MSCBD 5002: Knowledge Discovery and Data Mining

Instructor: Lei Chen

Acknowledgement: Slides modified by Dr. Lei Chen based on the slides provided by Jiawei Han, Micheline Kamber, and Jian Pei and Pete Barnum

©2012 Han, Kamber & Pei. All rights reserved.
Chapter 2: Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Measuring Data Similarity and Dissimilarity
- Summary
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 Examples

<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>

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>team</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coach</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>year</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>played</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ball</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>game</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wind</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lost</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>timeout</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>season</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Document 1**
  - TID 1: Bread, Coke, Milk
  - TID 2: Beer, Bread
  - TID 3: Beer, Coke, Diaper, Milk
  - TID 4: Beer, Bread, Diaper, Milk
  - TID 5: Coke, Diaper, Milk
Important Characteristics of Structured Data

- Dimensionality
  - Curse of dimensionality
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Distribution
  - Centrality and dispersion
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 (or dimensions, features, variables):** a data field, representing a characteristic or feature of a data object.
  - *E.g.*, *customer ID, name, address*

- **Types:**
  - Nominal
  - Binary
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled
Attribute Types

- **Nominal**: 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 (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., HIV positive)

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

- **Quantity** (integer or real-valued)
- **Interval**
  - Measured on a scale of *equal-sized units*
  - Values have order
    - E.g., *temperature in C° or F°, calendar dates*
  - No true zero-point
- **Ratio**
  - 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*
Discrete vs. Continuous Attributes

- **Discrete Attribute**
  - Has only a finite or countably 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
Chapter 2: Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Measuring Data Similarity and Dissimilarity
- Summary
Basic Statistical Descriptions of Data

- **Motivation**
  - To better understand the data: central tendency, variation and spread

- **Data dispersion characteristics**
  - median, max, min, quantiles, outliers, variance, etc.

- **Numerical dimensions** correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
    - Boxplot or quantile analysis on sorted intervals

- **Dispersion analysis on computed measures**
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube
Measuring the Central Tendency

- **Mean (algebraic measure) (sample vs. population):**
  \[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \]
  Note: \( n \) is sample size and \( N \) is population size.
  - Weighted arithmetic mean:
  - Trimmed mean: chopping extreme values

- **Median:**
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for grouped data):
    \[ \text{median} = L_1 + \left( \frac{n/2 - (\sum \text{freq})l}{\text{freq}_{\text{median}}} \right) \text{width} \]

- **Mode**
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal

- **Empirical formula:**
  \[ \text{mean} - \text{mode} = 3 \times (\text{mean} - \text{median}) \]
Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data
Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - **Quartiles**: $Q_1$ (25\textsuperscript{th} percentile), $Q_3$ (75\textsuperscript{th} percentile)
  - **Inter-quartile range**: $IQR = Q_3 - Q_1$
  - **Five number summary**: min, $Q_1$, median, $Q_3$, max
  - **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - **Outlier**: usually, a value higher/lower than 1.5 x IQR

- Variance and standard deviation (*sample*: $s$, *population*: $\sigma$)
  - **Variance**: (algebraic, scalable computation)
    \[
    s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \left[ \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right)^2 \right]
    \]
    \[
    \sigma^2 = \frac{1}{N} \sum_{i=1}^{n} (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^{n} x_i^2 - \mu^2
    \]
  - **Standard deviation** $s$ (*or* $\sigma$) is the square root of variance $s^2$ (*or* $\sigma^2$)
Boxplot Analysis

- **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
Properties of Normal Distribution Curve

- The normal (distribution) curve
  - From $\mu - \sigma$ to $\mu + \sigma$: contains about 68% of the measurements ($\mu$: mean, $\sigma$: standard deviation)
  - From $\mu - 2\sigma$ to $\mu + 2\sigma$: contains about 95% of it
  - From $\mu - 3\sigma$ to $\mu + 3\sigma$: contains about 99.7% of it
Graphic Displays of Basic Statistical Descriptions

- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis repres. frequencies
- **Quantile plot**: each value $x_i$ is paired with $f_i$, indicating that approximately 100 $f_i$% of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane
Histogram Analysis

- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- Differs from a bar chart in that it is the area of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent
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
Quantile Plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)

- Plots **quantile** information
  - For a data $x_i$ data sorted in increasing order, $f_i$ indicates that approximately $100 f_i\%$ of the data are below or equal to the value $x_i$
Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- View: Is there a shift in going from one distribution to another?
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.
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.
Positively and Negatively Correlated Data

- The left half fragment is positively correlated
- The right half is negatively correlated
Uncorrelated Data
Chapter 2: Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Measuring Data Similarity and Dissimilarity
- Summary
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
Visual Similarity

- Color
- Texture
Uses for Visual Similarity Measures

- Classification
  - Is it a horse?

- Image Retrieval
  - Show me pictures of horses.

- Unsupervised segmentation
  - Which parts of the image are grass?
Histogram Example
Cumulative Histogram
Adaptive Binning
Higher Dimensional Histograms

- Histograms generalize to any number of features
  - Colors
  - Textures
  - Gradient
  - Depth
Distance Metrics

\[
\begin{align*}
\text{Euclidian distance of 5 units} & \quad = \quad \text{Grayvalue distance of 50 values} \\
? & \quad = ?
\end{align*}
\]
Bin-by-bin

Bad!

Good!
Cross-bin

Bad!

Good!
Distance Measures

- Heuristic
  - Minkowski-form
  - Weighted-Mean-Variance (WMV)
- Nonparametric test statistics
  - $\chi^2$ (Chi Square)
  - Kolmogorov-Smirnov (KS)
  - Cramer/von Mises (CvM)
- Information-theory divergences
  - Kullback-Liebler (KL)
  - Jeffrey-divergence (JD)
- Ground distance measures
  - Histogram intersection
  - Quadratic form (QF)
  - Earth Movers Distance (EMD)
Heuristic Histogram Distances

- Minkowski-form distance $L_p$
  \[ D(I, J) = \left( \sum_i |f(i, I) - f(i, J)|^p \right)^{1/p} \]

- Special cases:
  - $L_1$: absolute, cityblock, or Manhattan distance
  - $L_2$: Euclidian distance
  - $L_{\infty}$: Maximum value distance
More Heuristic Distances

- Weighted-Mean-Variance
  - Only includes minimal information about the distribution

\[ D^r(I, J) = \frac{|\mu_r(I) - \mu_r(J)|}{|\sigma(\mu_r)|} + \frac{|\sigma_r(I) - \sigma_r(J)|}{|\sigma(\sigma_r)|} \]
Examples

- Using
  - Color (CIE Lab)
  - Color + XY
  - Texture (Gabor filter bank)
<table>
<thead>
<tr>
<th>L1 distance</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.jpg" alt="Image of L1 distance" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Jeffrey divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image2.jpg" alt="Image of Jeffrey divergence" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \chi^2 ) statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image3.jpg" alt="Image of ( \chi^2 ) statistics" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quadratic form distance</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image4.jpg" alt="Image of Quadratic form distance" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Earth Mover Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image5.jpg" alt="Image of Earth Mover Distance" /></td>
</tr>
</tbody>
</table>
Concluding thought

= it depends on the application
Data Matrix and Dissimilarity Matrix

- **Data matrix**
  - n data points with p dimensions
  - Two modes

- **Dissimilarity matrix**
  - n data points, but registers only the distance
  - A triangular matrix
  - Single mode
Proximity Measure for 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
Proximity Measure for Binary Attributes

- A contingency table for binary data

<table>
<thead>
<tr>
<th></th>
<th>Object ( i )</th>
<th>Object ( j )</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( q )</td>
<td>0</td>
<td>( q + r )</td>
</tr>
<tr>
<td>0</td>
<td>( s )</td>
<td>( t )</td>
<td>( s + t )</td>
</tr>
<tr>
<td>sum</td>
<td>( q + s )</td>
<td>( r + t )</td>
<td>( p )</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 \textit{asymmetric} binary variables):

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

- Note: Jaccard coefficient is the same as “coherence”:

\[
coherence(i, j) = \frac{\text{sup}(i, j)}{\text{sup}(i) + \text{sup}(j) - \text{sup}(i, j)} = \frac{q}{(q + r) + (q + s) - q}
\]
Dissimilarity between Binary Variables

- Example

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</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>M</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>F</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>M</td>
<td>Y</td>
<td>P</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

- Gender is a symmetric attribute
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N 0

\[
d(jack, mary) = \frac{0 + 1}{2 + 0 + 1} = 0.33
\]
\[
d(jack, jim) = \frac{1 + 1}{1 + 1 + 1} = 0.67
\]
\[
d(jim, mary) = \frac{1 + 2}{1 + 1 + 2} = 0.75
\]
Standardizing Numeric Data

- **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, “+” 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
Example:
Data Matrix and Dissimilarity Matrix

Data Matrix

<table>
<thead>
<tr>
<th>point</th>
<th>attribute1</th>
<th>attribute2</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 Matrix
(with Euclidean Distance)

<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>
**Distance on Numeric Data: 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
    \[d(i, j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + ... + |x_{i_p} - x_{j_p}|\]

- $h = 2$: (L$_2$ norm) Euclidean distance
  \[d(i, j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + ... + |x_{i_p} - x_{j_p}|^2)}\]

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

### Dissimilarity Matrices

#### Manhattan (L₁)

<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>

<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></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>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></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>

#### Supremum

<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></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x2</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x3</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>x4</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>
Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank
- 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
Attributes of Mixed Type

- A database may contain all attribute types
  - Nominal, symmetric binary, asymmetric binary, numeric, ordinal
- One may use a weighted formula to combine their effects
  \[
  d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}
  \]
- \( f \) is binary or nominal:
  \( d_{ij}^{(f)} = 0 \) if \( x_{if} = x_{jf} \), or \( d_{ij}^{(f)} = 1 \) otherwise
- \( f \) is numeric: use the normalized distance
- \( f \) is ordinal
  - Compute ranks \( r_{if} \) and
  - Treat \( z_{if} \) as interval-scaled
    \[
    z_{if} = \frac{r_{if} - 1}{M_f - 1}
    \]
Cosine Similarity

- 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>Document</th>
<th>team</th>
<th>coach</th>
<th>hockey</th>
<th>baseball</th>
<th>soccer</th>
<th>penalty</th>
<th>score</th>
<th>win</th>
<th>loss</th>
<th>season</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document1</td>
<td>5</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Document2</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Document3</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Document4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

- 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, 2, 0, 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*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481
  
  ||d_2|| = (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+0*0+1*1)^{0.5} = (17)^{0.5} = 4.12
  
  \cos(d_1, d_2) = 0.94
Chapter 2: Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary
Summary

- 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.
References

- W. Cleveland, Visualizing Data, Hobart Press, 1993
- U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- S. Santini and R. Jain,” Similarity measures”, IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999
- C. Yu et al., Visual data mining of multimedia data for social and behavioral studies, Information Visualization, 8(1), 2009