What is Data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- 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.

```
5  _______A________  1
6  ________B______  2
8  ________C______  3
10 __________D____  4
15 __________E____  5
```
Types of Attributes

- 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

Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: \( = \neq \)
  - Order: \(<\ >\)
  - Addition: \(+\ -\)
  - Multiplication: \(*\ /\)
  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & addition
  - Ratio attribute: all 4 properties
<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. ($\neq$, $\approx$)</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. ($\times$, $/$)</td>
<td>temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current</td>
<td>geometric mean, harmonic mean, percent variation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attribute Level</th>
<th>Transformation</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>Any permutation of values</td>
<td>If all employee ID numbers were reassigned, would it make any difference?</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>
</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>
</tr>
<tr>
<td>Ratio</td>
<td>$new_value = a \times old_value$</td>
<td>Length can be measured in meters or feet.</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.

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
Important Characteristics of Structured Data

- Dimensionality
  - Curse of Dimensionality

- Sparsity
  - Only presence counts

- Resolution
  - Patterns depend on the scale

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>Projection of x Load</th>
<th>Projection of y load</th>
<th>Distance</th>
<th>Load</th>
<th>Thickness</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.23</td>
<td>5.27</td>
<td>15.22</td>
<td>2.7</td>
<td>1.2</td>
</tr>
<tr>
<td>12.65</td>
<td>6.25</td>
<td>16.22</td>
<td>2.2</td>
<td>1.1</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>team</th>
<th>coach</th>
<th>year</th>
<th>play</th>
<th>score</th>
<th>season</th>
<th>in</th>
<th>at</th>
<th>home</th>
<th>lost</th>
<th>timeout</th>
<th>season</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document 1</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Document 2</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>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Document 3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</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

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

- Benzene Molecule: $C_6H_6$

![Benzene Molecule Diagram]

**Ordered Data**

- Sequences of transactions

```
( A B)  (D)  (C E)  
( B D)  (C)  (E)   
( C D)  (B)  (A E) 
```

An element of the sequence
Ordered Data

- Genomic sequence data

  GGTTCCGCCTCAGCCCCGCGCC
  CGCAGGGCCGCCGCCGCGCCGTC
  GAGAAGGGCCCGCTGCGGGCG
  GGGGAGGGCCGCCGCACCGACG
  CCAACCGAGTCCGACCAGGTGCC
  CCTCTGCTCGCTAGACCTGAG
  GCCAAGTAGAACACGCGAAGC
  TGGGCTGCCTGCCTGACAGCG

Ordered Data

- Spatio-Temporal Data

  Average Monthly Temperature of land and ocean
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 and “snow” on television screen

![Two Sine Waves](image1.png)  ![Two Sine Waves + Noise](image2.png)
Outliers

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

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
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (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

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

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.

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

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

- 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.
    - In sampling with replacement, the same object can be picked up more than once

- 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.

- 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 that captures the largest amount of variation in data

![Diagram of PCA](chart.png)
**Dimensionality Reduction: PCA**

- Find the eigenvectors of the covariance matrix
- The eigenvectors define the new space

**Dimensionality Reduction: ISOMAP**

By: Tenenbaum, de Silva, Langford (2000)

- Construct a neighbourhood graph
- For each pair of points in the graph, compute the shortest path distances – geodesic distances
Dimensionality Reduction: PCA

Dimensions = 206

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
  - 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

- Techniques:
  - 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
  - Filter approaches:
    - Features are selected before data mining algorithm is run
  - Wrapper approaches:
    - Use the data mining algorithm as a black box to find best subset of attributes

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
  - Mapping Data to New Space
  - Feature Construction
    - combining features
Mapping Data to a New Space

- Fourier transform
- Wavelet transform

Discretization Using Class Labels

- Entropy based approach

3 categories for both x and y
5 categories for both x and y
**Discretization Without Using Class Labels**

- Data Equal interval width
- Equal frequency K-means

**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
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

\[ dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2} \]

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

- Standardization is necessary, if scales differ.

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>

<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>
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

- \( r \rightarrow \infty \). “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

\begin{align*}
\text{L}_1 & : \quad \| p_1 \| = 0, \quad \| p_2 \| = 4, \quad \| p_3 \| = 2, \quad \| p_4 \| = 6 \\
\text{L}_2 & : \quad \| p_1 \| = 0, \quad \| p_2 \| = 2.828, \quad \| p_3 \| = 3.162, \quad \| p_4 \| = 5.099 \\
\text{L}_\infty & : \quad \| p_1 \| = 0, \quad \| p_2 \| = 2, \quad \| p_3 \| = 3, \quad \| p_4 \| = 5
\end{align*}

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>

Mahalanobis Distance

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

\[ \Sigma \text{ is the covariance matrix of the input data } X \]

\[ \Sigma_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j) \]

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)

Mahal(A,B) = 5
Mahal(A,C) = 4

### 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
  $SMC = \frac{\text{number of matches}}{\text{number of attributes}}$
  $= \frac{(M_{11} + M_{00})}{(M_{01} + M_{10} + M_{11} + M_{00})}$

  $J = \frac{\text{number of 11 matches}}{\text{number of not-both-zero attributes values}}$
  $= \frac{(M_{11})}{(M_{01} + M_{10} + M_{11})}$
SMC versus Jaccard: Example

\[
p = 10000000000 \\
q = 0000001001
\]

- \(M_{01} = 2\) (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)

\[
SMC = \frac{M_{11} + M_{00}}{M_{01} + M_{10} + M_{11} + M_{00}} = \frac{0 + 7}{2 + 1 + 0 + 7} = 0.7
\]

\[
J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} = \frac{0}{2 + 1 + 0} = 0
\]

Cosine Similarity

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

- Example:
  \[
  d_1 = 3205000200 \\
  d_2 = 1000000102
  \]

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

\[
||d_1|| = \sqrt{(3^2+2^2+0^2+0^2+5^2+0^2+0^2+0^2+0^2+2^2+0^2+0^2)} = \sqrt{42} = 6.481
\]

\[
||d_2|| = \sqrt{(1^2+1^2+0^2+0^2+0^2+0^2+0^2+0^2+0^2+1^2+0^2+2^2)} = \sqrt{6} = 2.245
\]

\[
\cos(d_1, d_2) = 0.3150
\]
**Extended Jaccard Coefficient (Tanimoto)**

- Variation of Jaccard for continuous or count attributes
  - Reduces to Jaccard for binary attributes

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

**Correlation**

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, p and q, and then take their dot product

\[
p_k' = \frac{p_k - \text{mean}(p)}{\text{std}(p)}
\]

\[
q_k' = \frac{q_k - \text{mean}(q)}{\text{std}(q)}
\]

\[
\text{correlation}(p, q) = p' \cdot q'
\]
**Visually Evaluating Correlation**

Scatter plots showing the similarity from –1 to 1.

**General Approach for Combining Similarities**

- Sometimes attributes are of many different types, but an overall similarity is needed.

1. For the k\textsuperscript{th} attribute, compute a similarity, \( s_k \), in the range \([0, 1]\).
2. Define an indicator variable, \( \delta_k \), for the k\textsuperscript{th} attribute as follows:

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

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

\[
similarity(p, q) = \frac{\sum_{k=1}^{n} \delta_k s_k}{\sum_{k=1}^{n} \delta_k}
\]
Using Weights to Combine Similarities

• May not want to treat all attributes the same.
  – Use weights $w_k$ which are between 0 and 1 and sum to 1.

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

\[
distance(p, q) = \left( \sum_{k=1}^{n} w_k |p_k - q_k|^r \right)^{1/r}
\]

Density

• Density-based clustering require a notion of density

• Examples:
  – Euclidean density
    • Euclidean density = number of points per unit volume
  – Probability density
  – Graph-based density
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.

![Figure 7.13. Cell-based density.](image)

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>11</td>
<td>18</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>14</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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.](image)