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

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

- **Dissimilarity matrix**
  - $n$ data points, but registers only the distance
  - A triangular matrix
  - Single mode

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

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
- Distance measure for symmetric binary variables:
  \[ d(i, j) = \frac{r + s}{q + r + s + t} \]
- Distance measure for asymmetric binary variables:
  \[ \text{sim}_{\text{Jaccard}}(i, j) = \frac{q}{q + r + s} \]

Dissimilarity between Binary Variables

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

<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>

Distance on Numeric Data: Minkowski Distance

- Minkowski distance: A popular distance measure
  \[ d(i, j) = \left( \sum_{k=1}^{p} |x_{ik} - y_{jk}|^h \right)^{1/h} \]
  where \( i = (x_{i1}, x_{i2}, ..., x_{ip}) \) and \( j = (y_{j1}, y_{j2}, ..., y_{jp}) \) 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

Example: Minkowski Distance

<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>Manhattan (L₁)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
</tr>
<tr>
<td>x1</td>
</tr>
<tr>
<td>x2</td>
</tr>
<tr>
<td>x3</td>
</tr>
<tr>
<td>x4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Euclidean (L₂)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
</tr>
<tr>
<td>x1</td>
</tr>
<tr>
<td>x2</td>
</tr>
<tr>
<td>x3</td>
</tr>
<tr>
<td>x4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Supremum</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
</tr>
<tr>
<td>x1</td>
</tr>
<tr>
<td>x2</td>
</tr>
<tr>
<td>x3</td>
</tr>
<tr>
<td>x4</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 \( x_i \) by their rank \( r_{ij} \in \{1, \ldots, M_j\} \)
  - map the range of each variable onto \([0, 1]\) by replacing \( i\)-th object in the \( f\)-th variable by
    \[
    z_{ij} = \frac{r_{ij} - 1}{M_j - 1}
    \]
- 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
  \[
  d(i, f) = \frac{\sum_{j=1}^{F} d_{ij}^2}{\sum_{j=1}^{F} d_{ij}^2}
  \]
- \( f\) is binary or nominal:
  \( d_{ij} = 0 \) if \( x_i = x_{j} \), or \( d_{ij} = 1 \) otherwise
- \( f\) is numeric: use the normalized distance
- \( f\) is ordinal:
  - Compute ranks \( r_d \) and
  - Treat \( z_d \) as interval-scaled
  \[
  z_{ij} = \frac{r_d - 1}{M_j - 1}
  \]

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>soccer</th>
<th>water</th>
<th>penalty</th>
<th>score</th>
<th>wins</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>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</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
  \[
  \cos(d_1, d_2) = \frac{d_1 \bullet d_2}{||d_1|| \cdot ||d_2||}
  \]
  where \( \bullet \) indicates vector dot product, \( ||d|| \): the length of vector \( d \)

Example: Cosine Similarity

- \( \cos(d_1, d_2) = \frac{d_1 \bullet d_2}{||d_1|| \cdot ||d_2||} \), where \( \bullet \) indicates vector dot product, \( ||d|| \): the length of vector \( d \)
- Ex: Find the similarity between documents 1 and 2.
  \[
  d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)
  \]
  \[
  d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)
  \]

\[
\begin{align*}
|d_1| &= \sqrt{5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 1^2 + 0^2 + 0^2 + 0^2} = 5.9165
|d_2| &= \sqrt{3^2 + 2^2 + 0^2 + 2^2 + 0^2 + 0^2 + 1^2 + 1^2 + 1^2 + 0^2} = 4.1231
\end{align*}
\]

\[
\begin{align*}
\cos(d_1, d_2) &= \frac{5 \cdot 3 + 0 \cdot 2 + 3 \cdot 0 + 2 \cdot 2 + 1 \cdot 1 + 0 \cdot 0 + 0 \cdot 0 + 2 \cdot 1 + 0 \cdot 1}{5.9165 \cdot 4.1231}
&= 0.848
\end{align*}
\]
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

Data Quality: Why Preprocess the Data?

- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, ...
  - Consistency: some modified but some not, dangling, ...
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

Major Tasks in Data Preprocessing

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., Occupation = "" (missing data)
  - noisy: containing noise, errors, or outliers
    - e.g., Salary = "-10" (an error)
  - inconsistent: containing discrepancies in codes or names, e.g.,
    - Age = "42", Birthday = "03/07/2010"
    - Was rating "1, 2, 3", now rating "A, B, C"
    - discrepancy between duplicate records
  - Intentional (e.g., disguised missing data)
    - Jan. 1 as everyone's birthday?
Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant: e.g., “unknown”, a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

Noisy Data

- **Noise**: random error or variance in a measured variable
- **Incorrect attribute values** may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- **Other data problems** which require data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

- **Binning**
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- **Regression**
  - smooth by fitting the data into regression functions
- **Clustering**
  - detect and remove outliers
- **Combined computer and human inspection**
  - detect suspicious values and check by human (e.g., deal with possible outliers)
Data Cleaning as a Process

- **Data discrepancy detection**
  - Use metadata (e.g., domain, range, dependency, distribution)
  - Check field overloading
  - Check uniqueness rule, consecutive rule and null rule
  - Use commercial tools
    - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
    - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)
- **Data migration and integration**
  - Data migration tools: allow transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface
  - Integration of the two processes
    - Iterative and interactive

Data Integration

- **Data integration**: Combines data from multiple sources into a coherent store
- **Schema integration**: e.g., A.cust-id = B.cust-
- **Entity identification problem**: Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- **Detecting and resolving data value conflicts**
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- Redundant data occur often when you integrate multiple databases
  - **Object identification**: The same attribute or object may have different names in different databases
  - **Derivable data**: One attribute may be a "derived" attribute in another table, e.g., annual revenue
- **Redundant attributes may be able to be detected by correlation analysis and covariance analysis**
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Correlation Analysis (Nominal Data)

- **$\chi^2$ (chi-square) test**
  \[
  \chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}
  \]
  - The larger the $\chi^2$ value, the more likely the variables are related
  - The cells that contribute the most to the $\chi^2$ value are those whose actual count is very different from the expected count
  - Correlation does not imply causality
    - # of hospitals and # of car-theft in a city are correlated
    - Both are causally linked to the third variable: population
Chi-Square Calculation: An Example

<table>
<thead>
<tr>
<th>Category</th>
<th>Play chess</th>
<th>Not play chess</th>
<th>Sum (row)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Like science fiction</td>
<td>250(90)</td>
<td>200(360)</td>
<td>450</td>
</tr>
<tr>
<td>Not like science fiction</td>
<td>50(210)</td>
<td>1000(840)</td>
<td>1050</td>
</tr>
<tr>
<td>Sum(col.)</td>
<td>300</td>
<td>1200</td>
<td>1550</td>
</tr>
</tbody>
</table>

- $\chi^2$ (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)
  $\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$
- It shows that like_science_fiction and play_chess are correlated in the group

Correlation Analysis (Numeric Data)

- Correlation coefficient (also called Pearson's product moment coefficient)
  $$r_{A,B} = \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{\sqrt{\sum_{i=1}^{n} (a_i - \bar{A})^2} \sqrt{\sum_{i=1}^{n} (b_i - \bar{B})^2}}$$
  where $n$ is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective means of A and B, $\sigma_A$ and $\sigma_B$ are the respective standard deviation of A and B, and $\sum(a_ip)$ is the sum of the AB cross-product.

- If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.
- $r_{A,B} = 0$: independent; $r_{A,B} < 0$: negatively correlated

Visually Evaluating Correlation

Scatter plots showing the similarity from -1 to 1.

Covariance (Numeric Data)

- Covariance is similar to correlation
  $$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{n}$$
  $$r_{A,B} = \frac{Cov(A, B)}{\sigma_A \sigma_B}$$

  where $n$ is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective mean or expected values of A and B, $\sigma_A$ and $\sigma_B$ are the respective standard deviation of A and B.

- Positive covariance: If $Cov_{A,B} > 0$, then A and B both tend to be larger than their expected values
- Negative covariance: If $Cov_{A,B} < 0$ then if A is larger than its expected value, B is likely to be smaller than its expected value
- Independence: $Cov_{A,B} = 0$ but the converse is not true:
  - Some pairs of random variables may have a covariance of 0 but are not independent. Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence.
Covariance: An Example

\[
\text{Cov}(A, B) = E((A - \bar{A})(B - \bar{B})) = \sum_{i=1}^{n} \frac{(a_i - \bar{A})(b_i - \bar{B})}{n}
\]

• It can be simplified in computation as

\[
\text{Cov}(A, B) = E(AB) - \bar{A}B
\]

• Suppose two stocks A and B have the following values in one week:
  (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).

• Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?
  \[E(A) = (2 + 3 + 5 + 4 + 6)/5 = 20/5 = 4\]
  \[E(B) = (5 + 8 + 10 + 11 + 14)/5 = 48/5 = 9.6\]
  \[\text{Cov}(A, B) = (2\times5+3\times8+5\times10+4\times11+6\times14)/5 - 4 \times 9.6 = 4\]

• Thus, A and B rise together since Cov(A, B) > 0.

Next Time

• More Data Preprocessing & Data Warehousing
• Finish reading Ch. 3, start Ch. 4