Chapter 2: Getting to Know Your Data

- Data Objects and Attribute Types
- Measuring Data Similarity and Dissimilarity
Types of Data Sets

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

- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures

- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data

- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data:
  - Video data:

<table>
<thead>
<tr>
<th>Name</th>
<th>FName</th>
<th>City</th>
<th>Age</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smith</td>
<td>John</td>
<td>3</td>
<td>35</td>
<td>$280</td>
</tr>
<tr>
<td>Doe</td>
<td>Jane</td>
<td>1</td>
<td>28</td>
<td>$325</td>
</tr>
<tr>
<td>Brown</td>
<td>Scott</td>
<td>3</td>
<td>41</td>
<td>$265</td>
</tr>
<tr>
<td>Howard</td>
<td>Shemp</td>
<td>4</td>
<td>48</td>
<td>$359</td>
</tr>
<tr>
<td>Taylor</td>
<td>Tom</td>
<td>2</td>
<td>22</td>
<td>$250</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Coke, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Beer, Bread</td>
</tr>
<tr>
<td>3</td>
<td>Beer, Coke, Diaper, Milk</td>
</tr>
<tr>
<td>4</td>
<td>Beer, Bread, Diaper, Milk</td>
</tr>
<tr>
<td>5</td>
<td>Coke, Diaper, Milk</td>
</tr>
</tbody>
</table>
Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples*, *examples*, *instances*, *data points*, *objects*, *tuples*.
- Data objects are described by **attributes**.
- Database rows -> data objects; columns -> attributes.
Attributes

- Attribute (dimension, feature, variable): a data field, representing a characteristic or feature of a data object.
  - *E.g.,* `customer_ID, name, address`

- Attribute types:
  - Non-numeric: symbolic, non-quantitative
    - Nominal (categorical)
      - Binary
    - Ordinal
  - Numeric: quantitative, a measurable quantity, integer or real
    - Interval-scaled
    - Ratio-scaled
Non-numeric Attribute Types

- **Nominal (categorical):** categories, states, or “names of things”
  - *Hair_color* = \{auburn, black, blond, brown, grey, red, white\}
  - marital status, occupation, ID numbers, zip codes

- **Binary**
  - Nominal attribute with only 2 states
  - **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., HIV positive)

- **Ordinal**
  - Values have a meaningful order (ranking) but difference between successive values is unknown.
  - *Size* = \{small, medium, large\}, letter grades, army rankings
Numeric Attribute Types

- **Interval-scaled**: difference between two values is meaningful
  - Measured on a scale of *equal-sized units*
    - E.g., temperature in C° or F°. pH, calendar dates
    - No true zero
      - neither 0C° nor 0F° indicates no heat
      - without zero, we cannot talk of one temperature value as being a multiple of another. We cannot say 10C° is twice as warm as 5C°

- **Ratio-scaled**: has all the properties of an interval variable, and also has a clear definition of zero (that means none)
  - e.g., temperature in Kelvin (0 kelvin does mean no heat), length, weight, counts, monetary quantities
  - We can speak of a value as being a multiple (or ratio) of another
    - 10K° is twice as high as 5K°
Discrete vs. Continuous Attributes

- Another way to categorize data types
  - **Discrete**
    - Has only a countable (finite or countably infinite) set of values
      - E.g., zip codes, 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**
    - Has real numbers (which are uncountable) 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
Chapter 2: Getting to Know Your Data

- Data Objects and Attribute Types
- Measuring Data Similarity and Dissimilarity
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
Data Matrix and Dissimilarity Matrix

- **Data matrix**
  - n x p, object-by-variable
  - n data points with p dimensions
  - Two modes – stores both objects and attributes

- **Dissimilarity matrix**
  - n x n, object-by-object
  - n data points, but registers only the distance
  - A triangular matrix
  - Single mode as it only stores dissimilarity values

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

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

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

- **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 nominal state
Binary Attributes

- Contingency table for binary data

- Distance for symmetric binary variables:
  - Similarity?

- Distance for asymmetric binary variables:
  - Jaccard coefficient (*similarity* measure for *asymmetric* binary variables):
### Binary Attributes: example

<table>
<thead>
<tr>
<th>Name</th>
<th>Fever</th>
<th>Cough</th>
<th>Test-1</th>
<th>Test-2</th>
<th>Test-3</th>
<th>Test-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jack</td>
<td>Y</td>
<td>N</td>
<td>P</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Mary</td>
<td>Y</td>
<td>N</td>
<td>P</td>
<td>N</td>
<td>P</td>
<td>N</td>
</tr>
<tr>
<td>Jim</td>
<td>Y</td>
<td>P</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

- All attributes are asymmetric binary
- Let the values Y and P be 1, and the value N be 0

\[
d(jack, mary) = \frac{0 + 1}{2 + 0 + 1} = 0.33
\]

\[
d(jack, jim) = \frac{1 + 1}{1 + 1 + 1} = 0.67
\]

\[
d(jim, mary) = \frac{1 + 2}{1 + 1 + 2} = 0.75
\]
Numeric Attributes: Standardization

- **Z-score:**
  \[ Z = \frac{x - \mu}{\sigma} \]
  - \( x \): raw score to be standardized, \( \mu \): mean of the population, \( \sigma \): standard deviation
  - the distance between the raw score and the population mean in units of the standard deviation
  - negative when the raw score is below the mean, positive when above

- An alternative way: Calculate the mean absolute deviation
  \[ s_f = \frac{1}{n} (|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|) \]
  where
  \[ m_f = \frac{1}{n} (x_{1f} + x_{2f} + \ldots + x_{nf}) \]

- Standardized measure (z-score):
  \[ Z_{if} = \frac{x_{if} - m_f}{s_f} \]

- Using mean absolute deviation is more robust than using standard deviation
Numeric Attributes: Minkowski Distance

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

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

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 (the distance so defined is also called L-\( h \) norm)

- **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
  \[
  d(i, j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}
  \]

- $h \to \infty$. "supremum" (L$_{\text{max}}$ norm, L$_\infty$ norm) distance.
  - This is the maximum difference between any component (attribute) of the vectors
    \[
    d(i, j) = \lim_{h \to \infty} \left( \sum_{f=1}^{p} |x_{i_f} - x_{j_f}|^h \right)^{1/h} = \max_{f} |x_{i_f} - x_{j_f}|
    \]
# Minkowski Distance: Example

## Point Attribute Example

<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>

## Manhattan (L₁)

<table>
<thead>
<tr>
<th>L</th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>x2</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>x3</td>
<td>3</td>
<td>6</td>
<td>0</td>
<td>0</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>L₂</th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>x2</td>
<td>3.61</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>x3</td>
<td>2.24</td>
<td>5.1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>x4</td>
<td>4.24</td>
<td>1</td>
<td>5.39</td>
<td>0</td>
</tr>
</tbody>
</table>

## Supremum (Lₘᵢₙ)
Ordinal Variables

- Can be treated like interval-scaled
  - replace $x_{if}$ by their rank $r_{if} \in \{1, \ldots, M_f\}$
  - 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, e.g., Euclidean distance
Ordinal Variables: Example

- Consider the data in the adjacent table:
- Here, the attribute Test has three states: fair, good and excellent, so $M_f = 3$
- For step 1, the four attribute values are assigned the ranks 3,1,2 and 3 respectively.
- Step 2 normalizes the ranking by mapping rank 1 to 0.0, rank 2 to 0.5 and rank 3 to 1.0
- For step 3, using Euclidean distance, a dissimilarity matrix is obtained as shown
- Therefore, students 1 and 2 are most dissimilar, as are students 2 and 4

<table>
<thead>
<tr>
<th>Student</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Excellent</td>
</tr>
<tr>
<td>2</td>
<td>Fair</td>
</tr>
<tr>
<td>3</td>
<td>Good</td>
</tr>
<tr>
<td>4</td>
<td>Excellent</td>
</tr>
</tbody>
</table>
Attributes of Mixed Types

- A database may contain multiple attribute types
- use a weighted formula to combine their effects

\[ d(i, j) = \frac{\sum_{f=1}^{p} \delta^{(f)} d^{(f)}_{ij}}{\sum_{f=1}^{p} \delta^{(f)}} \]

- \( f \) is binary or nominal:
  \[ d^{(f)}_{ij} = 0 \text{ if } x_{if} = x_{jf} \text{, or } d^{(f)}_{ij} = 1 \text{ otherwise} \]
- \( f \) is numeric: use the normalized distance
- \( f \) is ordinal
  - Compute ranks \( r_{if} \) and \( z_{if} = \frac{r_{if} - 1}{M_f - 1} \)
  - Treat \( z_{if} \) as numeric
- The indicator delta is generally set to 1, but
- If \( f \) is asymmetric binary and \( x_{if} = x_{jf} = 0 \), set the indicator to 0
  - recall we removed \( t \) from consideration for “Distance for asymmetric binary 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.

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

Applications: information retrieval, biologic taxonomy, gene feature mapping, ...

Cosine measure: If $d_1$ and $d_2$ 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, $||d||$: the length of vector $d$
Example: Cosine Similarity

- \( \cos(d_1, d_2) = \frac{(d_1 \cdot d_2)}{||d_1|| \cdot ||d_2||} \),
  where \( \cdot \) indicates vector dot product, \( ||d|| \): the length of vector \( d \)

- Ex: Find the similarity between documents 1 and 2.

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

  \[ d_1 \cdot d_2 = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25 \]
  \[ ||d_1|| = (5^2+0^2+3^2+0^2+2^2+0^2+1^2+0^2+1^2+0^2+0^2+0^2)^{0.5} = (42)^{0.5} = 6.481 \]
  \[ ||d_2|| = (3^2+0^2+2^2+0^2+1^2+1^2+1^2+0^2+1^2+0^2+0^2+1^2)^{0.5} = (17)^{0.5} = 4.12 \]

  \[ \cos(d_1, d_2) = 0.94 \]