

# Introduction to Data Mining

## Chapter 2 Data

by Michael Hahsler

Based in Slides by Tan,  
Steinbach, Karpadne, Kumar



This work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International](https://creativecommons.org/licenses/by-sa/4.0/) license



## R Code Examples

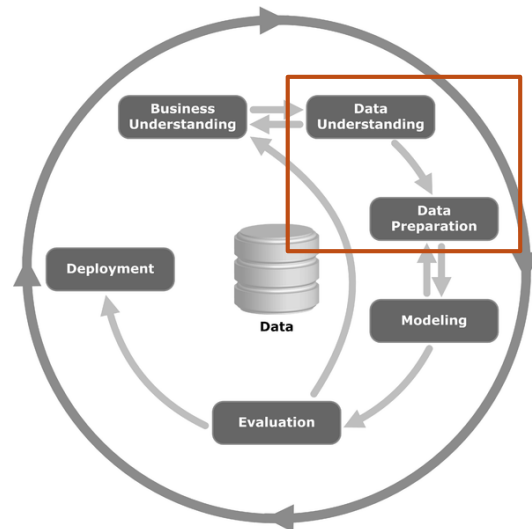
- Available R Code examples are indicated on slides by the R logo



- The Examples are available at [https://mhahsler.github.io/Introduction\\_to\\_Data\\_Mining\\_R\\_Examples/](https://mhahsler.github.io/Introduction_to_Data_Mining_R_Examples/)



## Tasks in the CRISP-DM Reference Model



## 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

- 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

| Objects | Attributes |        |                |                |       |
|---------|------------|--------|----------------|----------------|-------|
|         | Tid        | Refund | Marital Status | Taxable Income | Cheat |
|         | 1          | Yes    | Single         | 125K           | No    |
|         | 2          | No     | Married        | 100K           | No    |
|         | 3          | No     | Single         | 70K            | No    |
|         | 4          | Yes    | Married        | 120K           | No    |
|         | 5          | No     | Divorced       | 95K            | Yes   |
|         | 6          | No     | Married        | 60K            | No    |
|         | 7          | Yes    | Divorced       | 220K           | No    |
|         | 8          | No     | Single         | 85K            | Yes   |
|         | 9          | No     | Married        | 75K            | No    |
|         | 10         | No     | Single         | 90K            | Yes   |

| Attribute Type | Description                                                                                                                                                                    | Examples                                                                                  | Operations                                                                           |
|----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|
| Nominal        | The values of a nominal attribute are just different <b>names or labels</b> , i.e., nominal attributes provide only enough information to distinguish one object from another. | zip codes, employee ID numbers, eye color, sex: {male, female}                            | =, ≠<br><br>mode, entropy, contingency correlation, $\chi^2$ test                    |
| Ordinal        | The values of an ordinal attribute provide enough information to <b>order objects</b> .                                                                                        | zip codes, employee ID numbers, eye color, sex: {male, female}                            | Nominal + <, ><br><br>median, percentiles, rank correlation, run tests, sign tests   |
| Interval       | For interval attributes, the <b>differences between values</b> are meaningful, i.e., a unit of measurement exists.                                                             | calendar dates, temperature in Celsius or Fahrenheit                                      | Ordinal + +, -<br><br>mean, standard deviation, Pearson's correlation, t and F tests |
| Ratio          | For ratio variables, both differences and <b>ratios are meaningful</b> . Double the number means twice as much.                                                                | temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current | Interval + *, /<br><br>geometric mean, harmonic mean, percent variation              |

## 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

Categorical, Qualitative

Quantitative

## Discrete and Continuous Attributes

### Discrete Attribute

- Has only a finite or countably infinite set of values.
- Binary attributes are a special case of discrete attributes.
- Examples:** zip codes, counts, or the set of words in a collection of documents
- Representation:** Strings or integer variables (enumeration type).

### Continuous Attribute

- Has real numbers as attribute values
- Examples:** temperature, height, or weight.
- Representation:** floating-point variables. Computers represent real numbers in a discrete format with a finite number of digits.

## Examples

- What is the scale of measurement of:
  - Number of cars per minute (count data)
  - Age data grouped in:  
0-4 years, 5-9, 10-14, ...
  - Age data grouped in:  
<20 years, 21-30, 31-40, 41+



## 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

## Topics

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



## Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes (e.g., from a relational database)

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|-----|--------|----------------|----------------|-------|
| 1   | Yes    | Single         | 125K           | No    |
| 2   | No     | Married        | 100K           | No    |
| 3   | No     | Single         | 70K            | No    |
| 4   | Yes    | Married        | 120K           | No    |
| 5   | No     | Divorced       | 95K            | Yes   |
| 6   | No     | Married        | 60K            | No    |
| 7   | Yes    | Divorced       | 220K           | No    |
| 8   | No     | Single         | 85K            | Yes   |
| 9   | No     | Married        | 75K            | No    |
| 10  | No     | Single         | 90K            | Yes   |

## 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

n attributes

|           | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width |
|-----------|--------------|-------------|--------------|-------------|
| m objects | 5.6          | 2.7         | 4.2          | 1.3         |
|           | 6.5          | 3.0         | 5.8          | 2.2         |
|           | 6.8          | 2.8         | 4.8          | 1.4         |
|           | 5.7          | 3.8         | 1.7          | 0.3         |
|           | 5.5          | 2.5         | 4.0          | 1.3         |
|           | 4.8          | 3.0         | 1.4          | 0.1         |
|           | 5.2          | 4.1         | 1.5          | 0.1         |



## 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.

Terms

|            | team | coach | play | ball | score | game | win | lost | timeout | season |
|------------|------|-------|------|------|-------|------|-----|------|---------|--------|
| Document 1 | 3    | 0     | 5    | 0    | 2     | 6    | 0   | 2    | 0       | 2      |
| Document 2 | 0    | 7     | 0    | 2    | 1     | 0    | 0   | 3    | 0       | 0      |
| Document 3 | 0    | 1     | 0    | 0    | 1     | 2    | 2   | 0    | 3       | 0      |

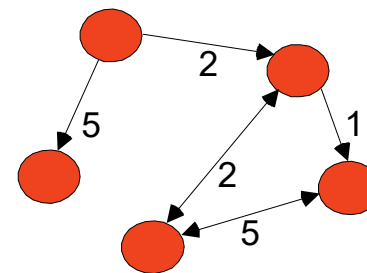
## Transaction Data

- 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.

| TID | Items                     |
|-----|---------------------------|
| 1   | Bread, Coke, Milk         |
| 2   | Beer, Bread               |
| 3   | Beer, Coke, Diaper, Milk  |
| 4   | Beer, Bread, Diaper, Milk |
| 5   | Coke, Diaper, Milk        |

## Graph Data

- Examples: Generic graph and HTML Links

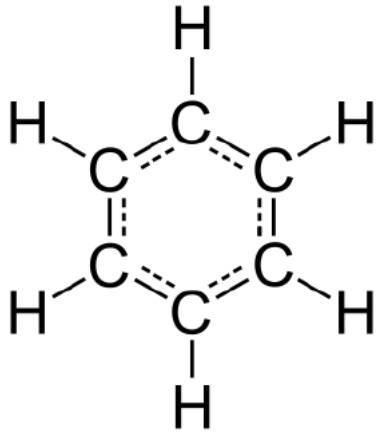


```

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

## Chemical Data

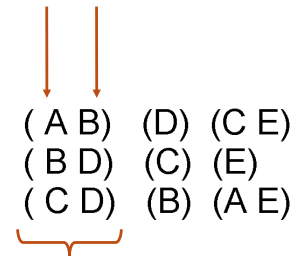
- Benzene Molecule: C<sub>6</sub>H<sub>6</sub>



## Ordered Data

- Sequences of transactions

Items/Events



An element of the sequence

## Ordered Data

- Genomic sequence data

GGTTCCGCCTTCAGCCCCGCGCC  
CGCAGGGCCCCGCCCCGCGCCGTC  
GAGAAGGGCCCGCCTGGCGGGCG  
GGGGAGGCGGGGCCCGCCGAGC  
CCAACCGAGTCCGACCAGGTGCC  
CCCTCTGCTCGGCCTAGACCTGA  
GCTCATTAGGCGGCAGCGGACAG  
GCCAAGTAGAACACGCGAAGCGC  
TGGGCTGCCTGCTGCGACCAGGG

Subsequences

## Ordered Data: Time Series Data

S&P 500 Index

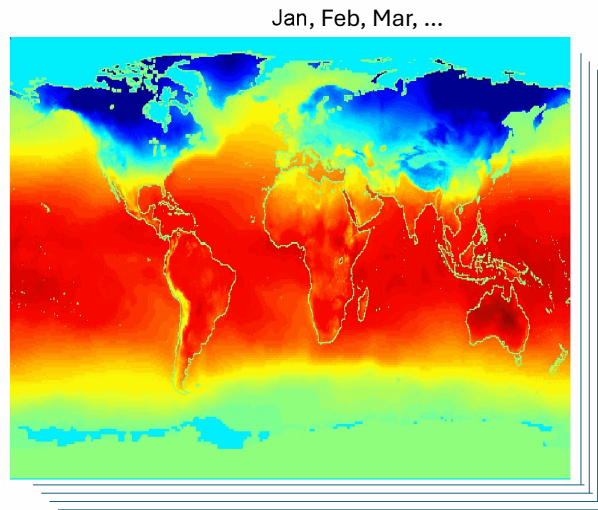
April 1, 2016 – March 31, 2017



Source: FactSet

## Ordered Data: Spatio-Temporal

**Average Monthly  
Temperature of  
land and ocean**



## Topics

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

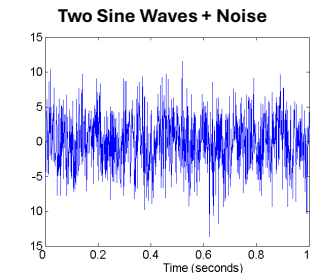
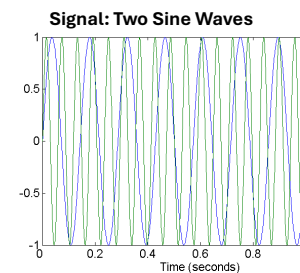


## Data Quality

- What kinds of data quality problems exist?
  - Noise and outliers
  - Missing values
  - Duplicate data
- How can we detect problems with the data?
  - Statistics
  - Visualization
- What can we do about these problems?
  - Mark value as missing
  - Remove object

## Noise

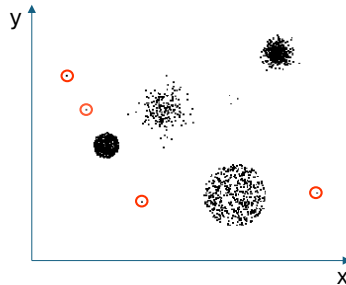
- Noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone, “snow” on television screen, measurement errors.



- Find less noisy data
- Sometimes we can de-noise (signal processing)

## Outliers

- Outliers are data objects with feature values that are considerably different than most of the other data objects in the data set.
- Reasons:
  - A true outlier is a special object (e.g., a genius's IQ score).
  - May be the result of a measurement mistake.



- Typical treatment: Statistical outlier detection +
  - Make outlying feature missing, or
  - Remove the complete outlier object

## 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)
  - Value was a mistake and set to missing.
- 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)

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|-----|--------|----------------|----------------|-------|
| 1   | Yes    | Single         | 125K           | No    |
| 2   | No     | Married        | 100K           | No    |
| 3   | No     | Single         | NA             | No    |
| 4   | Yes    | Married        | 120K           | No    |
| 5   | No     | Divorced       | 95K            | Yes   |
| 6   | No     | Married        | 60K            | No    |
| 7   | Yes    | Divorced       | 220K           | No    |
| 8   | No     | Single         | 85K            | Yes   |
| 9   | No     | Married        | 75K            | No    |
| 10  | No     | Single         | 90K            | Yes   |

## Duplicate Data

- Data set may include data objects that are duplicates, or "close 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

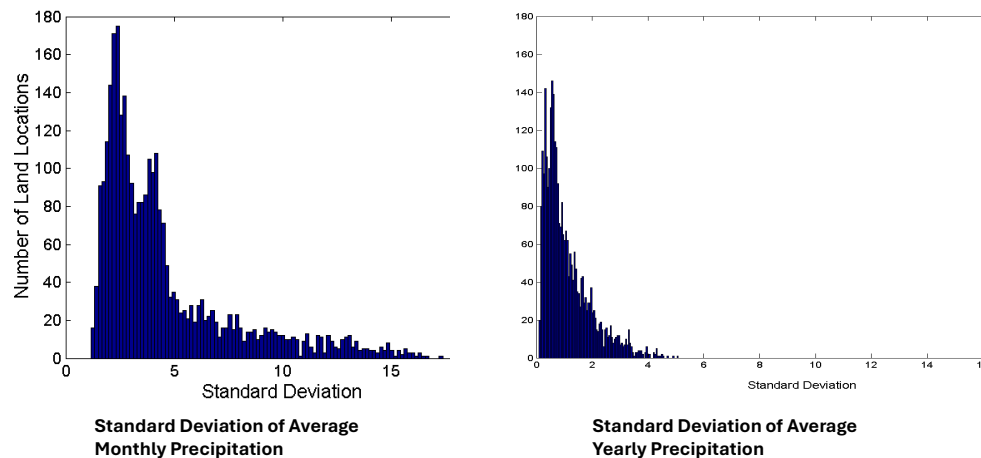
## Aggregation

- 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



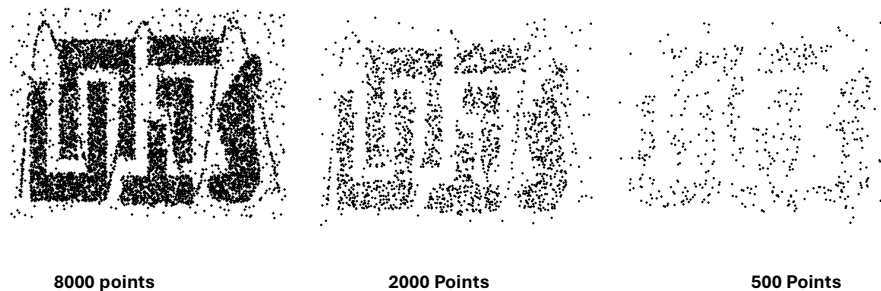
## Sampling

- 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 (e.g., does not fit into memory or is too slow).

## 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.

## Sample Size



## Types of Sampling

### Replacement?

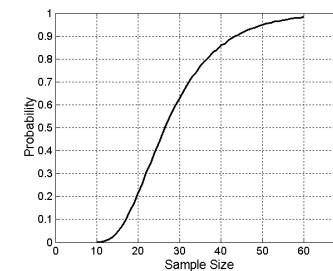
- **Sampling without replacement**  
As each item is selected, it is removed from the population.
- **Sampling with replacement**  
Objects are not removed from the population as they are selected for the sample. Note: the same object can be picked up more than once.

### Selection?

- **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

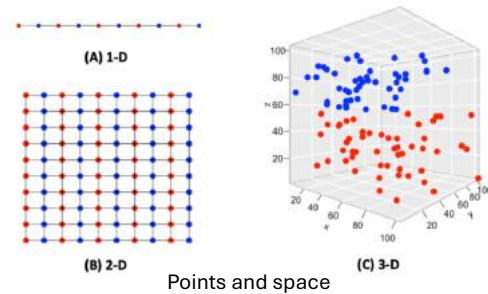
- What sample size is necessary to get at least one object from each of 10 groups.



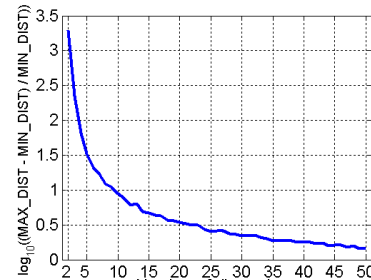
- Sample size determination:
  - Statistics: confidence interval for parameter estimate or desired statistical power of test.
  - Machine learning: often more is better, cross-validated accuracy.

## Curse of Dimensionality

- When dimensionality increases, the size of the data space grows exponentially.
- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful
  - Density  $\rightarrow 0$
  - All points tend to have the same Euclidean distance to each other.

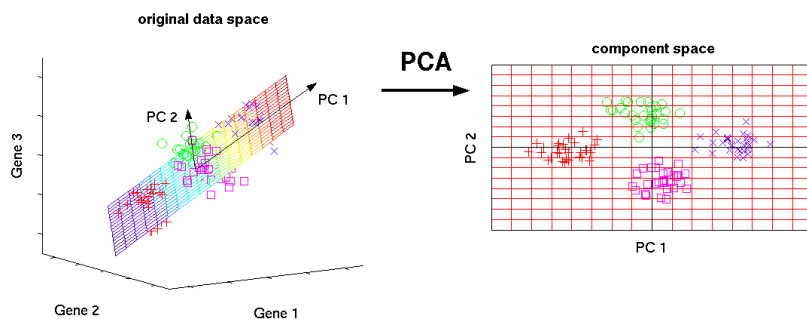


**Experiment:** Randomly generate 500 points. Compute difference between max and min distance between any pair of points



## Dimensionality Reduction: Principal Components Analysis (PCA)

- Goal:** Map points to a lower dimensional space while preserving distance information.

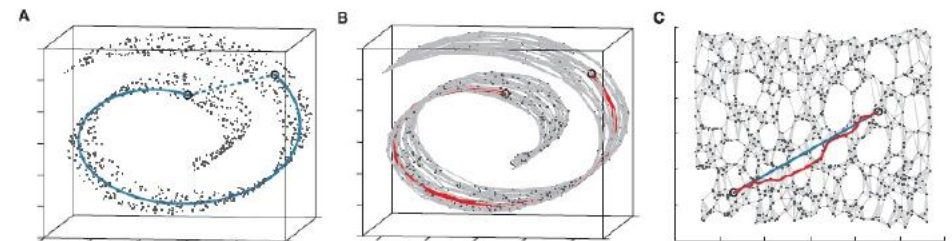


- Method:** Find a projection (new axes) that captures the largest amount of variation in data. This can be done using eigenvectors of the covariance matrix or SVD (singular value decomposition).

## 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: ISOMAP

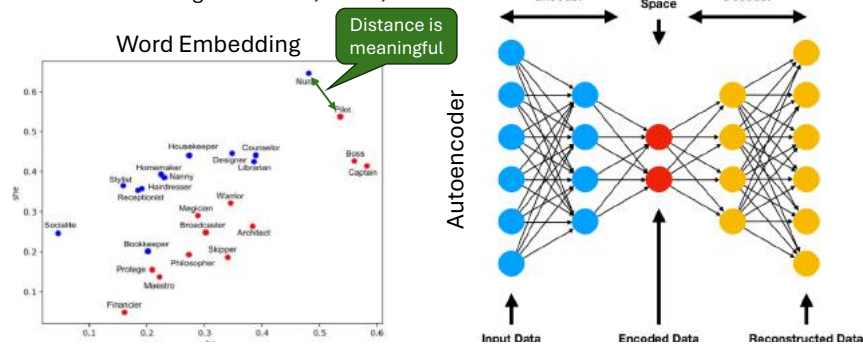


- Goal:** Unroll the “swiss roll!” (i.e., preserve distances on the roll)
- Method:** Use a non-metric space, i.e., distances are not measured by Euclidean distance, but along the surface of the roll (geodesic distances).
  - Construct a neighbourhood graph (k-nearest neighbors or within a radius).
  - For each pair of points in the graph, compute the shortest path distances = geodesic distances.
  - Create a lower dimensional embedding using the geodesic distances (multi-dimensional scaling; MDS)



## Low-dimensional Embedding

- General notion of representing objects described in one space (i.e., set of features) in a different space using a map  $f: X \rightarrow Y$
- PCA is an example where  $Y$  is the space spanned by the principal components and objects close in the original space  $X$  are embedded in space  $Y$ .
- Low-dimensional embeddings can be produced with various other methods:
  - T-SNE: T-distributed Stochastic Neighbor Embedding; non-linear for visualization of high-dimensional datasets.
  - Autoencoders (deep learning): non-linear
  - Word embedding: Word2vec, GloVe, BERT

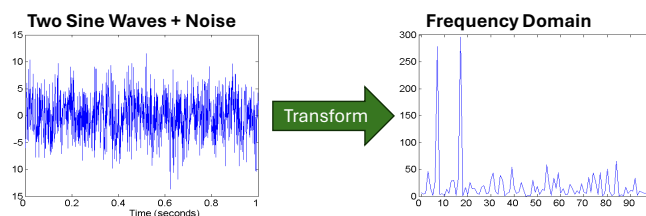


## Feature Creation

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 / Feature Engineering
  - Combining features (interactions: multiply features)
  - Example: Calculate the body mass index from height and weight
- Mapping Data to New Space
  - Example: Fourier transform/Wavelet transform



## Feature Subset Selection

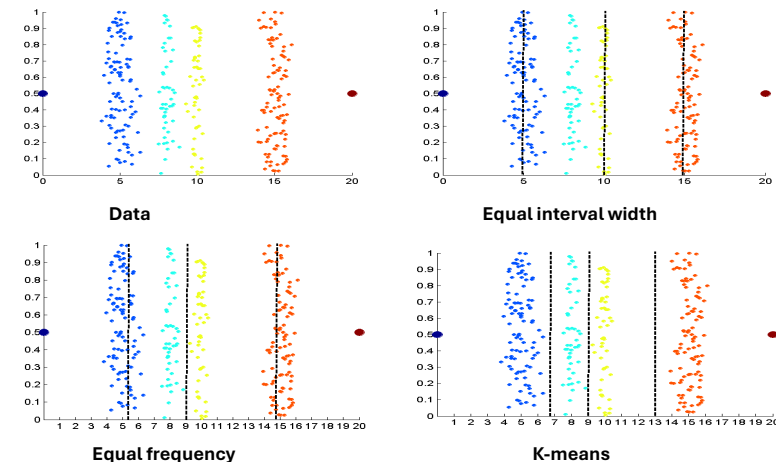
= Remove features (columns):

- Redundant features
  - duplicate information contained in multiple features (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
  - Example: students' ID is often irrelevant to the task of predicting students' GPA

### Methods

- 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)
- Brute-force approach:
  - Try all possible feature subsets as input to data mining algorithm and choose the best.
- Wrapper approaches:
  - Use the data mining algorithm as a black box to find best subset of attributes (often using greedy search)

## Mapping Data to New Space: Unsupervised Discretization



## Attribute Transformation: Normalization

- 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
    - The z-score normalizes data roughly to an interval of  $[-3,3]$ .

$$x' = \frac{x - \bar{x}}{s_x}$$

$\bar{x}$  ... column (attribute) mean

$s_x$  ... column (attribute) standard deviation



## 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

## Topics

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



## Similarity/Dissimilarity for Simple Attributes

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

| Attribute Type    | Dissimilarity                                                                                         | Similarity                                                                      |
|-------------------|-------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------|
| Nominal           | $d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$                       | $s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$ |
| Ordinal           | $d = \frac{ p-q }{n-1}$<br>(values mapped to integers 0 to $n-1$ , where $n$ is the number of values) | $s = 1 - \frac{ p-q }{n-1}$                                                     |
| Interval or Ratio | $d =  p - q $                                                                                         | $s = -d, s = \frac{1}{1+d}$ or $s = 1 - \frac{d - \min.d}{\max.d - \min.d}$     |

$$s = f(d)$$

$f$  can be any strictly decreasing function.

## Euclidean Distance

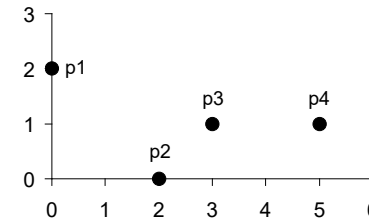
| point | x | y |
|-------|---|---|
| p     | 0 | 2 |
| q     | 2 | 0 |

- Euclidean Distance (for quantitative attribute vectors)

$$d_E = \sqrt{\sum_{k=1}^n (p_k - q_k)^2} = \|\mathbf{p} - \mathbf{q}\|_2$$

- Where  $\mathbf{p}$  and  $\mathbf{q}$  are two objects represented by vectors.  $n$  is the number of dimensions (attributes) of the vectors and  $p_k$  and  $q_k$  are, respectively, the  $k$ th attributes (components) or data objects  $p$  and  $q$ .
- $\|\cdot\|_2$  is the  $L^2$  vector norm (i.e., length of a vector in Euclidean space).
- Note:** If ranges differ between components of  $\mathbf{p}$  then standardization (z-scores) is necessary to avoid one variable to dominate the distance.

## Euclidean Distance



| point | x | y |
|-------|---|---|
| p1    | 0 | 2 |
| p2    | 2 | 0 |
| p3    | 3 | 1 |
| p4    | 5 | 1 |

|    | p1   | p2   | p3   | p4   |
|----|------|------|------|------|
| p1 | 0.00 | 2.83 | 3.16 | 5.10 |
| p2 | 2.83 | 0.00 | 1.41 | 3.16 |
| p3 | 3.16 | 1.41 | 0.00 | 2.00 |
| p4 | 5.10 | 3.16 | 2.00 | 0.00 |

Distance Matrix

## Minkowski Distance

| point | x | y |
|-------|---|---|
| p     | 0 | 2 |
| q     | 2 | 0 |

- Minkowski Distance is a generalization of Euclidean Distance

$$d_M = \left( \sum_{k=1}^n |p_k - q_k|^r \right)^{\frac{1}{r}} = \|\mathbf{p} - \mathbf{q}\|_r$$

- Where  $\mathbf{p}$  and  $\mathbf{q}$  are two objects represented by vectors.  $n$  is the number of dimensions (attributes) of the vectors and  $p_k$  and  $q_k$  are, respectively, the  $k$ th attributes (components) or data objects  $p$  and  $q$ .
- Note:** If ranges differ then standardization (z-scores) is necessary to avoid one variable to dominate the distance.

## 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 ( $L^2$  norm)
- $r = \infty$ . “supremum” (maximum 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 Distances

Distance Matrix

| point | x | y |
|-------|---|---|
| p1    | 0 | 2 |
| p2    | 2 | 0 |
| p3    | 3 | 1 |
| p4    | 5 | 1 |

| $L^1$ | p1 | p2 | p3 | p4 |
|-------|----|----|----|----|
| p1    | 0  | 4  | 4  | 6  |
| p2    | 4  | 0  | 2  | 4  |
| p3    | 4  | 2  | 0  | 2  |
| p4    | 6  | 4  | 2  | 0  |

| $L^2$ | p1   | p2   | p3   | p4   |
|-------|------|------|------|------|
| p1    | 0.00 | 2.83 | 3.16 | 5.10 |
| p2    | 2.83 | 0.00 | 1.41 | 3.16 |
| p3    | 3.16 | 1.41 | 0.00 | 2.00 |
| p4    | 5.10 | 3.16 | 2.00 | 0.00 |

| $L^\infty$ | p1 | p2 | p3 | p4 |
|------------|----|----|----|----|
| p1         | 0  | 2  | 3  | 5  |
| p2         | 2  | 0  | 1  | 3  |
| p3         | 3  | 1  | 0  | 2  |
| p4         | 5  | 3  | 2  | 0  |



## Cosine Similarity

For two vectors  $A$  and  $B$ , the cosine similarity is defined as

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

Example:

$$A = [3 \ 2 \ 0 \ 5 \ 0 \ 0 \ 0 \ 2 \ 0 \ 0]$$

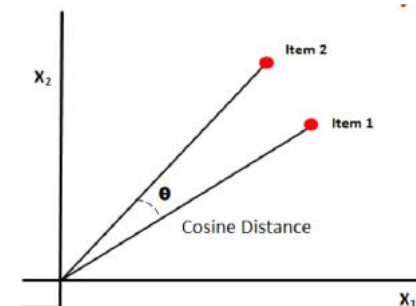
$$B = [1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 2]$$

$$A \cdot B = 3 \cdot 1 + 2 \cdot 0 + 0 \cdot 0 + 5 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 2 \cdot 1 + 0 \cdot 0 + 0 \cdot 2 = 5$$

$$\|A\| = (3^2 + 2^2 + 0^2 + 5^2 + 0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2)^{0.5} = 6.481$$

$$\|B\| = (1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 0^2 + 2^2)^{0.5} = 2.245$$

$$s_{\cosine} = .3150$$



**Information retrieval:** Cosine similarity is often used for **word count vectors** to compare documents. It compares the distribution of words.

## Similarity Between Binary Vectors

- A common situation is that objects,  $p$  and  $q$ , have only binary attributes

- Compute similarities using the following quantities

$f_{01}$  = the number of attributes where  $p$  was 0 and  $q$  was 1

$f_{10}$  = the number of attributes where  $p$  was 1 and  $q$  was 0

$f_{00}$  = the number of attributes where  $p$  was 0 and  $q$  was 0

$f_{11}$  = the number of attributes where  $p$  was 1 and  $q$  was 1

- Simple Matching and Jaccard Coefficients**

$s_{SMC}$  = number of matches / number of attributes

$$= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$$

$s_J$  = number of 11 matches / number of not-both-zero attribute values

$$= (f_{11}) / (f_{01} + f_{10} + f_{11})$$

**Note:** Jaccard ignores 0s!

## SMC versus Jaccard: Example

$$p = [1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]$$

$$q = [0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 1]$$

$f_{01} = 2$  (the number of attributes where  $p$  was 0 and  $q$  was 1)

$f_{10} = 1$  (the number of attributes where  $p$  was 1 and  $q$  was 0)

$f_{00} = 7$  (the number of attributes where  $p$  was 0 and  $q$  was 0)

$f_{11} = 0$  (the number of attributes where  $p$  was 1 and  $q$  was 1)

$$s_{SMC} = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}} = (0 + 7) / (2 + 1 + 0 + 7) = 0.7$$

$$s_J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} = 0 / (2 + 1 + 0) = 0$$

## 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
- Calculation:

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:

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

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

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



## Common Properties of a Similarity

- Similarities, also have some well-known properties.

$$s(p, q) = 1 \text{ (or maximum similarity) only if } p = q.$$

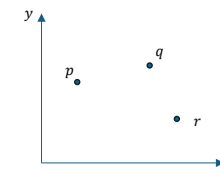
$$s(p, q) = s(q, p) \text{ for all } p \text{ and } q. \text{ (Symmetry)}$$

where  $s(p, q)$  is the similarity between points (data objects),  $p$  and  $q$ .

## Common Properties of a Distance

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

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



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

- A distance that satisfies these properties is a **metric** and forms a **metric space**.
- **Note:** Some **dissimilarities** violate property 3 (Triangular inequality). They are not a metric and some people would not call them a distance.

## Exercise

|   | x | y |
|---|---|---|
| A | 2 | 1 |
| B | 4 | 3 |
| C | 1 | 1 |

- Manually calculate the following using the equations in the slides:
  - 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
- Check your results by using the dist function in R. You will need to Google to find out how to compute the Cosine similarity in R.

## Topics

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

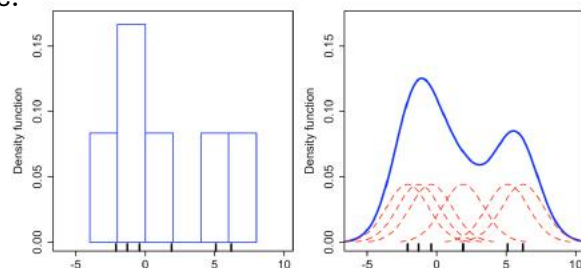


## Kernel Density Estimation (KDE)

- KDE is a non-parametric way to estimate the probability density function of a random variable.

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$

- $K$  is the kernel (a non-negative function that integrates to one) and  $h > 0$  is a smoothing parameter called the bandwidth. Often a Gaussian kernel is used.
- Example:



## Density

- Density-based clustering requires a notion of density
- Examples:
  - Probability density (function) = describes the likelihood of a random variable taking a given value
  - Euclidean density = number of points per unit volume
  - ~~Graph-based density = number of edges compared to a complete graph~~
  - ~~Density of a matrix = proportion of non-zero entries.~~

## Euclidean Density – Cell-based

- A Simple approach is to divide region into rectangular cells of equal volume and define density as # of points in each cell.

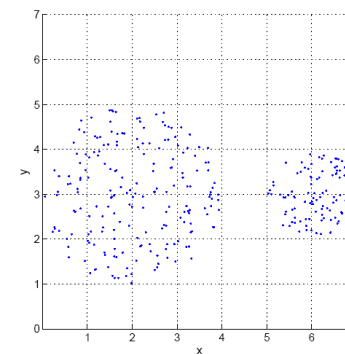


Figure 7.13. Cell-based density.

|    |    |    |    |   |    |    |
|----|----|----|----|---|----|----|
| 0  | 0  | 0  | 0  | 0 | 0  | 0  |
| 0  | 0  | 0  | 0  | 0 | 0  | 0  |
| 4  | 17 | 18 | 6  | 0 | 0  | 0  |
| 14 | 14 | 13 | 13 | 0 | 18 | 27 |
| 11 | 18 | 10 | 21 | 0 | 24 | 31 |
| 3  | 20 | 14 | 4  | 0 | 0  | 0  |
| 0  | 0  | 0  | 0  | 0 | 0  | 0  |

Table 7.6. Point counts for each grid cell.

## Euclidean Density – Center-based

- Euclidean density is the number of points within a specified radius around the point. This is also called the **neighborhood** of the point.

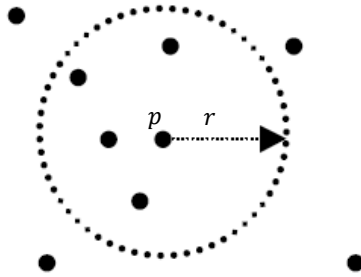


Figure 7.14. Illustration of center-based density.



### You should know now about...

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

