If the attribute is **numeric** → histogram

![Price ($) vs. # items sold graph]

- Price $:
  - 40-59
  - 60-79
  - 80-99
  - 100-119
  - 120-139
- # items sold:
  - 0
  - 1000
  - 2000
  - 3000
  - 4000
  - 5000
Histograms Often Tell More than Boxplots

- The two histograms shown in the left may have the same boxplot representation.
  - The **same values** for: min, Q1, median, Q3, max, **But they have rather different data distributions**
Scatter plot

- Provides a first look at bivariate data to see clusters of points, outliers, etc.
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane.
The left half fragment is positively correlated

The right half is negatively correlated
Uncorrelated Data
1. Data Objects and Attribute Types
2. Descriptive Data Summarization
3. Measuring Data Similarity and Dissimilarity
**Data Similarity and Dissimilarity**

- **Similarity**
  - Numerical measure of how alike two data objects are
  - Value is higher when objects are more alike
  - Often falls in the range \([0,1]\)

- **Dissimilarity** (e.g., distance)
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies

- **Proximity** refers to a similarity or dissimilarity
Data Matrix and Dissimilarity Matrix

- **Data matrix**
  - n data points with p dimensions
  - Two modes: rows and columns represent different entities

\[
\begin{bmatrix}
  x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\
  \vdots & \ddots & \vdots & \ddots & \vdots \\
  x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\
  \vdots & \ddots & \vdots & \ddots & \vdots \\
  x_{n1} & \cdots & x_{nf} & \cdots & x_{np}
\end{bmatrix}
\]

- **Dissimilarity matrix**
  - n data points, but registers only the distance
  - A triangular matrix
  - Single mode: row and columns represent the same entity

\[
\begin{bmatrix}
  0 \\
  d(2,1) & 0 \\
  d(3,1) & d(3,2) & 0 \\
  \vdots & \vdots & \vdots \\
  d(n,1) & d(n,2) & \cdots & \cdots & 0
\end{bmatrix}
\]
Nominal Attributes

- Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)

- **Method 1**: Simple matching
  - \( m \): # of matches, \( p \): total # of variables

\[
d(i, j) = \frac{p - m}{p}
\]

- **Method 2**: Use a large number of binary attributes
  - creating a new binary attribute for each of the \( M \) nominal states
Binary Attributes

- A contingency table for binary data

<table>
<thead>
<tr>
<th>Object i</th>
<th>Object j</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>q</td>
<td>r</td>
</tr>
<tr>
<td>0</td>
<td>s</td>
</tr>
<tr>
<td>sum</td>
<td>q+s</td>
</tr>
</tbody>
</table>

- Distance measure for symmetric binary variables

\[ d(i, j) = \frac{r+s}{q+r+s+t} \]

- Distance measure for asymmetric binary variables

\[ d(i, j) = \frac{r+s}{q+r+s} \]

- Jaccard coefficient (\textit{similarity} measure for asymmetric binary variables)

\[ sim(i, j) = \frac{q}{q+r+s} = 1 - d(i, j) \]
Numeric Attributes

The measurement unit used for interval-scale attributes can have an effect on the similarity
- E.g., kilograms vs. pounds for weight

Need of standardizing the data
- Convert the original measurements to unit-less variables
- For measurements of each variable $f$:
  - Calculate the mean absolute deviation, $s_f$

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|)$$

- Calculate the standardized measurement, or z-score

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

- Using the mean absolute deviation reduces the effect of outliers
- Outliers remain detectable (non squared deviation)
Distance on Numeric Data

- **Minkowski distance**: A popular distance measure

\[ d(i, j) = h \sqrt{(|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \ldots + |x_{ip} - x_{jp}|^h)} \]

where \( i = (x_{i1}, x_{i2}, \ldots, x_{ip}) \) and \( j = (x_{j1}, x_{j2}, \ldots, x_{jp}) \) are two \( p \)-dimensional data objects, and \( h \) is the order

- **Properties**
  - \( d(i, j) > 0 \) if \( i \neq j \), and \( d(i, i) = 0 \) (Positive definiteness)
  - \( d(i, j) = d(j, i) \) (Symmetry)
  - \( d(i, j) \leq d(i, k) + d(k, j) \) (Triangle Inequality)

- A distance that satisfies these properties is a **metric**
Special Cases of Minkowski Distance

- **$h = 1$:** Manhattan (city block, $L_1$ norm) distance
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors
  
  \[
  d(i, j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_p} - x_{j_p}|
  \]

- **$h = 2$:** ($L_2$ norm) Euclidean distance
  
  \[
  d(i, j) = \sqrt{|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2}
  \]
Example: Minkowski Distance

Data Objects

<table>
<thead>
<tr>
<th>point</th>
<th>attribute 1</th>
<th>attribute 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>x2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>x3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>x4</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Dissimilarity Matrices

**Manhattan (L₁)**

<table>
<thead>
<tr>
<th></th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x2</td>
<td>5</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x3</td>
<td>3</td>
<td>6</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>x4</td>
<td>6</td>
<td>1</td>
<td>7</td>
<td>0</td>
</tr>
</tbody>
</table>

**Euclidean (L₂)**

<table>
<thead>
<tr>
<th></th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x2</td>
<td>3.61</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x3</td>
<td>2.24</td>
<td>5.1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>x4</td>
<td>4.24</td>
<td>1</td>
<td>5.39</td>
<td>0</td>
</tr>
</tbody>
</table>
An ordinal variable can be discrete or continuous.

Order is important, e.g., rank. $r_{if} \in \{1, \ldots, M_f\}$

Can be treated like interval-scaled:
- replace $x_{if}$ by their rank
- map the range of each variable onto $[0, 1]$ by replacing $i$-th object in the $f$-th variable by
  
  $$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

compute the dissimilarity using methods for interval-scaled variables.
A document can be represented by thousands of attributes, each recording the frequency of a particular word (such as keywords) or phrase in the document.

<table>
<thead>
<tr>
<th></th>
<th>team</th>
<th>coach</th>
<th>baseball</th>
<th>soccer</th>
<th>penalty</th>
<th>score</th>
<th>win</th>
<th>loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc1</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Doc2</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Doc3</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Doc4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

**Other vector objects**: gene features in micro-arrays, ...

**Cosine measure**: If $d1$ and $d2$ are two vectors (e.g., term-frequency vectors), then

\[
\cos(d_1, d_2) = \frac{(d_1 \cdot d_2)}{|d_1||d_2|}
\]

where $\cdot$ indicates **vector dot product**, $|$ $d$ $|$: **length of vector** $d$
Cosine Similarity

**Example:** Find the *similarity* between documents 1 and 2

\[ d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0) \]
\[ d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1) \]

**Compute** \( d_1 \cdot d_2 \)

\[ d_1 \cdot d_2 = 5 \times 3 + 0 \times 0 + 3 \times 2 + 0 \times 0 + 2 \times 1 + 0 \times 1 + 0 \times 1 + 2 \times 1 + 0 \times 0 + 0 \times 1 = 25 \]

**Compute** \( ||d_1|| \) \( ||d_2|| \)

\[ ||d_1|| = \sqrt{5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2} = \sqrt{42} \approx 6.481 \]

\[ ||d_2|| = \sqrt{3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2} = \sqrt{17} \approx 4.12 \]

\[ \cos(d_1, d_2) = \frac{d_1 \cdot d_2}{||d_1|| \cdot ||d_2||} = \frac{25}{6.481 \times 4.12} \approx 0.94 \]
Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled

Many types of data sets, e.g., numerical, text, graph, Web, image.

Gain insight into the data by:
- Basic statistical data description: central tendency, dispersion, graphical displays
- Data visualization: map data onto graphical primitives
- Measure data similarity

Above steps are the beginning of data preprocessing.

Many methods have been developed but still an active area of research.