Chapter 2. Getting to Know Your Data

Jiawei Han, Computer Science, Univ. Illinois at Urbana-Champaign, 2106
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
Types of Data Sets: (1) Record Data

- Relational records
  - Relational tables, highly structured
- Data matrix, e.g., numerical matrix, crosstabs

<table>
<thead>
<tr>
<th>Active Outdoor Crochet Glove</th>
<th>China</th>
<th>England</th>
<th>France</th>
<th>Japan</th>
<th>USA</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active Outdoor Lyre Glove</td>
<td>12.00</td>
<td>4.00</td>
<td>1.00</td>
<td>246.00</td>
<td>257.00</td>
<td></td>
</tr>
<tr>
<td>Inflex Crochet Glove</td>
<td>3.00</td>
<td>6.00</td>
<td>6.00</td>
<td>102.00</td>
<td>145.00</td>
<td></td>
</tr>
<tr>
<td>Inflex Lyre Glove</td>
<td>2.00</td>
<td>6.00</td>
<td>6.00</td>
<td>143.00</td>
<td>145.00</td>
<td></td>
</tr>
<tr>
<td>Triumph Pro Helmet</td>
<td>5.00</td>
<td>25.00</td>
<td>25.00</td>
<td>325.00</td>
<td>340.00</td>
<td></td>
</tr>
<tr>
<td>Triumph Vispide Helmet</td>
<td>5.00</td>
<td>20.00</td>
<td>20.00</td>
<td>374.00</td>
<td>400.00</td>
<td></td>
</tr>
<tr>
<td>Stratus Adult Indent</td>
<td>8.00</td>
<td>8.00</td>
<td>7.00</td>
<td>281.00</td>
<td>290.00</td>
<td></td>
</tr>
<tr>
<td>Stratus Youth Indent</td>
<td>3.00</td>
<td>7.00</td>
<td>7.00</td>
<td>14.00</td>
<td>21.00</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>14.00</td>
<td>49.00</td>
<td>54.00</td>
<td>2.00</td>
<td>3,972.00</td>
<td>2,086.00</td>
</tr>
</tbody>
</table>

- Transaction data

<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>

- Document data: Term-frequency vector (matrix) of text documents
Types of Data Sets: (2) Graphs and Networks

- Transportation network
- World Wide Web
- Molecular Structures
- Social or information networks
Types of Data Sets: (3) Ordered Data

- Video data: sequence of images
- Temporal data: time-series
- Sequential Data: transaction sequences
- Genetic sequence data
Types of Data Sets: (4) Spatial, image and multimedia Data

- Spatial data: maps
- Image data:
- Video data:
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 (e.g., red, blue)
  - Binary (e.g., {true, false})
  - Ordinal (e.g., {freshman, sophomore, junior, senior})
  - Numeric: quantitative
    - Interval-scaled: 100°C is interval scales
    - Ratio-scaled: 100°K is ratio scaled since it is twice as high as 50°K

- **Q1**: Is student ID a nominal, ordinal, or interval-scaled data?
- **Q2**: What about eye color? Or color in the color spectrum of physics?
Attribute Types

- **Nominal**: categories, states, or “names of things”
  - $\text{Hair\_color} = \{\text{auburn, black, blond, brown, grey, red, white}\}$
  - 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
  - $\text{Size} = \{\text{small, medium, large}\}$, 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
- Data Visualization
- 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, ...

- **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: (1) Mean

- **Mean (algebraic measure) (sample vs. population):**
  
  \[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \]  
  \[ \mu = \frac{\sum x}{N} \]

- **Weighted arithmetic mean:**
  
  \[ \bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \]

- **Trimmed mean:**
  
  Chopping extreme values (e.g., Olympics gymnastics score computation)
Measuring the Central Tendency: (2) Median

- **Median:**
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*):

<table>
<thead>
<tr>
<th>age</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–5</td>
<td>200</td>
</tr>
<tr>
<td>6–15</td>
<td>450</td>
</tr>
<tr>
<td>16–20</td>
<td>300</td>
</tr>
<tr>
<td>21–50</td>
<td>1500</td>
</tr>
<tr>
<td>51–80</td>
<td>700</td>
</tr>
<tr>
<td>81–110</td>
<td>44</td>
</tr>
</tbody>
</table>

Approximate median:  

\[
\text{median} = L_1 + \left(\frac{n/2 - (\sum \text{freq})_1}{\text{freq}_{\text{median}}}\right) \times \text{width}
\]  

Sum before the median interval  

Interval width \((L_2 - L_1)\)  

Low interval limit
Measuring the Central Tendency: (3) Mode

- Mode: Value that occurs most frequently in the data

- Unimodal
  - Empirical formula:
    \[ mean - mode = 3 \times (mean - median) \]

- Multi-modal
  - Bimodal
  - Trimodal
Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data
Properties of Normal Distribution Curve

Represent data dispersion, spread

99.7% of the data are within 3 standard deviations of the mean

95% within 2 standard deviations

68% within 1 standard deviation

Represent central tendency

μ - 3σ  μ - 2σ  μ - σ  μ  μ + σ  μ + 2σ  μ + 3σ
Variance and standard deviation \((sample: s, population: \sigma)\)

- **Variance**: (algebraic, scalable computation)
  - Q: Can you compute it incrementally and efficiently?

\[
s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \left[ \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right)^2 \right]
\]

\[
\sigma^2 = \frac{1}{N} \sum_{i=1}^{n} (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^{n} x_i^2 - \mu^2
\]

- **Standard deviation** \(s\) (or \(\sigma\)) is the square root of variance \(s^2\) (or \(\sigma^2\))
Graphic Displays of Basic Statistical Descriptions

- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis represent frequencies
- **Quantile plot**: each value $x_i$ is paired with $f_i$, indicating that approximately $100f_i\%$ of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariate 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
Measuring the Dispersion of Data: Quartiles & Boxplots

- **Quartiles**: $Q_1$ (25th percentile), $Q_3$ (75th percentile)
- **Inter-quartile range**: $IQR = Q_3 - Q_1$
- **Five number summary**: min, $Q_1$, median, $Q_3$, max
- **Boxplot**: Data is represented with a box
  - $Q_1$, $Q_3$, IQR: The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - Median ($Q_2$) 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
    - **Outlier**: usually, a value higher/lower than $1.5 \times IQR$
Visualization of Data Dispersion: 3-D Boxplots
Histogram Analysis

- Histogram: Graph display of tabulated frequencies, shown as bars

- Differences between histograms and bar charts
  - Histograms are used to show distributions of variables while bar charts are used to compare variables
  - Histograms plot binned quantitative data while bar charts plot categorical data
  - Bars can be reordered in bar charts but not in histograms
  - 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
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 is 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
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary
Data Visualization

- Why data visualization?
  - Gain insight into an information space by mapping data onto graphical primitives
  - Provide qualitative overview of large data sets
  - Search for patterns, trends, structure, irregularities, relationships among data
  - Help find interesting regions and suitable parameters for further quantitative analysis
  - Provide a visual proof of computer representations derived

- Categorization of visualization methods:
  - Pixel-oriented visualization techniques
  - Geometric projection visualization techniques
  - Icon-based visualization techniques
  - Hierarchical visualization techniques
  - Visualizing complex data and relations
Pixel-Oriented Visualization Techniques

- For a data set of \( m \) dimensions, create \( m \) windows on the screen, one for each dimension.
- The \( m \) dimension values of a record are mapped to \( m \) pixels at the corresponding positions in the windows.
- The colors of the pixels reflect the corresponding values.

(a) Income  (b) Credit Limit  (c) transaction volume  (d) age
Laying Out Pixels in Circle Segments

- To save space and show the connections among multiple dimensions, space filling is often done in a circle segment.

(b) Laying out pixels in circle segment
Geometric Projection Visualization Techniques

- Visualization of geometric transformations and projections of the data
- Methods
  - Direct visualization
  - Scatterplot and scatterplot matrices
  - Landscapes
  - Projection pursuit technique: Help users find meaningful projections of multidimensional data
  - Prosection views
  - Hyperslice
  - Parallel coordinates
Direct Data Visualization

Ribbons with Twists Based on Vorticity

data courtesy of NCSA, University of Illinois at Urbana-Champaign
Scatterplot Matrices

Matrix of scatterplots (x-y-diagrams) of the k-dim. data [total of $(k^2/2 - k)$ scatterplots]
Landscapes

- Visualization of the data as a perspective landscape

- The data needs to be transformed into a (possibly artificial) 2D spatial representation which preserves the characteristics of the data

News articles visualized as a landscape
Parallel Coordinates

- n equidistant axes which are parallel to one of the screen axes and correspond to the attributes
- The axes are scaled to the [minimum, maximum]: range of the corresponding attribute
- Every data item corresponds to a polygonal line which intersects each of the axes at the point which corresponds to the value for the attribute
Parallel Coordinates of a Data Set
Announcements: Homework #1 and 4th Credit Project

- CS412: The First Homework
  - Assignment #1 is ready and is distributed today!
  - Please check lecture page linking to the assignment #1

- Information About the Project for the 4th Credit
  - This project is part of WSDM 2017 Cup (http://www.wsdm-cup-2017.org/triple-scoring.html)
  - Please choose one from the following two competition tasks.
    - Choice #1: Triple Scoring: Computing relevance scores for triples from type-like relations
    - Choice #2: Vandalism Detection for Wikipages

- Submission: You can team up and each team will submit one program to the WSDM 2017 Cup evaluation system—Grading based on WSDM 2017 Cup evaluation results
  - The information about groups and registrations will be given later
Project #1: Triple Scoring: Relevance Scores for Triples

- **Triple Scoring**: Computing relevance scores for triples from type-like relations

- **Example**:
  - The triple “Johnny_Depp profession Actor” should get a high score, because acting is Depp’s main profession, whereas “Quentin_Tarantino profession Actor” should get a low score, because Tarantino is more of a director than an actor. Such scores are a basic ingredient for ranking results in entity search.

- **Training data (given by cup organizers)**
  - A training set consisting of triples and their relevance scores (in the range of [0, 1]), as obtained from human judges
  - Additional information that can be used for distant supervision learning, such as text corpus

- **The objective is to predict the relevance scores for the given triples**: The prediction accuracy will be evaluated against ground truth from human judges
Project #2: Vandalism Detection for Wikipages

- **Background:** Wikidata is the new, large-scale knowledge base of the Wikimedia Foundation which can be edited by anyone. Its knowledge is increasingly used within Wikipedia as well as in all kinds of information systems, which imposes high demands on its integrity. Nevertheless, Wikidata frequently gets vandalized, exposing all its users to the risk of spreading vandalized and falsified information.

- **Task:** Given a Wikidata revision, compute a vandalism score denoting the likelihood of this revision being vandalism (or similarly damaging).

- **Data**
  - **Training:** We will be provided with a training corpus, consisting of Wikidata revisions and whether they are considered vandalism.
  - **Testing:** There will be a test data which is not published during the contest, but to be used in final evaluation.

- **Submission:** You may team up to work on this project. If there are multiple teams working on this project, we may ensemble different teams' results to generate one model and submit to WSDM Cup's competition, based on your agreement. Grading will be based on your performance and final report.
Icon-Based Visualization Techniques

- Visualization of the data values as features of icons
- Typical visualization methods
  - Chernoff Faces
  - Stick Figures
- General techniques
  - Shape coding: Use shape to represent certain information encoding
  - Color icons: Use color icons to encode more information
  - Tile bars: Use small icons to represent the relevant feature vectors in document retrieval
Chernoff Faces

- A way to display variables on a two-dimensional surface, e.g., let x be eyebrow slant, y be eye size, z be nose length, etc.

- The figure shows faces produced using 10 characteristics--head eccentricity, eye size, eye spacing, eye eccentricity, pupil size, eyebrow slant, nose size, mouth shape, mouth size, and mouth opening): Each assigned one of 10 possible values, generated using Mathematica (S. Dickson)


A census data figure showing age, income, gender, education, etc.

A 5-piece stick figure (1 body and 4 limbs w. different angle/length)
Hierarchical Visualization Techniques

- Visualization of the data using a hierarchical partitioning into subspaces

- Methods
  - Dimensional Stacking
  - Worlds-within-Worlds
  - Tree-Map
  - Cone Trees
  - InfoCube
Dimensional Stacking

- Partitioning of the n-dimensional attribute space in 2-D subspaces, which are ‘stacked’ into each other
- Partitioning of the attribute value ranges into classes. The important attributes should be used on the outer levels.
- Adequate for data with ordinal attributes of low cardinality
- But, difficult to display more than nine dimensions
- Important to map dimensions appropriately
Visualization of oil mining data with longitude and latitude mapped to the outer x-, y-axes and ore grade and depth mapped to the inner x-, y-axes
- Assign the function and two most important parameters to innermost world
- Fix all other parameters at constant values - draw other (1 or 2 or 3 dimensional worlds choosing these as the axes)
- Software that uses this paradigm
  - N–vision: Dynamic interaction through data glove and stereo displays, including rotation, scaling (inner) and translation (inner/outer)
  - Auto Visual: Static interaction by means of queries
Tree-Map

- Screen-filling method which uses a hierarchical partitioning of the screen into regions depending on the attribute values
- The x- and y-dimension of the screen are partitioned alternately according to the attribute values (classes)
- A 3-D visualization technique where hierarchical information is displayed as nested semi-transparent cubes
- The outermost cubes correspond to the top level data, while the subnodes or the lower level data are represented as smaller cubes inside the outermost cubes, etc.
Three-D Cone Trees

- 3D cone tree visualization technique works well for up to a thousand nodes or so
- First build a 2D circle tree that arranges its nodes in concentric circles centered on the root node
- Cannot avoid overlaps when projected to 2D
- Graph from Nadeau Software Consulting website: Visualize a social network data set that models the way an infection spreads from one person to the next
Visualizing Complex Data and Relations: Tag Cloud

- **Tag cloud**: Visualizing user-generated tags
- The importance of tag is represented by font size/color
- Popularly used to visualize word/phrase distributions

KDD 2013 Research Paper Title Tag Cloud

Newsmap: Google News Stories in 2005
Visualizing Complex Data and Relations: Social Networks

- Visualizing non-numerical data: social and information networks

A typical network structure

A social network

organizing information networks
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
Similarity, Dissimilarity, and Proximity

- **Similarity measure** or **similarity function**
  - A real-valued function that quantifies the similarity between two objects
  - Measure how two data objects are alike: The higher value, the more alike
  - Often falls in the range $[0,1]$: 0: no similarity; 1: completely similar

- **Dissimilarity** (or **distance**) measure
  - Numerical measure of how different two data objects are
  - In some sense, the inverse of similarity: The lower, the more alike
  - Minimum dissimilarity is often 0 (i.e., completely similar)
  - Range $[0,1]$ or $[0, \infty)$, depending on the definition

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

- Data matrix
  - A data matrix of n data points with l dimensions
- Dissimilarity (distance) matrix
  - n data points, but registers only the distance \( d(i, j) \) (typically metric)
  - Usually symmetric, thus a triangular matrix
  - Distance functions are usually different for real, boolean, categorical, ordinal, ratio, and vector variables
  - Weights can be associated with different variables based on applications and data semantics
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 (by 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
  \[ d(i, j) = \sqrt[p]{\sum_{l=1}^{p} |x_{il} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{il} - x_{jl}|^p} \]
  where \( i = (x_{i1}, x_{i2}, ..., x_{il}) \) and \( j = (x_{j1}, x_{j2}, ..., x_{jl}) \) are two \( l \)-dimensional data objects, and \( p \) is the order (the distance so defined is also called L-\( p \) norm)

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

- A distance that satisfies these properties is a metric

- Note: There are nonmetric dissimilarities, e.g., set differences
Special Cases of Minkowski Distance

- \( p = 1 \): (\( L_1 \) norm) Manhattan (or city block) distance
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors
    \[
d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{il} - x_{jl}|
    \]

- \( p = 2 \): (\( L_2 \) norm) Euclidean distance
  \[
d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{il} - x_{jl}|^2}
  \]

- \( p \to \infty \): (\( L_{\max} \) norm, \( L_{\infty} \) norm) “supremum” distance
  - The maximum difference between any component (attribute) of the vectors
    \[
d(i, j) = \lim_{p \to \infty} \left( |x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{il} - x_{jl}|^p \right)^{\frac{1}{p}} = \max_{f=1}^{l} |x_{if} - x_{jf}|
    \]
Example: Minkowski Distance at Special Cases

<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 ($L_1$)

\[
L_1 \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 5 \\ 3 \\ 6 \end{pmatrix}
\]

Euclidean ($L_2$)

\[
L_2 \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 3.61 \\ 0 \\ 2.24 \\ 4.24 \end{pmatrix}
\]

Supremum ($L_\infty$)

\[
L_\infty \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 3 \\ 2 \\ 3 \end{pmatrix}
\]
Proximity Measure for Binary Attributes

- A contingency table for binary data

<table>
<thead>
<tr>
<th></th>
<th>Object 1</th>
<th>Object 0</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object 1</td>
<td>1</td>
<td>q</td>
<td>r</td>
</tr>
<tr>
<td>Object 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 (similarity measure for asymmetric binary variables):

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

- Note: Jaccard coefficient is the same as (a concept discussed in Pattern Discovery)

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

<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 (not counted in)
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N be 0
- Distance: \( d(i, j) = \frac{r + s}{q + r + s} \)

\[
\begin{align*}
    d(\text{jack}, \text{mary}) &= \frac{0 + 1}{2 + 0 + 1} = 0.33 \\
    d(\text{jack}, \text{Jim}) &= \frac{1 + 1}{1 + 1 + 1} = 0.67 \\
    d(\text{Jim}, \text{mary}) &= \frac{1 + 2}{1 + 1 + 2} = 0.75
\end{align*}
\]
Proximity Measure for Categorical Attributes

- Categorical data, also called nominal attributes
  - Example: Color (red, yellow, blue, green), profession, etc.

- **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
Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank (e.g., freshman, sophomore, junior, senior)
- Can be treated like interval-scaled
  - Replace *an ordinal variable value* by its 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
    $$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$
  - Example: freshman: 0; sophomore: 1/3; junior: 2/3; senior 1
    - Then distance: $d($freshman, senior$) = 1$, $d($junior, senior$) = 1/3$
  - Compute the dissimilarity using methods for interval-scaled variables
Attributes of Mixed Type

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

- A **document** can be represented by a bag of terms or a long vector, with each attribute recording the frequency of a particular term (such as word, keyword, or phrase) in the document

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Document</td>
<td>team</td>
<td>coach</td>
<td>hockey</td>
<td>baseball</td>
<td>soccer</td>
<td>penalty</td>
</tr>
<tr>
<td>Document 1</td>
<td>5</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Document 2</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Document 3</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Document 4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

- Other vector objects: Gene features in micro-arrays
- Applications: Information retrieval, biologic taxonomy, gene feature mapping, etc.
- Cosine measure: If \( d_1 \) and \( d_2 \) are two vectors (e.g., term-frequency vectors), then

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

where \( \cdot \) indicates vector dot product, \( ||d|| \): the length of vector \( d \)
Example: Calculating Cosine Similarity

Calculating Cosine Similarity:
\[ \cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\| d_1 \| \times \| d_2 \|} \]

where \( \cdot \) indicates vector dot product, \( \| \cdot \| \): the length of vector \( d \)

Ex: Find the similarity between documents 1 and 2.

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

First, calculate vector dot product

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

Then, calculate \( \| d_1 \| \) and \( \| d_2 \| \)

\[ \| d_1 \| = \sqrt{5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2} = 6.481 \]
\[ \| d_2 \| = \sqrt{3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2} = 4.12 \]

Calculate cosine similarity:
\[ \cos(d_1, d_2) = \frac{25}{6.481 \times 4.12} = 0.94 \]
Announcements: Meeting of the 4th Credit Project

- **CS412: Assignment #1** was distributed last Tuesday!
  - The due date is Sept. 15. No late homework will be accepted!!
- **Waitlist is cleared:** We took 50 additional students into the video only session
  - Please find your status with Holly. You are either in or out (wait for Spring 2017)
- **Meeting for Project for the 4th Credit**
  - You can change from 4 to 3 credit or from 3 to 4 credits by sending me e-mails
  - **Meeting time and location:** 10-11am Friday (tomorrow!) at 0216 SC
  - This project is part of WSDM 2017 Cup
  - Choice #1: **Triple Scoring:** Computing relevance scores for triples from type-like relations
  - Choice #2: **Vandalism Detection** for Wikipages
  - Tas/PhD student/postdoc will give you the details in the Friday meeting! **Must attend if you want to do the 4th credit project!!!**
KL Divergence: Comparing Two Probability Distributions

- The Kullback-Leibler (KL) divergence: Measure the difference between two probability distributions over the same variable $x$
- From information theory, closely related to relative entropy, information divergence, and information for discrimination
- $D_{KL}(p(x) \| q(x))$: divergence of $q(x)$ from $p(x)$, measuring the information lost when $q(x)$ is used to approximate $p(x)$

$$D_{KL}(p(x) \| q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$$

Continuous form

$$D_{KL}(p(x) \| q(x)) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx$$

Acknowledgment: Wikipedia entry: *The Kullback-Leibler (KL) divergence*
More on KL Divergence

- The KL divergence measures the expected number of extra bits required to code samples from $p(x)$ ("true" distribution) when using a code based on $q(x)$, which represents a theory, model, description, or approximation of $p(x)$.
- The KL divergence is not a distance measure, not a metric: asymmetric, not satisfy triangular inequality ($D_{KL}(P \ Q)$ does not equal $D_{KL}(Q \ P)$).
- In applications, $P$ typically represents the "true" distribution of data, observations, or a precisely calculated theoretical distribution, while $Q$ typically represents a theory, model, description, or approximation of $P$.
- The Kullback–Leibler divergence from $Q$ to $P$, denoted $D_{KL}(P \ Q)$, is a measure of the information gained when one revises one's beliefs from the prior probability distribution $Q$ to the posterior probability distribution $P$. In other words, it is the amount of information lost when $Q$ is used to approximate $P$.
- The KL divergence is sometimes also called the information gain achieved if $P$ is used instead of $Q$. It is also called the relative entropy of $P$ with respect to $Q$.

\[ D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)} \]
Subtlety at Computing the KL Divergence

- Base on the formula, \( D_{KL}(P,Q) \geq 0 \) and \( D_{KL}(P \mid \mid Q) = 0 \) if and only if \( P = Q \)
- How about when \( p = 0 \) or \( q = 0 \)?
  - \( \lim_{p \to 0} p \log p = 0 \)
  - when \( p 
eq 0 \) but \( q = 0 \), \( D_{KL}(p \mid \mid q) \) is defined as \( \infty \), i.e., if one event \( e \) is possible (i.e., \( p(e) > 0 \)), and the other predicts it is absolutely impossible (i.e., \( q(e) = 0 \)), then the two distributions are absolutely different
- However, in practice, \( P \) and \( Q \) are derived from frequency distributions, not counting the possibility of unseen events. Thus smoothing is needed
- Example: \( P : (a : 3/5, b : 1/5, c : 1/5) \). \( Q : (a : 5/9, b : 3/9, d : 1/9) \)
  - need to introduce a small constant \( \epsilon \), e.g., \( \epsilon = 10^{-3} \)
  - The sample set observed in \( P \), \( SP = \{a, b, c\} \), \( SQ = \{a, b, d\} \), \( SU = \{a, b, c, d\} \)
  - Smoothing, add missing symbols to each distribution, with probability \( \epsilon \)
  - \( P' : (a : 3/5 - \epsilon/3, b : 1/5 - \epsilon/3, c : 1/5 - \epsilon/3, d : \epsilon) \)
  - \( Q' : (a : 5/9 - \epsilon/3, b : 3/9 - \epsilon/3, c : \epsilon, d : 1/9 - \epsilon/3) \)
  - \( D_{KL}(P' \mid \mid Q') \) can then be computed easily

\[
D_{KL}(p(x) \mid \mid q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}
\]
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