Supervised Learning

Chapter II: Data and Preprocessing

Johannes Jurgovsky

Outline

Data and Preprocessing

1. Objects and Attributes
2. Types of Data Sets
3. Preprocessing

1. Objects and Attributes

Objects and Attributes:

- An object $o \in O$ is described by a set of attributes.
  An object is also known as record, point, case, sample, entity, or instance.
- An attribute A is a property of an object.
  An attribute is also known as variable, field, characteristic, or feature.
- A measurement scale is a system (often a convention) of assigning a numerical or symbolic value to an attribute of an object.

Attributes

Objects

ID	Check	Status	Income	Risk
1	+	single	125 000	No
2	-	married	100 000	No
3	-	single	70 000	No
4	+	married	120 000	No
5	-	divorced	95 000	Yes
6	-	married	60 000	No
7	+	divorced	220 000	No
8	-	single	85 000	Yes
9	-	married	75 000	No
10	-	single	90 000	Yes

Objects and Attributes:

- An object $o \in O$ is described by a set of attributes.
  An object is also known as record, point, case, sample, entity, or instance.
- An attribute A is a property of an object.
  An attribute is also known as variable, field, characteristic, or feature.
- A measurement scale is a system (often a convention) of assigning a numerical or symbolic value to an attribute of an object.

Attributes

Objects

ID	Check	Status	Income	Risk
1	+	single	125 000	No
2	-	married	100 000	No
3	-	single	70 000	No
4	+	married	120 000	No
5	-	divorced	95 000	Yes
6	-	married	60 000	No
7	+	divorced	220 000	No
8	-	single	85 000	Yes
9	-	married	75 000	No
10	-	single	90 000	Yes

Objects and Attributes:

- An object $o \in O$ is described by a set of attributes.
  An object is also known as record, point, case, sample, entity, or instance.
- An attribute A is a property of an object.
  An attribute is also known as variable, field, characteristic, or feature.
- A measurement scale is a system (often a convention) of assigning a numerical or symbolic value to an attribute of an object.

Attributes

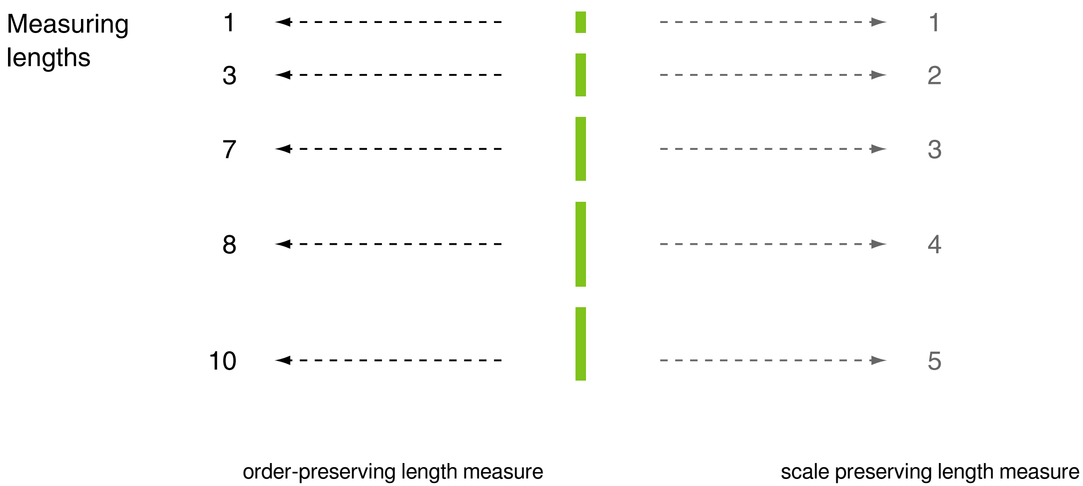
Objects

ID	Check	Status	Income	Risk
1	+	single	125 000	No
2	-	married	100 000	No
3	-	single	70 000	No
4	+	married	120 000	No
5	-	divorced	95 000	Yes
6	-	married	60 000	No
7	+	divorced	220 000	No
8	-	single	85 000	Yes
9	-	married	75 000	No
10	-	single	90 000	Yes

Objects and Attributes:

- Attribute values may vary from one object to another or one time to another.
- The 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.

The way an attribute is measured may not preserve all of the attribute's properties:



Objects and Attributes:

Types of Attributes

Type		Comparison	Statistics	Examples
categorical (qualitative)	nominal	values are names, only information to distinguish objects $=$ $\neq$	mode, entropy, contingency, correlation, χ^2 test	zip codes, employee IDs, eye color, gender: {male, female}

Objects and Attributes:

Types of Attributes

Type		Comparison	Statistics	Examples
categorical (qualitative)	nominal	values are names, only information to distinguish objects $= \quad \neq$	mode, entropy, contingency, correlation, χ^2 test	zip codes, employee IDs, eye color, gender: {male, female}
	ordinal	enough information to order objects $< \quad > \quad \leq \quad \geq$	median, percentiles, rank correlation	hardness of minerals, grades, street numbers, quality: {good, better, best}

Objects and Attributes:

Types of Attributes

Type		Comparison	Statistics	Examples
categorical (qualitative)	nominal	values are names, only information to distinguish objects $=$ $\neq$	mode, entropy, contingency, correlation, χ^2 test	zip codes, employee IDs, eye color, gender: {male, female}
	ordinal	enough information to order objects $<$ $>$ $\leq$ $\geq$	median, percentiles, rank correlation	hardness of minerals, grades, street numbers, quality: {good, better, best}
numerical (quantitative)	interval	differences are meaningful, a unit of measurement exists $+$ $-$	mean, standard deviation, Pearson's correlation, t-test, F-test	calendar dates, temperature in Celsius, temperature in Fahrenheit

Objects and Attributes:

Types of Attributes

Type		Comparison	Statistics	Examples
categorical (qualitative)	nominal	values are names, only information to distinguish objects $= \quad \neq$	mode, entropy, contingency, correlation, χ^2 test	zip codes, employee IDs, eye color, gender: {male, female}
	ordinal	enough information to order objects $< \quad > \quad \leq \quad \geq$	median, percentiles, rank correlation	hardness of minerals, grades, street numbers, quality: {good, better, best}
numerical (quantitative)	interval	differences are meaningful, a unit of measurement exists $+ \quad -$	mean, standard deviation, Pearson's correlation, t-test, F-test	calendar dates, temperature in Celsius, temperature in Fahrenheit
	ratio	differences and ratios are meaningful; Zero $\approx$ None $* \quad /$	percent variation, geometric mean, harmonic mean	temperature in Kelvin, monetary quantities, counts, age, length, electrical current

Objects and Attributes:
Types of Attributes

Type		Permissible transformation	Comment
categorical (qualitative)	nominal	any one-to-one mapping, permutation of values	A reassignment of employee ID numbers will not make any difference.

Objects and Attributes:

Types of Attributes

Type		Permissible transformation	Comment
categorical (qualitative)	nominal	any one-to-one mapping, permutation of values	A reassignment of employee ID numbers will not make any difference.
	ordinal	any order-preserving change of values: $x \to f(x)$, where f is monotonic	An attribute encompassing the notion of "{good, better, best}" can be represented equally well by the values $\{1, 2, 3\}$.

Objects and Attributes:

Types of Attributes

Type		Permissible transformation	Comment
categorical (qualitative)	nominal	any one-to-one mapping, permutation of values	A reassignment of employee ID numbers will not make any difference.
	ordinal	any order-preserving change of values: $x \rightarrow f(x)$, where f is monotonic	An attribute encompassing the notion of "{good, better, best}" can be represented equally well by the values $\{1, 2, 3\}$.
numerical (quantitative)	interval	$x \rightarrow a \cdot x + b$, where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).

Objects and Attributes:

Types of Attributes

Type		Permissible transformation	Comment
categorical (qualitative)	nominal	any one-to-one mapping, permutation of values	A reassignment of employee ID numbers will not make any difference.
	ordinal	any order-preserving change of values: $x \to f(x)$, where f is monotonic	An attribute encompassing the notion of "{good, better, best}" can be represented equally well by the values $\{1, 2, 3\}$.
numerical (quantitative)	interval	$x \to a \cdot x + b$, where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
	ratio	$x \to a \cdot x$, where a is a constant	Length can be measured in meters or feet.

Remarks:

- Identifying, considering, and measuring an attribute A of an object O is the heart of model formation and always goes along with a sort of abstraction. Formally, this abstraction is operationalized by a model formation function $\alpha : O \rightarrow X$. (⇒ Chapter ML Introduction)

- The terms "attribute" and "feature" can be used synonymously. However, a slight distinction is the following: attributes are often associated with objects, O, while features usually designate the dimensions of the feature space, X.

- The type of an attribute is also referred to as the type of a *measurement scale* or *level of measurement*.

- We call a transformation of an attribute *permissible* if its meaning is unchanged after the transformation.

- Distinguish between *discrete* attributes and *continuous* attributes. The former can only take a finite or countably infinite set of values, the latter can be measured in infinitely small units. Be careful when deriving from this distinction an attribute's type.

- We will encode attributes of interval type or ratio type by real numbers. Note that attributes of nominal type and ordinal type can also be encoded by real numbers.

- Particular learning methods require particular attribute types.

2. Types of Data Sets

Types of Data Sets:

Types of Data Sets

Data sets may not be a homogeneous collection of objects but come along with differently intricate characteristics:

1. Inhomogeneity of attributes:

2. Inhomogeneity of objects:

3. Curse of dimensionality:

4. Resolution:

Types of Data Sets:

Types of Data Sets

Data sets may not be a homogeneous collection of objects but come along with differently intricate characteristics:

1. Inhomogeneity of attributes:
   Consider the combination of different attribute types within a single object.

2. Inhomogeneity of objects:
   Consider the combination of different objects in a single data set.

3. Curse of dimensionality:
   Attribute number and object density stand in exponential relation.

4. Resolution:
   The number of objects or attributes may be given at different resolutions.

Types of Data Sets:

Types of Data Sets: Record Data

Collection of records, each of which consists of a fixed set of attributes:

ID	Check	Status	Income	Risk
1	+	single	125 000	No
2	-	married	100 000	No
3	-	single	70 000	No
4	+	married	120 000	No
5	-	divorced	95 000	Yes
6	-	married	60 000	No
7	+	divorced	220 000	No
8	-	single	85 000	Yes
9	-	married	75 000	No
10	-	single	90 000	Yes

- If all elements in a data set have the same fixed set of numeric attributes, they can be thought of as points in a multi-dimensional space.
- Such data can be represented by a matrix, where each row stores an object and each column stores an attribute.

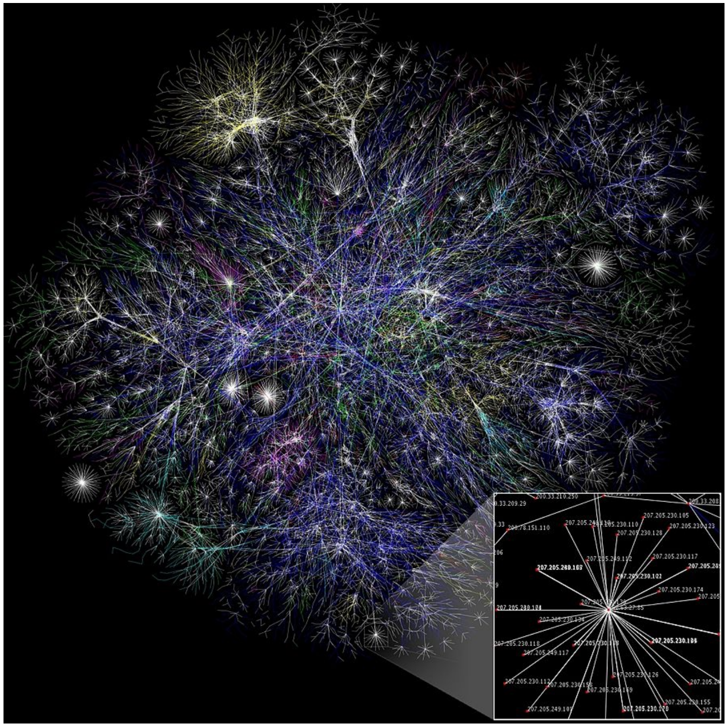
  Example: term-document matrices in information retrieval.

Types of Data Sets:

Types of Data Sets: Graph Data

Graph Data: objects contain special nominal attribute representing links

Example: Internet network visualised as graph:



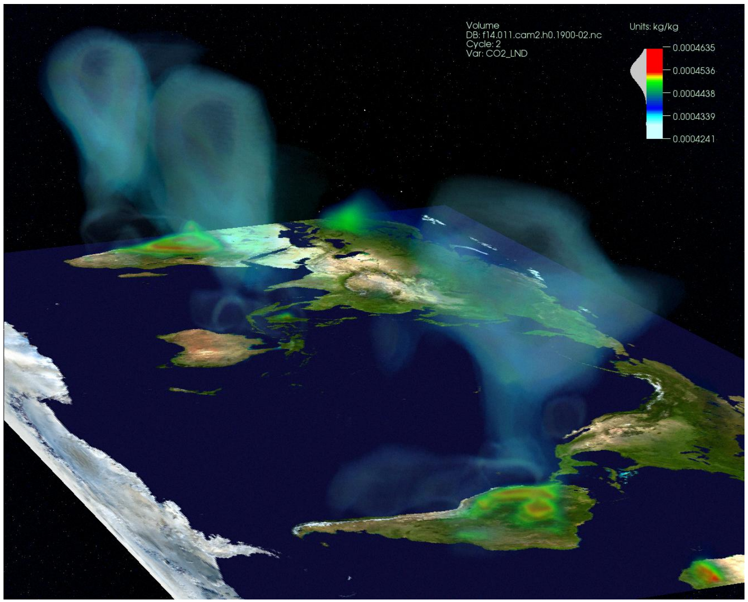
Types of Data Sets:

Types of Data Sets: Fields / Grids

Fields / Grids: Data arranged in spatial (e.g. space) relationships to each others. Objects have attributes representing a position / spatial information.

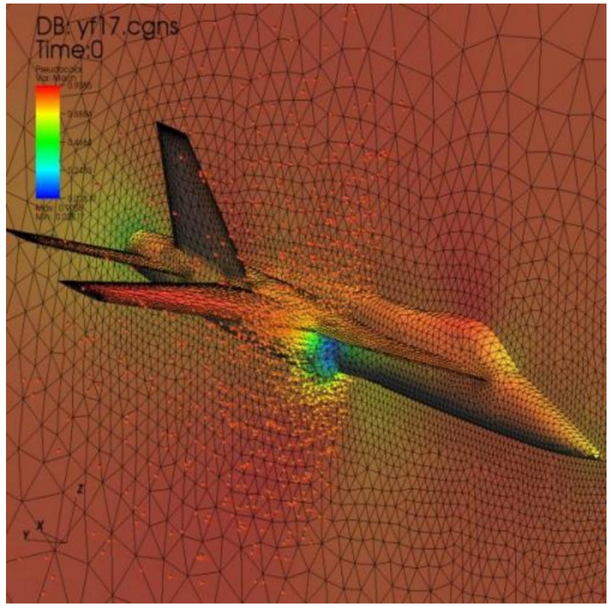
Example:

Climate Visualisation



Source Wikipedia

Material properties of a plane



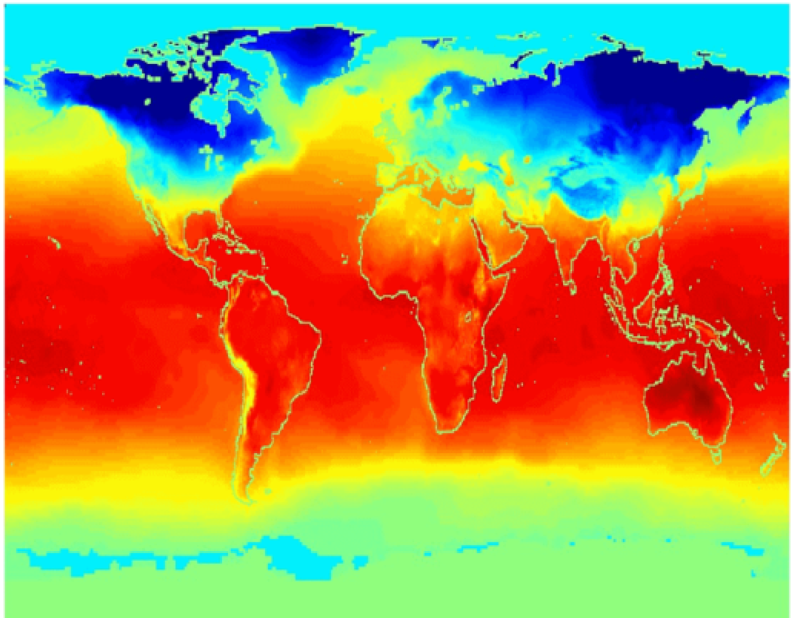
Source Wikipedia

Types of Data Sets:

Types of Data Sets: Spatio-temporal data

Spatio-temporal data sets: Attributes representing spatial information (e.g. longitude/latitude) and temporal information (e.g. time)

Example: Average monthly temperature of land and ocean (= spatio-temporal data):



Types of Data Sets:

Data Quality

When repeating measurements of a quantity, measurement errors and data collection errors may occur during the measurement process. Questions:

1. What kinds of data quality problems exist?
2. How to detect data quality problems?
3. How to address data quality problems?

Types of Data Sets:
Data Quality

When repeating measurements of a quantity, measurement errors and data collection errors may occur during the measurement process. Questions:

1. What kinds of data quality problems exist?
2. How to detect data quality problems?
3. How to address data quality problems?

Definition 1 (Precision, Bias, Accuracy)

Given a set of repeated measurements of the same quantity. Then, the closeness of the measurements to one another is called *precision*, the mean deviation from the true value is called *bias*, and the (overall) closeness to the true value is called *accuracy*.

Types of Data Sets:

Data Quality

When repeating measurements of a quantity, measurement errors and data collection errors may occur during the measurement process. Questions:

1. What kinds of data quality problems exist?
2. How to detect data quality problems?
3. How to address data quality problems?

Definition 2 (Precision, Bias, Accuracy)

Given a set of repeated measurements of the same quantity. Then, the closeness of the measurements to one another is called *precision*, the mean deviation from the true value is called *bias*, and the (overall) closeness to the true value is called *accuracy*.

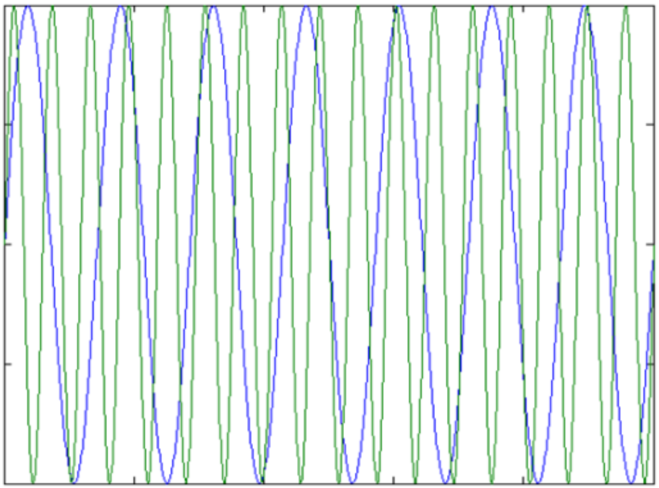
Examples for data quality problems:

- noise, artifacts, outliers
- missing values, duplicate data

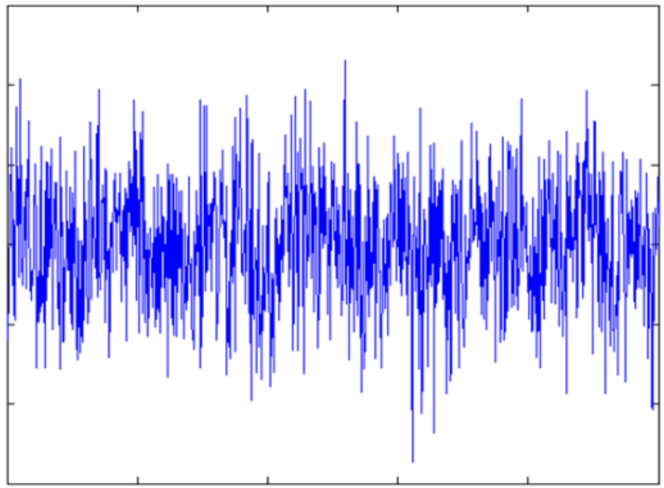
Types of Data Sets:

Data Quality: Noise

Noise refers to random modifications of attributes that often have a spatial or temporal characteristics:



sine waves



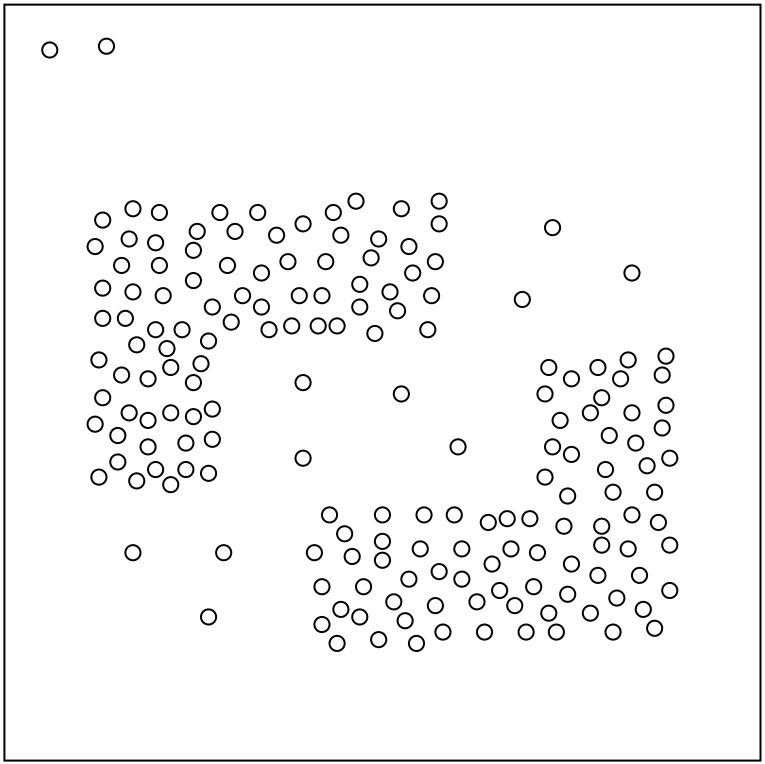
sine waves with noise

Artifacts refer to more deterministic distortions of a measurement process.

Types of Data Sets:

Data Quality: Outliers

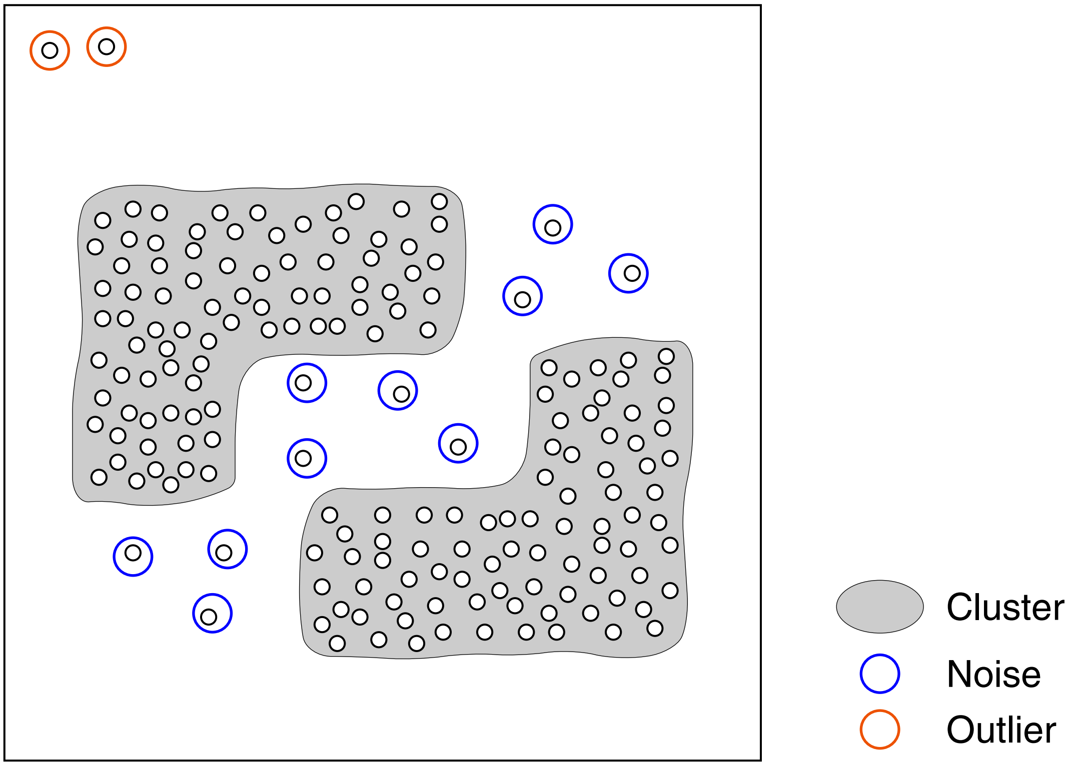
Outliers are members in the data set with characteristics that are considerably different (rare) than most of the other elements:



Types of Data Sets:

Data Quality: Outliers

Outliers are members in the data set with characteristics that are considerably different (rare) than most of the other elements:



Types of Data Sets:
Data Quality: Missing Values

Main reasons for missing values:

1. Information is not collected.
   Example: people decline to give their age or weight.

2. Attributes may not be applicable to all elements in O.
   Example: annual income is not applicable to children.

Strategies for handling missing values:

- eliminate members of the data
- estimate missing values (expected value, mode)
- ignore the missing value during analysis

3. Preprocessing

Preprocessing:

Preprocessing adapts objects and/or features to prepare the data for subsequent processing. The following adaptations can be distinguished on a conceptual level:

- aggregation of objects in O
- sampling of object set O
- sampling of feature space X
- selection of attributes (features) [attributes versus features]
- transformation of attributes (features)
- discretization and binarization of attributes (features)
- dimensionality reduction of feature space X

Preprocessing:
Aggregation

Example: Customer purchase information over different department stores.

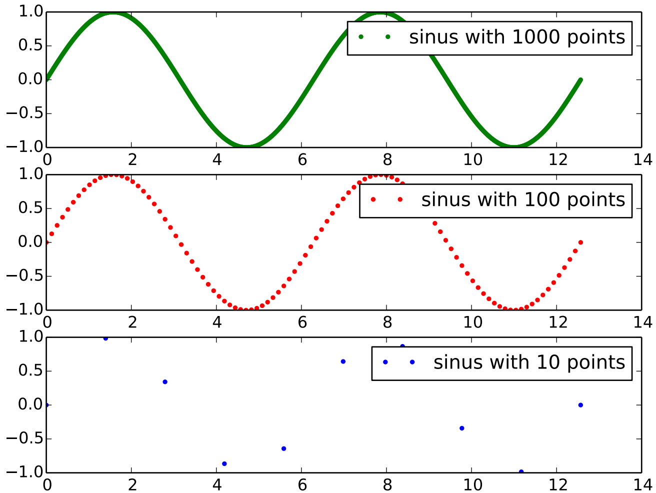
Transaction ID	Item	Store Location	Date	Price
⋮	⋮	⋮	⋮	⋮
101123	Watch	Berlin	13/10/21	EUR 100
101123	Shoes	Rosenheim	13/11/21	EUR 20
101123	Book	Passau	13/11/21	EUR 10
⋮	⋮	⋮	⋮	⋮

- Possible Aggregations on Store Location, Date, Price Range etc. by using some aggregation function like sum or average
- Reduction in processing resources
- Reduces variation which might lead to more stable estimates
- Loss in details
- Typical operation for Data Warehouse systems

Preprocessing:

Sampling of object set O

Example: Sampling of data in a sine wave



- Different kinds of sampling (random sampling with/without replacement, stratified sampling, progressive sampling etc.)
- Sampling means loss of information
- Sampling a proper amount of object retains their overall structure
- Useful for very large data sets

Preprocessing:
Dimensionality Reduction

Data sets may contain a large number of attributes (features) as for example words as features for documents

Problems of high dimensional data sets:

- Many irrelevant or noisy or correlated features
- Curse of Dimensionality
- Higher demand on computing resources
- Low understandability

Preprocessing:

Dimensionality Reduction

Data sets may contain a large number of attributes (features) as for example words as features for documents

Problems of high dimensional data sets:

- Many irrelevant or noisy or correlated features
- Curse of Dimensionality
- Higher demand on computing resources
- Low understandability

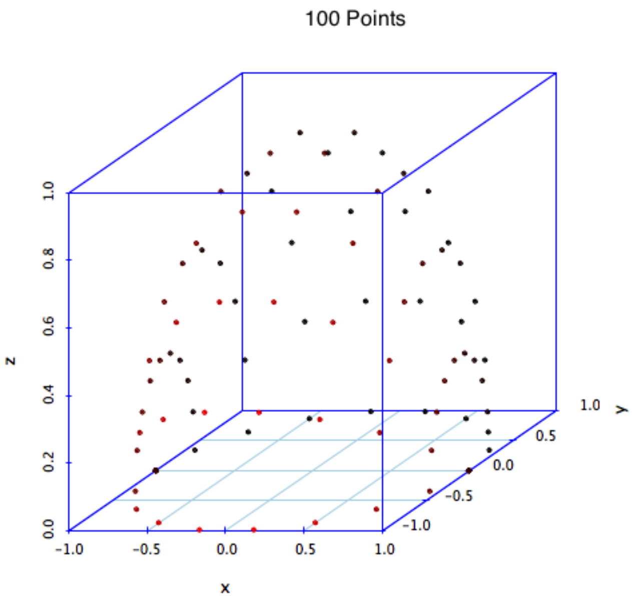
Techniques for reducing the dimensionality

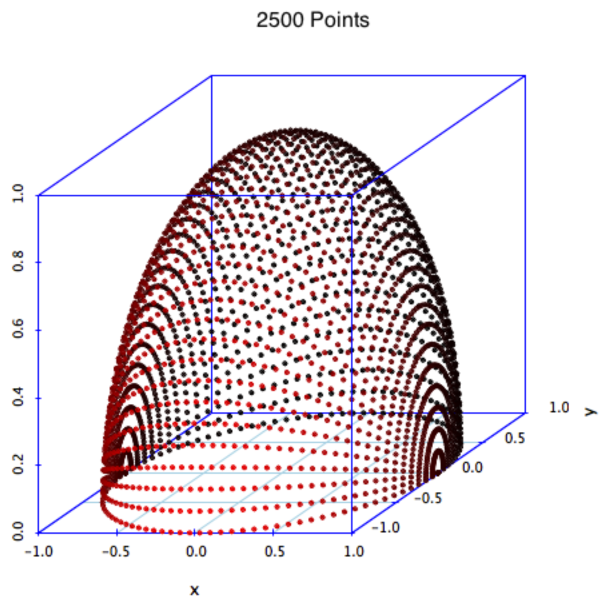
- Dimensionality Reduction Techniques: Create new attributes through a transformation: $X_{low} = T(X)$
- Feature (Subset) Selection: Select a subset of attributes (features)
- Feature Transformation/Creation Domain specific creation of attributes (e.g. image processing)

Preprocessing:
Dimensionality Reduction

Curse of Dimensionality

- As dimensionality increases, data is becoming increasingly sparse.
- More data points are needed to learn from data (i.e. separating relevant from irrelevant patterns)
- High dimensionality is problematic for a large number of learning algorithms





Preprocessing:
Dimensionality Reduction

Curse of Dimensionality: More formally, consider n objects with d binary attributes. We aim to label the objects as positive or negative. How many possible labelings exists?

Preprocessing:
Dimensionality Reduction

Curse of Dimensionality: More formally, consider n objects with d binary attributes. We aim to label the objects as positive or negative. How many possible labelings exists?

Every object can be either positive or negative. Hence we have 2^n different labelings.

But how many objects do we have?

Preprocessing:
Dimensionality Reduction

Curse of Dimensionality: More formally, consider n objects with d binary attributes. We aim to label the objects as positive or negative. How many possible labelings exists?

Every object can be either positive or negative. Hence we have 2^n different labelings.

But how many objects do we have?

We have $n = 2^d$ different objects.

So in total we have 2^{2^d} different labelings. The number of possible labelings grows exponentially with the number of attributes!

Preprocessing:
Dimensionality Reduction

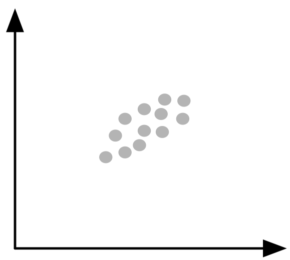
Dimensionality Reduction Methods

- Linear Methods (Projections known from Linear Algebra, e.g. Principal Component Analysis, Singular Value Decomposition, Random Projections)
- Nonlinear Methods (e.g., Locally-linear embedding, Multidimensional Scaling, Self-Organizing Maps)

Dimensionality Reduction Methods

- Linear Methods (Projections known from Linear Algebra, e.g. Principal Component Analysis, Singular Value Decomposition, Random Projections)
- Nonlinear Methods (e.g., Locally-linear embedding, Multidimensional Scaling, Self-Organizing Maps)

Example PCA - Principal Component Analysis: Select direction with the largest variance.



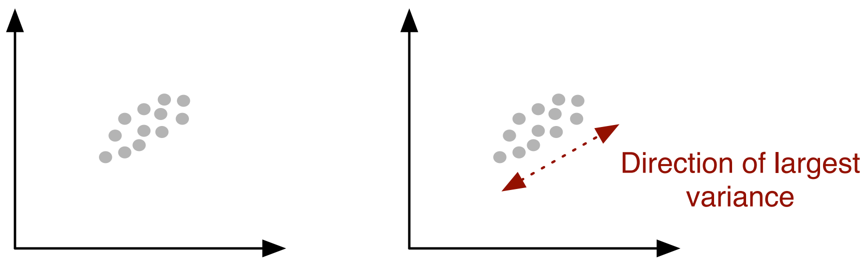
Preprocessing:

Dimensionality Reduction

Dimensionality Reduction Methods

- Linear Methods (Projections known from Linear Algebra, e.g. Principal Component Analysis, Singular Value Decomposition, Random Projections)
- Nonlinear Methods (e.g., Locally-linear embedding, Multidimensional Scaling, Self-Organizing Maps)

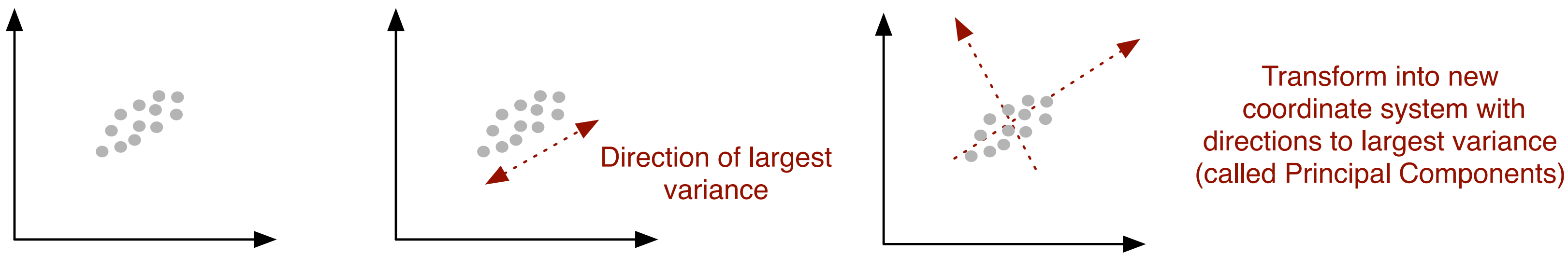
Example PCA - Principal Component Analysis: Select direction with the largest variance.



Dimensionality Reduction Methods

- Linear Methods (Projections known from Linear Algebra, e.g. Principal Component Analysis, Singular Value Decomposition, Random Projections)
- Nonlinear Methods (e.g., Locally-linear embedding, Multidimensional Scaling, Self-Organizing Maps)

Example PCA - Principal Component Analysis: Select direction with the largest variance.



Preprocessing:

Feature Subset Selection

Use only a subset of the features to overcome

- **Redundant features:** Duplicate or correlated features (e.g. birthdate and age)
- **Irrelevant features:** Features that do not contain information (e.g. student ID for predicting average grade)

Approach:

- **domain knowledge:** use of specific knowledge on the task/domain
- **automatic approach:** try different combinations of features

Preprocessing:

Feature Subset Selection

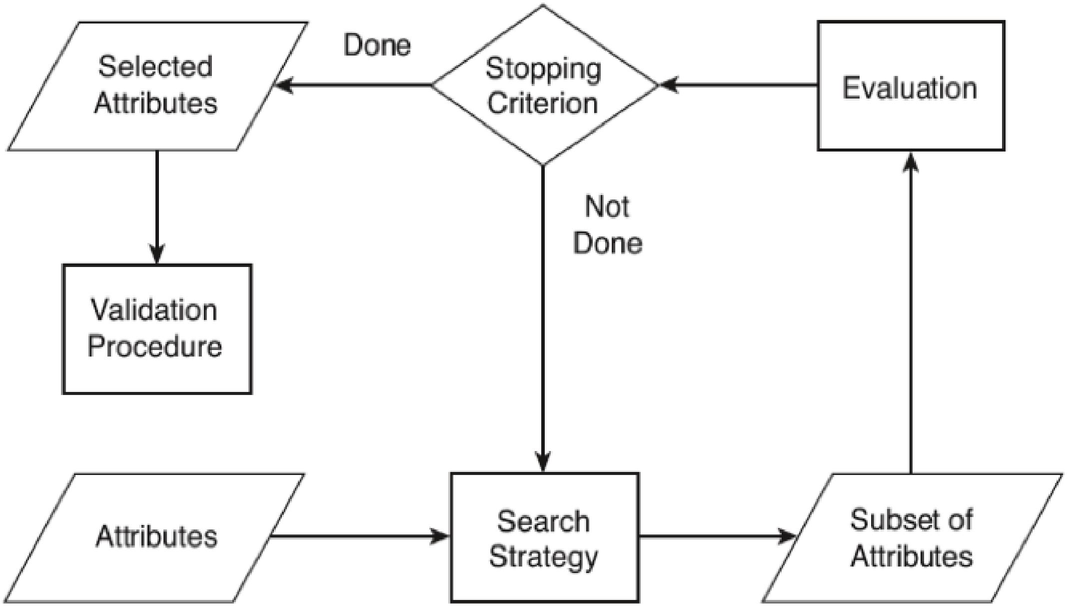
Given n features yields 2^n possible features subsets

Three standard approaches for testing

- **Embedded approach:** Feature selection is part of the learning algorithm (e.g. Decision Trees)
- **Filter approaches:** Features are selected before the learning algorithm is run using some statistical measure (e.g. correlation with dependent variable). Mostly assesses features individually (no subsets).
- **Wrapper approaches:** Consider feature selection as an outer loop around the learning algorithm. Choose a set of features, run the learning algorithm and record its performance.

Preprocessing:
Feature Subset Selection

Flow-chart for filter and wrapper approaches ([Tan et. al. 2005]):



- Search over all possible feature subsets
- Tradeoff between computational complexity and optimal subsets
- Search strategy and evaluation procedure most critical parts

Preprocessing:

Feature Creation

Creation of new attributes from the original ones.

Examples:

- Creating color histograms for face detection from pixel attributes
- Transforming an audio signal to its frequency spectrum (e.g. Fourier Transformation)
- Calculating the density of an object out of the mass and volume attributes

Feature creation is a highly domain specific process and reflects the properties of the model formulation function α

Preprocessing:

Discretization and Binarization

- **Discretization:** Transform numeric values of an attribute into categorical values
- **Binarization:** Transform numeric and categorical values of an attribute into binary values (i.e. nominal one)

Note that we do not change the attributes, but their values

Preprocessing:
Discretization and Binarization

- **Discretization:** Transform numeric values of an attribute into categorical values
  - Unsupervised Discretization
    Divide the continuous attribute based on statistical properties (e.g. histogram, cluster analysis)
  - Supervised Discretization
    Additional information like class labels $c(x)$ are used to find suitable intervalls (e.g. create bins that maximize class purity).
- **Binarization:** Transform numeric and categorical values of an attribute into binary values (i.e. nominal one)
  - Symmetric binary features
    Assign integer values to categorical values and convert the into a binary representation
  - Asymmetric binary features
    Introduce one feature per categorical value

Preprocessing:
Value Transformation

Value Transformation adapts the value range of an attribute for every object.

Two transformation could be distinguished:

- **Simple functions** like $\sqrt{x}$
- **Normalization or standardization**
  - Addresses the problem of having different scales/units between attributes
  - *Gaussian Normlization (z-score)*:

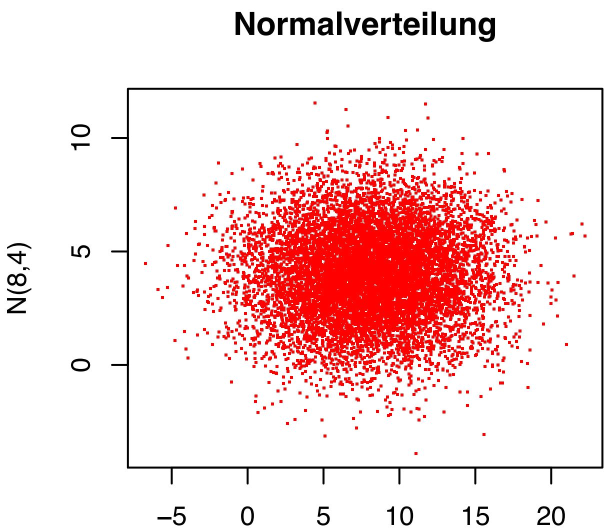
$$x_i' = \frac{x_i - \overline{x}}{s}$$

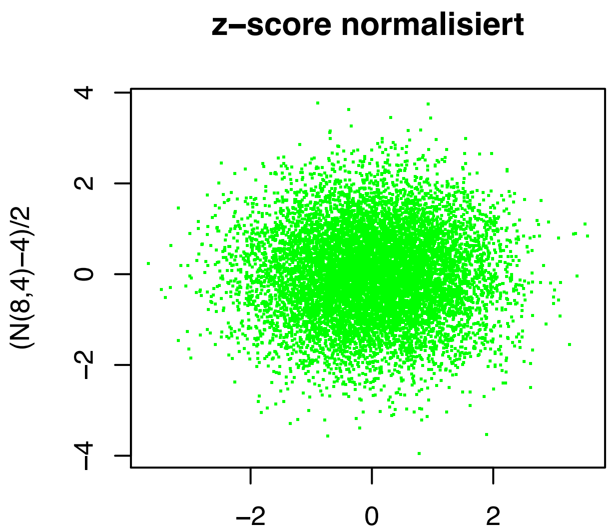
    with $\overline{x}$ being the mean of some attribute and s_k being the standard deviation of the attribute
  - *Min-Max Normalization*:

$$x_i' = \frac{x_i - min(x_i)}{max(x_i) - min(x_i)}$$

Preprocessing:
Value Transformation

Example Gaussian Normalization:





Preprocessing:
Value Transformation

Example Min-Max Normalization and its sensitivity to outliers: