Data Mining: Data

Lecture Notes for Chapter 2
Slides by Tan, Steinbach, Kumar adapted by Michael Hahsler

Look for accompanying R code on the course web site.
Topics

• Attributes/Features
• Types of Data Sets
• Data Quality
• Data Preprocessing
• Similarity and Dissimilarity
• Density
What is Data?

• Collection of data objects and their attributes

• An attribute (in Data Mining and Machine learning often "feature") is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic

• A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

<table>
<thead>
<tr>
<th>Tid</th>
<th>Refund</th>
<th>Marital Status</th>
<th>Taxable Income</th>
<th>Cheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Single</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Married</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Single</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Married</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Divorced</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Married</td>
<td>60K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Divorced</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Single</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Married</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Single</td>
<td>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Attribute Values

• Attribute values are numbers or symbols assigned to an attribute

• Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    • Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    • Example: Attribute values for ID and age are integers
    • But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value
Measurement of Length

The way you measure an attribute is somewhat may not match the attributes properties.
Types of Attributes - Scales

There are different types of attributes

- **Nominal**
  - Examples: ID numbers, eye color, zip codes

- **Ordinal**
  - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}

- **Interval**
  - Examples: calendar dates, temperatures in Celsius or Fahrenheit.

- **Ratio**
  - Examples: temperature in Kelvin, length, time, counts
<table>
<thead>
<tr>
<th>Attribute Type</th>
<th>Description</th>
<th>Examples</th>
<th>Operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. ($=, \neq$)</td>
<td>zip codes, employee ID numbers, eye color, sex: {male, female}</td>
<td>mode, entropy, contingency correlation, $\chi^2$ test</td>
</tr>
<tr>
<td>Ordinal</td>
<td>The values of an ordinal attribute provide enough information to order objects. ($&lt;, &gt;$)</td>
<td>hardness of minerals, {good, better, best}, grades, street numbers</td>
<td>median, percentiles, rank correlation, run tests, sign tests</td>
</tr>
<tr>
<td>Interval</td>
<td>For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. ($+,-$)</td>
<td>calendar dates, temperature in Celsius or Fahrenheit</td>
<td>mean, standard deviation, Pearson's correlation, $t$ and $F$ tests</td>
</tr>
<tr>
<td>Ratio</td>
<td>For ratio variables, both differences and ratios are meaningful. ($\times, \div$)</td>
<td>temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current</td>
<td>geometric mean, harmonic mean, percent variation</td>
</tr>
<tr>
<td>Attribute Level</td>
<td>Transformation</td>
<td>Comments</td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>----------------</td>
<td>----------</td>
<td></td>
</tr>
<tr>
<td>Nominal</td>
<td>Any permutation of values</td>
<td>If all employee ID numbers were reassigned, would it make any difference?</td>
<td></td>
</tr>
<tr>
<td>Ordinal</td>
<td>An order preserving change of values, i.e., $new_value = f(old_value)$ where $f$ is a monotonic function.</td>
<td>An attribute encompassing the notion of good, better best can be represented equally well by the values ${1, 2, 3}$ or by ${0.5, 1, 10}$.</td>
<td></td>
</tr>
<tr>
<td>Interval</td>
<td>$new_value = a \times old_value + b$ where $a$ and $b$ are constants</td>
<td>Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).</td>
<td></td>
</tr>
<tr>
<td>Ratio</td>
<td>$new_value = a \times old_value$</td>
<td>Length can be measured in meters or feet.</td>
<td></td>
</tr>
</tbody>
</table>
Discrete and Continuous Attributes

• Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes

• Continuous Attribute
  - Has real numbers as attribute values
  - Examples: 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.
Topics

• Attributes/Features
• Types of Data Sets
• Data Quality
• Data Preprocessing
• Similarity and Dissimilarity
• Density
Types of data sets

- **Record**
  - Data Matrix
  - Document Data
  - Transaction Data

- **Graph**
  - World Wide Web
  - Molecular Structures

- **Ordered**
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data
Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

<table>
<thead>
<tr>
<th>Tid</th>
<th>Refund</th>
<th>Marital Status</th>
<th>Taxable Income</th>
<th>Cheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Single</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Married</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Single</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Married</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Divorced</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Married</td>
<td>60K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Divorced</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Single</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Married</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Single</td>
<td>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute.
- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute.

<table>
<thead>
<tr>
<th>Sepal.Length</th>
<th>Sepal.Width</th>
<th>Petal.Length</th>
<th>Petal.Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.6</td>
<td>2.7</td>
<td>4.2</td>
<td>1.3</td>
</tr>
<tr>
<td>6.5</td>
<td>3.0</td>
<td>5.8</td>
<td>2.2</td>
</tr>
<tr>
<td>6.8</td>
<td>2.8</td>
<td>4.8</td>
<td>1.4</td>
</tr>
<tr>
<td>5.7</td>
<td>3.8</td>
<td>1.7</td>
<td>0.3</td>
</tr>
<tr>
<td>5.5</td>
<td>2.5</td>
<td>4.0</td>
<td>1.3</td>
</tr>
<tr>
<td>4.8</td>
<td>3.0</td>
<td>1.4</td>
<td>0.1</td>
</tr>
<tr>
<td>5.2</td>
<td>4.1</td>
<td>1.5</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Document Data

Each document becomes a `term' vector,

- each term is a component (attribute) of the vector,
- the value of each component is the number of times the corresponding term occurs in the document.

<table>
<thead>
<tr>
<th></th>
<th>Term 1</th>
<th>Term 2</th>
<th>...</th>
<th>Term 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document 1</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Document 2</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Document 3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Transaction Data

A special type of record data, where
- each record (transaction) involves a set of items.
- For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

<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>
Graph Data

Examples: Generic graph and HTML Links

<ol>
  <li><a href="papers/papers.html#bbbb">Data Mining</a></li>
  <li><a href="papers/papers.html#aaaa">Graph Partitioning</a></li>
  <li><a href="papers/papers.html#aaaa">Parallel Solution of Sparse Linear System of Equations</a></li>
  <li><a href="papers/papers.html#ffff">N-Body Computation and Dense Linear System Solvers</a></li>
</ol>
Chemical Data

Benzene Molecule: $\text{C}_6\text{H}_6$
Ordered Data

Sequences of transactions

An element of the sequence

Items/Events

( A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)
Ordered Data

Genomic sequence data

GGTTCCGCCTTCAGCCCCC"CGCC
CGCAGGGCCGCCCCGC"CGC"CTC
GAGAAGGGGCCC"CGCCTGGCGGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCCGAC"CTGGACATGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGCGAGCGGACAG
GCCAAGTAGAACA"CGCGAAGCGC
TGGGCTGCG"CTGCTGC"CGACCAGGG
Ordered Data

Spatio-Temporal Data

Average Monthly Temperature of land and ocean
Topics

- Attributes/Features
- Types of Data Sets
- Data Quality
- Data Preprocessing
- Similarity and Dissimilarity
- Density
Data Quality

• What kinds of data quality problems?
• How can we detect problems with the data?
• What can we do about these problems?

• Examples of data quality problems:
  - Noise and outliers
  - missing values
  - duplicate data
Noise

Noise refers to modification of original values

- Examples: distortion of a person’s voice when talking on a poor phone, “snow” on television screen, measurement errors.

- Find less noisy data
- De-noise (signal processing)
Outliers

Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set

- Outlier detection + remove outliers
Missing Values

• Reasons for missing values
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)

• Handling missing values
  - Eliminate data objects with missing value
  - Eliminate feature with missing values
  - Ignore the Missing Value During Analysis
  - Estimate missing values = Imputation
    (e.g., replace with mean or weighted mean where all possible values are weighted by their probabilities)
Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources

- Examples:
  - Same person with multiple email addresses

- Data cleaning
  - Process of dealing with duplicate data issues
  - ETL tools typically support deduplication
Topics

- Attributes/Features
- Types of Data Sets
- Data Quality
- Data Preprocessing
- Similarity and Dissimilarity
- Density
Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation
Aggregation

• Combining two or more attributes (or objects) into a single attribute (or object)

• Purpose
  - Data reduction
    • Reduce the number of attributes or objects
  - Change of scale
    • Cities aggregated into regions, states, countries, etc
  - More “stable” data
    • Aggregated data tends to have less variability (e.g., reduce seasonality by aggregation to yearly data)
Aggregation

Variation of Precipitation in Australia

Standard Deviation of Average Monthly Precipitation

Standard Deviation of Average Yearly Precipitation
Sampling

• Sampling is the main technique employed for data selection.
  - It is often used for both the preliminary investigation of the data and the final data analysis.

• Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.

• Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.
Sampling ...

- The key principle for effective sampling is the following:
  - using a sample will work almost as well as using the entire data sets, if the sample is representative
  - A sample is representative if it has approximately the same property (of interest) as the original set of data
Types of Sampling

- **Sampling without replacement**
  - As each item is selected, it is removed from the population

- **Sampling with replacement**
  - Objects are not removed from the population as they are selected for the sample. Note: the same object can be picked up more than once

- **Simple Random Sampling**
  - There is an equal probability of selecting any particular item

- **Stratified sampling**
  - Split the data into several partitions; then draw random samples from each partition
Sample Size

8000 points

2000 Points

500 Points
Sample Size

- What sample size is necessary to get at least one object from each of 10 groups.
Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies.
- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful.
  - Density $\to 0$
  - All points tend to have the same Euclidean distance to each other.

*Randomly generate 500 points*
*Compute difference between max and min distance between any pair of points*
Dimensionality Reduction

• Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise

• Techniques
  - Principle Component Analysis
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques
Dimensionality Reduction: PCA

Goal is to find a projection (new axes) that captures the largest amount of variation in data.

Methods: Find the eigenvectors of the covariance matrix or use SVD.
**Goal:** Unroll the "swiss roll!"

**Solution:** Use a non-metric space, i.e., distances are not measured by Euclidean distance.

- Construct a neighbourhood graph
- For each pair of points in the graph, compute the shortest path distances – geodesic distances
Feature Subset Selection

- Another way to reduce dimensionality of data

- **Redundant features**
  - duplicate much or all of the information contained in one or more other attributes (are correlated)
  - Example: purchase price of a product and the amount of sales tax paid

- **Irrelevant features**
  - contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA
Feature Subset Selection

- Brute-force approach:
  - Try all possible feature subsets as input to data mining algorithm

- Embedded approaches:
  - Feature selection occurs naturally as part of the data mining algorithm (e.g., regression, decision trees)

- Filter approaches:
  - Features are selected before data mining algorithm is run (e.g. highly correlated features)

- Wrapper approaches:
  - Use the data mining algorithm as a black box to find best subset of attributes (often using greedy search)
Feature Creation

Create new attributes that can capture the important information in a data set much more efficiently than the original attributes.

Three general methodologies:

- Feature Extraction
  - Domain-specific (e.g., face recognition in image mining)

- Feature Construction
  - combining features (interactions: multiply features)

- Mapping Data to New Space
Mapping Data to a New Space

Fourier transform
Wavelet transform

Two Sine Waves

Two Sine Waves + Noise

Frequency
Discretization Without Using Class Labels

Data

Equal interval width

Equal frequency

K-means
Discretization Using Class Labels

- **Entropy based approach** (entropy is a measure of unorder). Bisect data to reduce entropy.

For x and y:
- 3 categories for both x and y
- 5 categories for both x and y
Attribute Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
  - Simple functions: \( x^k, \log(x), e^x, |x| \)
  - Standardization and Normalization

\[
x' = \frac{x - \bar{x}}{sd(x)}
\]
Topics

- Attributes/Features
- Types of Data Sets
- Data Quality
- Data Preprocessing
- Similarity and Dissimilarity
- Density
Similarity and Dissimilarity

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

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

- **Proximity** refers to a similarity or dissimilarity.
## Similarity/Dissimilarity for Simple Attributes

$p$ and $q$ are the attribute values for two data objects.

<table>
<thead>
<tr>
<th>Attribute Type</th>
<th>Dissimilarity</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>$d = \begin{cases} 0 &amp; \text{if } p = q \ 1 &amp; \text{if } p \neq q \end{cases}$</td>
<td>$s = \begin{cases} 1 &amp; \text{if } p = q \ 0 &amp; \text{if } p \neq q \end{cases}$</td>
</tr>
<tr>
<td>Ordinal</td>
<td>$d = \frac{</td>
<td>p-q</td>
</tr>
<tr>
<td>Interval or Ratio</td>
<td>$d =</td>
<td>p-q</td>
</tr>
</tbody>
</table>

**Table 5.1.** Similarity and dissimilarity for simple attributes
Euclidean Distance

- Euclidean Distance

\[ \text{dist} = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2} = \| p - q \|_2 \]

Where \( n \) is the number of dimensions (attributes) and \( p_k \) and \( q_k \) are, respectively, the \( k^{\text{th}} \) attributes (components) or data objects \( p \) and \( q \).

- Standardization is necessary, if scales differ.
Euclidean Distance

Distance Matrix

<table>
<thead>
<tr>
<th>point</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>p2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>p3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>p4</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>
Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

\[
dist = \left( \sum_{k=1}^{n} |p_k - q_k|^r \right)^{\frac{1}{r}}
\]

Where \( r \) is a parameter, \( n \) is the number of dimensions (attributes) and \( p_k \) and \( q_k \) are, respectively, the kth attributes (components) or data objects \( p \) and \( q \).
Minkowski Distance: Examples

- $r = 1$. City block (Manhattan, taxicab, $L_1$ norm) distance.
  - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors

- $r = 2$. Euclidean distance

- $r \to \infty$. “supremum” ($L_{\text{max}}$ norm, $L_{\infty}$ norm) distance.
  - This is the maximum difference between any component of the vectors

- Do not confuse $r$ with $n$, i.e., all these distances are defined for all numbers of dimensions.
### Minkowski Distance

#### Distance Matrix

<table>
<thead>
<tr>
<th>point</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>p2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>p3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>p4</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

#### L1 Distance Matrix

<table>
<thead>
<tr>
<th></th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
<th>p4</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>p2</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>p3</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>p4</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

#### L2 Distance Matrix

<table>
<thead>
<tr>
<th></th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
<th>p4</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>0</td>
<td>2.828</td>
<td>3.162</td>
<td>5.099</td>
</tr>
<tr>
<td>p2</td>
<td>2.828</td>
<td>0</td>
<td>1.414</td>
<td>3.162</td>
</tr>
<tr>
<td>p3</td>
<td>3.162</td>
<td>1.414</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>p4</td>
<td>5.099</td>
<td>3.162</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

#### L∞ Distance Matrix

<table>
<thead>
<tr>
<th></th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
<th>p4</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>p2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>p3</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>p4</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Distance Matrix
Mahalanobis Distance

\[ \text{mahalanobis}(p, q) = (p - q) \Sigma^{-1} (p - q)^T \]

\( \Sigma \) is the covariance matrix of the input data \( X \)

\[ \Sigma_{j,k} = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k) \]

Measures how many standard deviations two points are away from each other → scale invariant measure

**Example:** For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.
Mahalanobis Distance

Covariance Matrix:

\[
\Sigma = \begin{bmatrix}
0.3 & 0.2 \\
0.2 & 0.3 \\
\end{bmatrix}
\]

A: (0.5, 0.5)
B: (0, 1)
C: (1.5, 1.5)

\[\text{Mahal}(A,B) = 5\]
\[\text{Mahal}(A,C) = 4\]

Data varies in direction A-C more than in A-B!
Cosine Similarity

• If \( d_1 \) and \( d_2 \) are two count vectors, then
  \[
s_{\cos}(d_1, d_2) = \frac{(d_1 \cdot d_2)}{||d_1|| \; ||d_2||},
\]
  where \( \cdot \) indicates vector dot product and \( ||d|| \) is the length of vector \( d \).

• Example:
  \[
d_1 = 3205000200
  \]
  \[
d_2 = 1000000102
  \]
  \[
d_1 \cdot d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5
  \]
  \[
  ||d_1|| = (3^2+2^2+0^2+5^2+0^2+0^2+0^2+2^2+0^2+0^2) = (42)^{0.5} = 6.481
  \]
  \[
  ||d_2|| = (1^2+0^2+0^2+0^2+0^2+0^2+0^2+1^2+0^2+2^2) = (6)^{0.5} = 2.245
  \]
  \[
s_{\cos}(d_1, d_2) = .3150
  \]

Cosine similarity is often used for word count vectors to compare documents.
Common Properties of a Distance

Distances, such as the Euclidean distance, have some well known properties.

1. \( d(p, q) \geq 0 \) for all \( p \) and \( q \) and \( d(p, q) = 0 \) only if \( p = q \). (Positive definiteness)
2. \( d(p, q) = d(q, p) \) for all \( p \) and \( q \). (Symmetry)
3. \( d(p, r) \leq d(p, q) + d(q, r) \) for all points \( p, q, \) and \( r \). (Triangle Inequality)

where \( d(p, q) \) is the distance (dissimilarity) between points (data objects), \( p \) and \( q \).

A distance that satisfies these properties is a metric.
Common Properties of a Similarity

- Similarities, also have some well known properties.

1. \( s(p, q) = 1 \) (or maximum similarity) only if \( p = q \).

2. \( s(p, q) = s(q, p) \) for all \( p \) and \( q \). (Symmetry)

where \( s(p, q) \) is the similarity between points (data objects), \( p \) and \( q \).
Similarity Between Binary Vectors

• Common situation is that objects, $p$ and $q$, have only binary attributes

• Compute similarities using the following quantities
  
  $M_{01}$ = the number of attributes where $p$ was 0 and $q$ was 1
  $M_{10}$ = the number of attributes where $p$ was 1 and $q$ was 0
  $M_{00}$ = the number of attributes where $p$ was 0 and $q$ was 0
  $M_{11}$ = the number of attributes where $p$ was 1 and $q$ was 1

• Simple Matching and Jaccard Coefficients
  
  $s_{SMC} = \frac{\text{number of matches}}{\text{number of attributes}}$
  
  $s_{SMC} = \frac{M_{11} + M_{00}}{M_{01} + M_{10} + M_{11} + M_{00}}$

  $s_J = \frac{\text{number of 11 matches}}{\text{number of not-both-zero attributes values}}$
  
  $s_J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$

  Jaccard ignores 0s!
SMC versus Jaccard: Example

\[ p = 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \]
\[ q = 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 1 \]

\[ M_{01} = 2 \] (the number of attributes where \( p \) was 0 and \( q \) was 1)
\[ M_{10} = 1 \] (the number of attributes where \( p \) was 1 and \( q \) was 0)
\[ M_{00} = 7 \] (the number of attributes where \( p \) was 0 and \( q \) was 0)
\[ M_{11} = 0 \] (the number of attributes where \( p \) was 1 and \( q \) was 1)

\[ s_{SMC} = (M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00}) = (0+7) / (2+1+0+7) = 0.7 \]

\[ s_J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0 \]
Extended Jaccard Coefficient (Tanimoto)

Variation of Jaccard for continuous or count attributes:

\[ T(p, q) = \frac{p \cdot q}{\|p\|^2 + \|q\|^2 - p \cdot q} \]

where \( \cdot \) is the dot product between two vectors and \( \|\cdot\|^2 \) is the Euclidean norm (length of the vector).

Reduces to Jaccard for binary attributes
Exercise

<table>
<thead>
<tr>
<th></th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- Calculate the Euclidean and the Manhattan distances between A and C and A and B
- Calculate the Cosine similarity between A and C and A and B
Dis(similarities) With Mixed Types

- Sometimes attributes are of many different types (nominal, ordinal, ratio, etc.), but an overall similarity is needed.

- Gower's (dis)similarity:
  - Ignores missing values
  - Final (dis)similarity is a weighted sum of variable-wise (dis)similarities

1. For the $k^{th}$ attribute, compute a similarity, $s_k$, in the range $[0, 1]$.

2. Define an indicator variable, $\delta_k$, for the $k^{th}$ attribute as follows:

\[
\delta_k = \begin{cases} 
0 & \text{if the } k^{th} \text{ attribute is a binary asymmetric attribute and both objects have a value of 0, or if one of the objects has a missing values for the } k^{th} \text{ attribute} \\
1 & \text{otherwise}
\end{cases}
\]

3. Compute the overall similarity between the two objects using the following formula:

\[
similarity(p, q) = \frac{\sum_{k=1}^{n} \delta_k s_k}{\sum_{k=1}^{n} \delta_k}
\]
Correlation

- Correlation measures the linear relationship between two variables.
- To compute **Pearson** correlation (Pearson's Product Moment Correlation), we standardize data objects, \( p \) and \( q \), and then take their dot product.

\[
\rho = \frac{\text{cov}(X, Y)}{sd(X)sd(Y)}
\]

\[
r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sum (y_i - \bar{y})^2}
\]

- Correlation is often used as a measure of similarity.
Visually Evaluating Correlation

Scatter plots showing the similarity from \(-1\) to \(1\).
Rank Correlation

- Measure the degree of similarity between two ratings (e.g., ordinal data).
- Is more robust against outliers and does not assume normality of data like Pearson Correlation.
- Measures (all are between -1 and 1)
  - **Spearman's Rho**: Pearson correlation between ranked variables.
  - **Kendall's Tau**
    \[
    \tau = \frac{N_s - N_d}{\frac{1}{2}n(n-1)}
    \]
    - **Goodman and Kruskal's Gamma**
    \[
    \gamma = \frac{N_s - N_d}{N_s + N_d}
    \]
Topics

• Attributes/Features
• Types of Data Sets
• Data Quality
• Data Preprocessing
• Similarity and Dissimilarity
• Density
Density

- Density-based clustering require a notion of density

- **Examples:**
  - **Euclidean density** = number of points per unit volume
  - **Probability density** (function) = describes the likelihood of a random variable taking a given value
  - **Graph-based density** = number of edges compared to a complete graph
Euclidean Density - Cell-based

- Simplest approach is to divide region into a number of rectangular cells of equal volume and define density as # of points the cell contains

![Graph showing cell-based density](image)

**Figure 7.13.** Cell-based density.

**Table 7.6.** Point counts for each grid cell.
Euclidean Density - Center-based

- Euclidean density is the number of points within a specified radius of the point

Figure 7.14. Illustration of center-based density.