

# DATA MINING 1

## Data Preparation

---

Dino Pedreschi, Riccardo Guidotti

*Revisited slides from Lecture Notes for Chapter 2 “Introduction to Data Mining”, 2nd Edition by Tan, Steinbach, Karpatne, Kumar*



# Data Understanding vs Data Preparation

---

**Data understanding** provides general information about data

- the existence of **missing values**
- the existence of **outliers**
- the character of attributes
- **dependencies** between attributes.

**Data preparation** uses this information to

- select attributes
- reduce the data dimension
- select records
- treat missing values
- treat outliers
- integrate, unify and transform data
- improve data quality

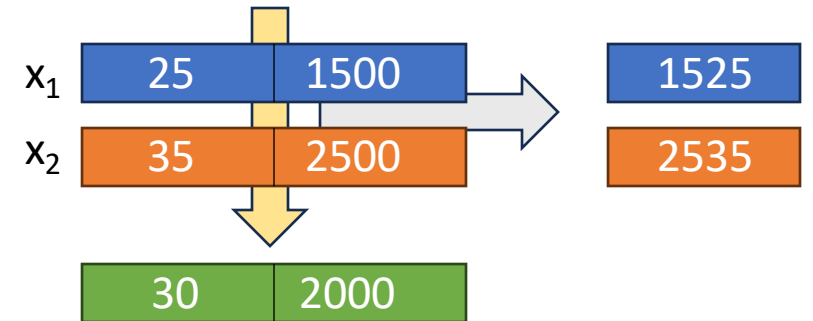
# Data Preparation

---

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

# Aggregation

- Combining two or more attributes (or points) into a single attribute (or point)
- Purpose
  - **Data reduction**
    - Reduce the number of attributes or objects
  - **Change of scale**
    - Cities aggregated into regions, states, countries, etc.
    - Days aggregated into weeks, months, or years
  - **More “stable” data**
    - Aggregated data tends to have less variability



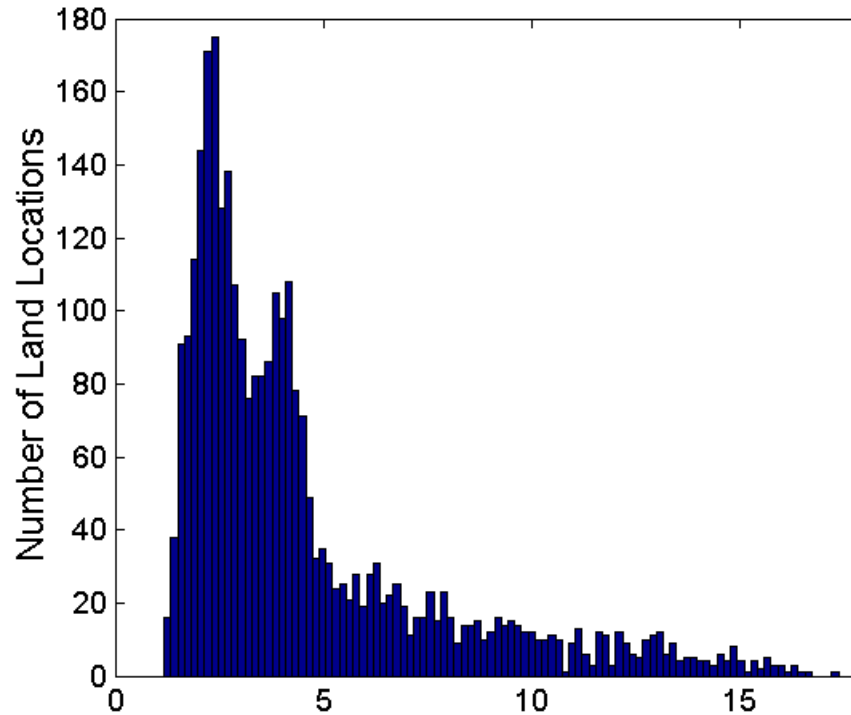
# Example: Precipitation in Australia

---

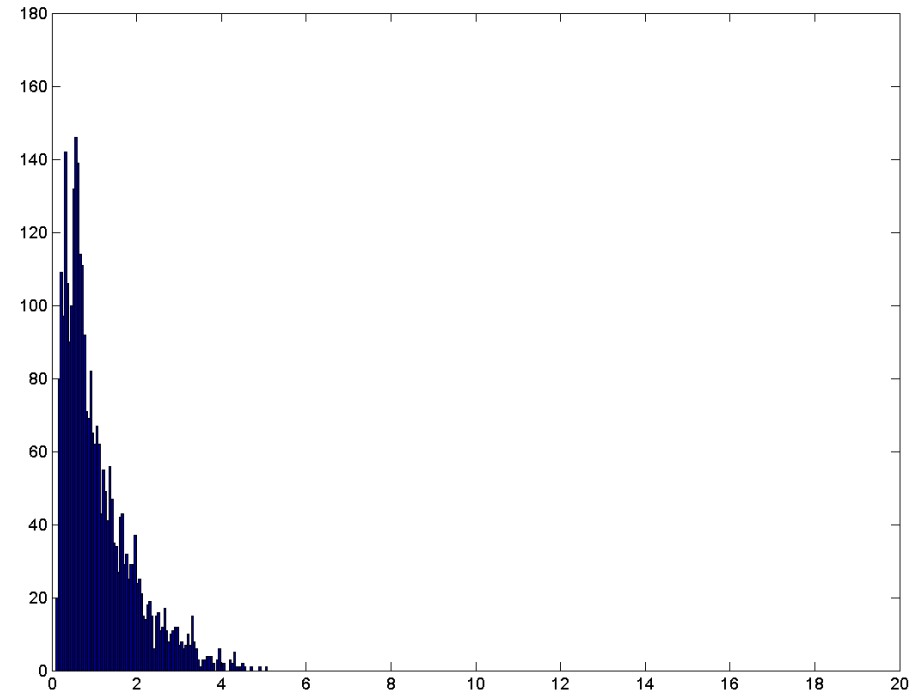
- This example is based on precipitation in Australia from 1982 to 1993.  
The next slide shows
  - A histogram for the standard deviation of average monthly precipitation for specific locations in Australia, and
  - A histogram for the standard deviation of the average yearly precipitation for the same locations.
- The average **yearly precipitation has less variability** than the **average monthly precipitation**.
- All precipitation measurements (and their standard deviations) are in centimeters.

# Example: Precipitation in Australia ...

## Variation of Precipitation in Australia



**Standard Deviation of Average Monthly Precipitation**



**Standard Deviation of Average Yearly Precipitation**

# Data Reduction

---

## Reducing the amount of data

- Reduce the number of **records** (rows)
  - Data Sampling
  - Clustering
- Reduce the number of **attributes** (columns)
  - Select a subset of attributes
  - Generate a new (a smaller) set of attributes

# Sampling

---

- Sampling is the main technique employed for data reduction.
  - It is often used for both the preliminary investigation of the data and the final data analysis.
- Sampling is typically used in data mining because **processing** the entire set of data of interest is too expensive or time consuming.



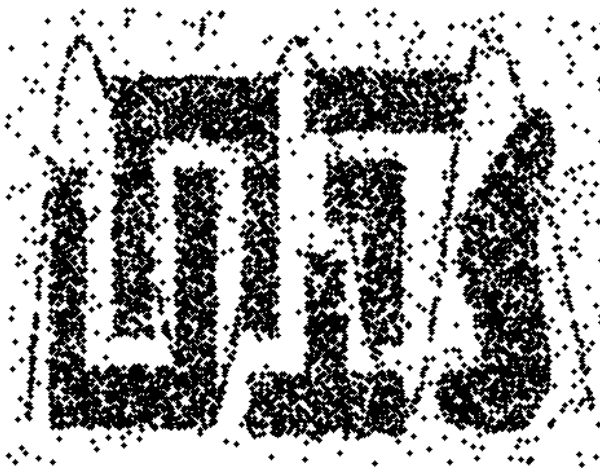
# Sampling ...

---

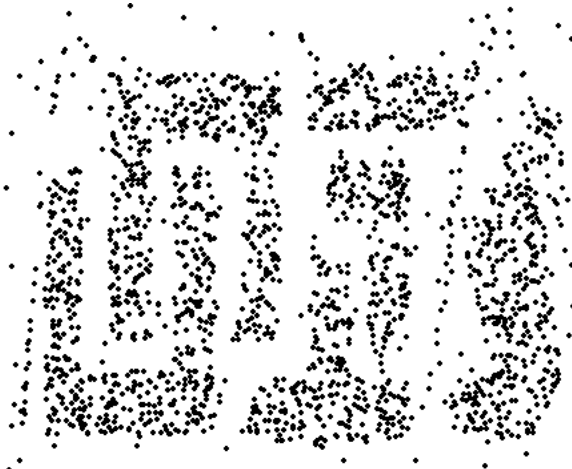
- The key principle for effective sampling is the following:
  - Using a sample will work almost as well as using the entire data set, **if the sample is representative**
  - A sample is **representative** if it has approximately the **same properties** (of interest) as the original set of data

# Sample Size

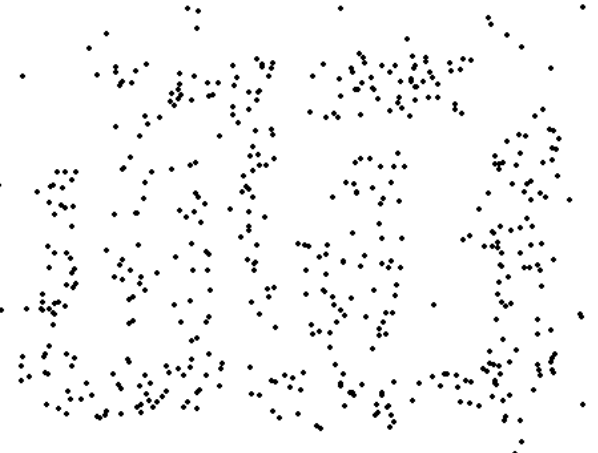
---



8000 points



2000 Points



500 Points

# 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

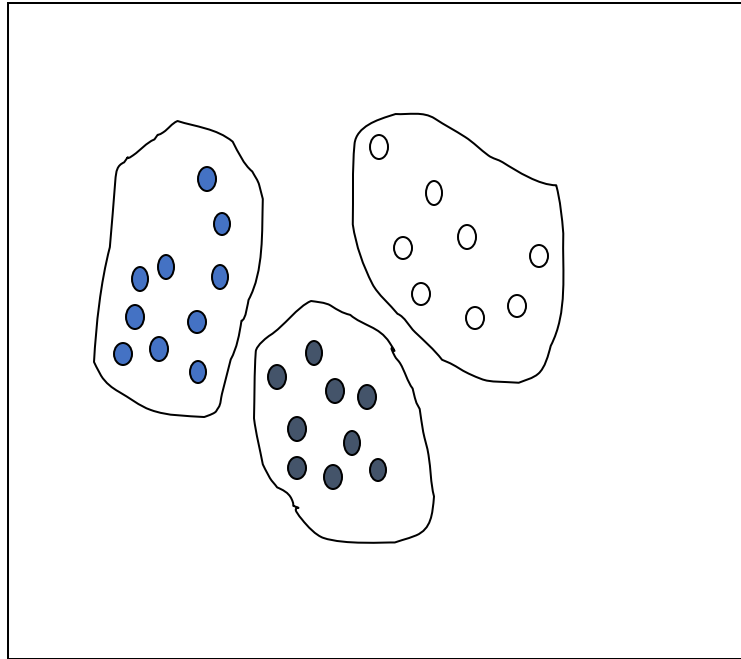
- Stratified sampling

- Split the data into several partitions; then draw random samples from each partition
- Approximation of the percentage of each class
- Suitable for distribution with peaks: each peak is a **layer**

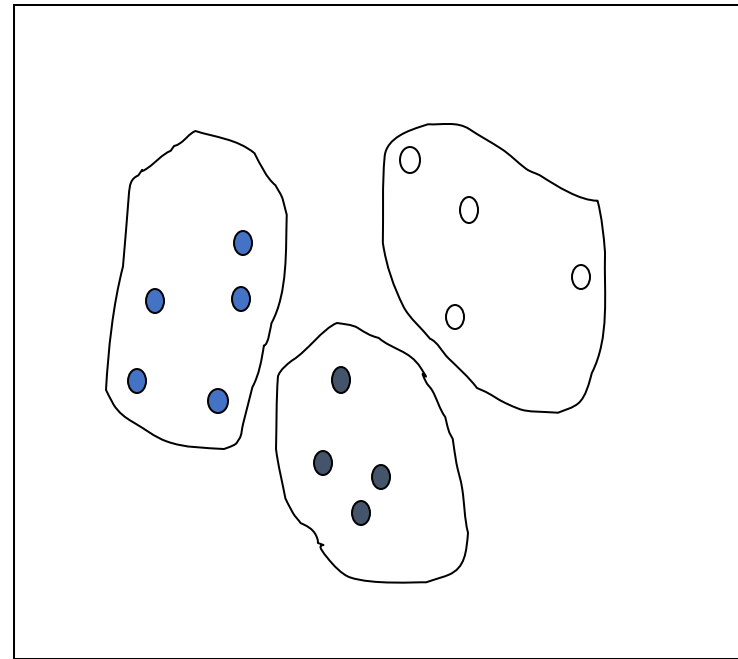
# Stratified Sampling

---

**Raw Data**



**Cluster/Stratified Sample**



# Reduction of Dimensionality

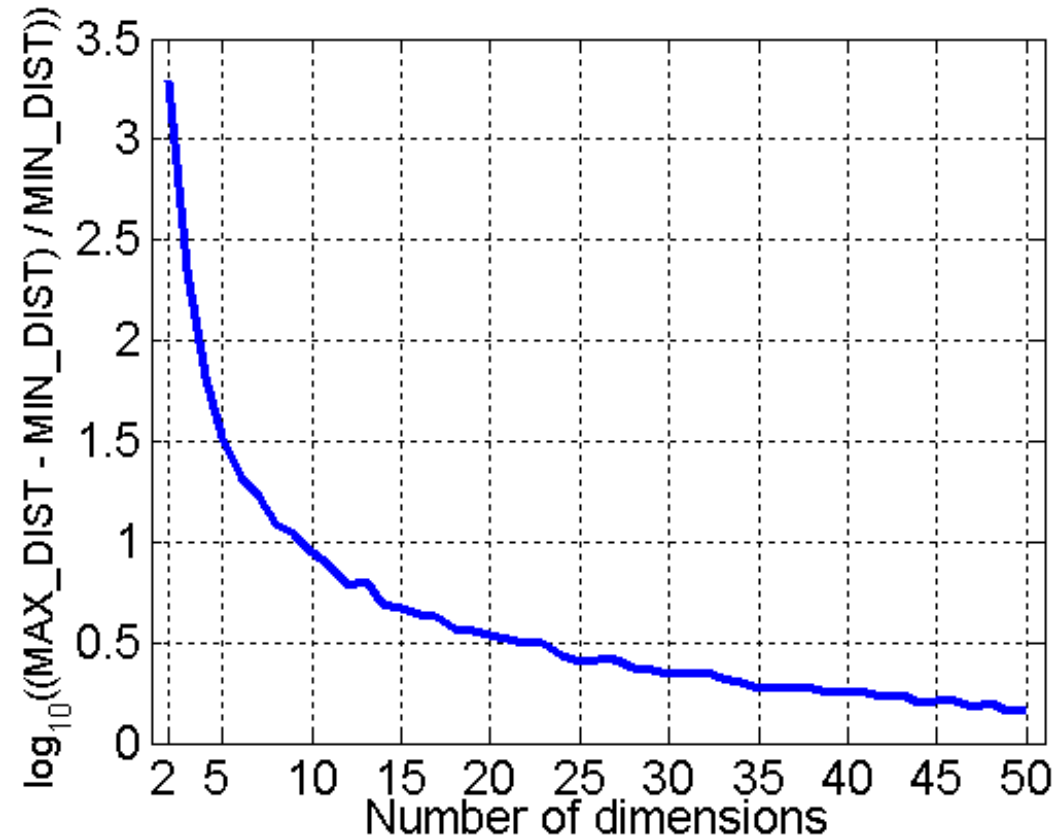
---

**Selection of a subset of attributes** that is as small as possible and sufficient for the data analysis.

- removing (more or less) **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
- removing **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

# Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies
- Definitions of density and distance between points, which are critical for clustering and outlier detection, become less meaningful



# Dimensionality Reduction

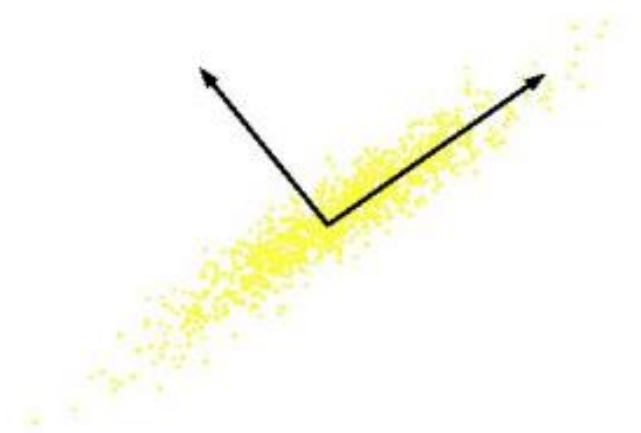
---

- 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
  - **Principal Components Analysis (PCA)**
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

# Principal Component Analysis

---

- The goal of PCA is to **find a new set of dimensions** (attributes or features) that better **captures the variability of the data**.
- The **first dimension** is chosen to capture as **much of the variability** as possible.
- The **second dimension** is orthogonal to the first and, subject to that constraint, captures as much of **the remaining variability** as possible, and so on.
- It is a **linear transformation** that chooses a new coordinate system for the data set





# Steps of the approach

---

- **Step 1:** Standardize the dataset.
- **Step 2:** Calculate the covariance matrix for the features in the dataset.
- **Step 3:** Calculate the eigenvalues and eigenvectors for the covariance matrix.
- **Step 4:** Sort eigenvalues and their corresponding eigenvectors and pick  $k$  eigenvalues and form a matrix of eigenvectors.
- **Step 5:** Transform the original matrix.

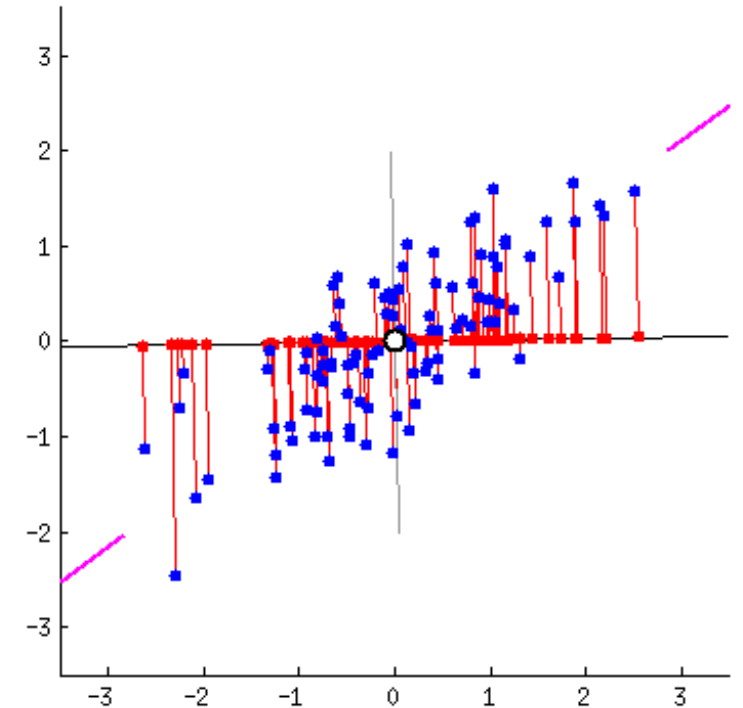
# Covariance key factor in PCA

---

- Variance and Covariance are a measure of the “**spread**” of a set of points around their **center of mass** (mean)
- **Variance** – measure of the deviation from the mean for points in one dimension e.g. heights
- **Covariance** as a measure of how much each of the dimensions vary from the mean with respect to each other.
- Covariance is measured between 2 dimensions to see if there is a relationship between the 2 dimensions
  - e.g. number of hours studied & marks obtained.
- The covariance between one dimension and itself is the variance

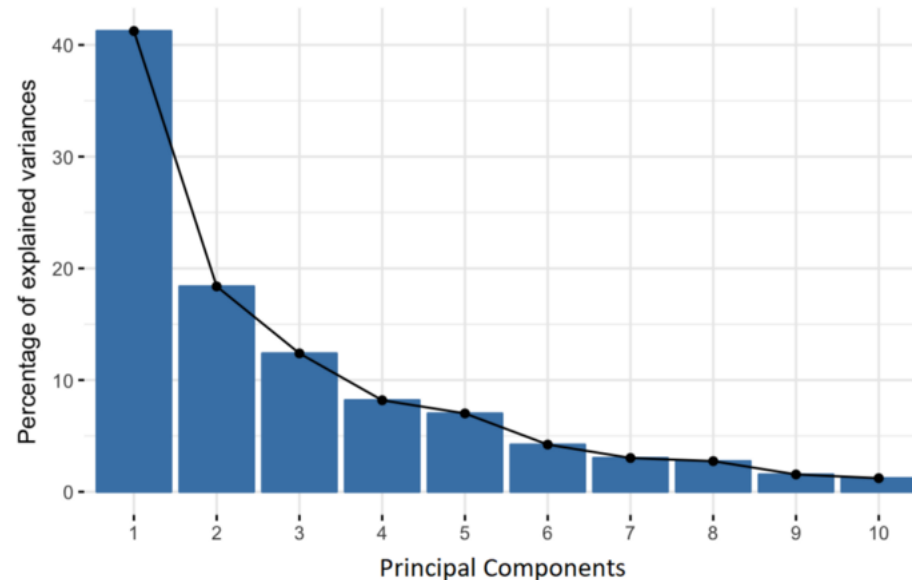
# How to construct PC?

- The first principal component accounts for the **largest possible variance** in the data set.
  - I want to fix the black line such that the spread on them of the red points, i.e., the original points projected on the black line, is maximised.
  - The second principal component is calculated in the same way, with the condition that **it is uncorrelated with the first principal component** and that it accounts for the **next highest variance**.
- 
- More details will be provided in DM2



# Identify the Principal Components

Given 10-dimensional data you get 10 principal components but only the first PCs capture most of the variability of the data



Discarding the components with low information and considering the remaining components as your new variables.

# Removing Irrelevant/Redundant Features

---

- For **removing irrelevant features**, a **performance measure** is needed that indicates how well a feature or subset of features performs w.r.t. the considered data analysis task
- For removing **redundant features**, either a **performance measure** for subsets of features or a **correlation measure** is needed.

# Reduction of Dimensionality

---

## Filter Methods

- Selection after analyzing the **significance** and **correlation** with other attributes
- Selection is independent of any data mining task
- The operation is a pre-processing

## Wrapper Methods

- Selecting the top-ranked features using as reference a DM task
- Incremental Selection of the “best” attributes
- “Best” = with respect to a specific measure of statistical significance (e.g.: information gain)

## Embedded Methods

- Selection as part of the data mining algorithm
- During the operation of the DM algorithm, the algorithm itself decides which attributes to use and which to ignore (e.g. Decision tree)

# Wrapper Feature Selection Techniques

---

- **Selecting the top-ranked features:** Choose the features with the best evaluation when single features are evaluated.
- **Selecting the top-ranked subset:** Choose the subset of features with the best performance. This requires exhaustive search and is impossible for larger numbers of features. (For 20 features there are already more than one million possible subsets.)
- **Forward selection:** Start with the empty set of features and add features one by one. In each step, add the feature that yields the best improvement of the performance.
- **Backward elimination:** Start with the full set of features and remove features one by one. In each step, remove the feature that yields to the least decrease in performance.

# 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 construction
    - Domain-dependent
    - Example: dividing mass by volume to get density
  - Feature Projection
    - Transforms the data from the high-dimensional space to a space of fewer dimensions



# Feature Creation: features needed for task

## Find the best workers in a company.

- Attributes :
  - the tasks, a worker has finished within each month,
  - the number of hours he has worked each month,
  - the number of hours that are normally needed to finish each task.
- These attributes *contain* information about the efficiency of the worker.
- But instead using these three “raw” attributes, it might be more useful to define a new attribute *efficiency*.
- $\text{efficiency} = \frac{\text{hours actually spent to finish the tasks}}{\text{hours normally needed to finish the tasks}}$

# Feature Creation: features needed for task

---

- **Task:** face recognition in images
- Images are only set of contiguous pixels
- They are not suitable for many types of classification algorithms
- Process to provide **higher level features**
  - presence or absence of certain types of areas that are highly correlated with the presence of human faces
  - a much broader set of classification techniques can be applied to this problem

# Feature Projection or Extraction

---

- It transforms the data in the high-dimensional space to a space of fewer dimensions.
- The data transformation may be linear, or nonlinear.
- Approaches:
  - Principal Component Analysis (PCA)
  - Singular Value Decomposition (SVD)
  - Non-negative matrix factorization (NMF)
  - Linear Discriminant Analysis (LDA)
  - Autoencoder

# Data Cleaning

---

- How to handle anomalous values
- How to handle outliers
- Data Transformations

# Anomalous Values

---

- **Missing values**
  - NULL, ?
- **Unknown Values**
  - Values without a real meaning
- **Not Valid Values**
  - Values not significant

# Manage Missing Values

---

1. Elimination of records
2. Substitution of values

**Note:** it can influence the original distribution of numerical values

- Use mean/median/mode
- Estimate missing values **using the probability distribution** of existing values
- Data Segmentation and using mean/mode/median of each **segment**
- Data Segmentation and using **the probability distribution within the segment**
- Build a model of **classification/regression** for computing missing values

# Discretization

---

- **Discretization** is the process of converting a continuous attribute into an ordinal attribute
  - A potentially infinite number of values are mapped into a small number of categories
  - Discretization is commonly used in classification
  - Many classification algorithms work best if both the independent and dependent variables have only a few values
- When you make a histogram for a continuous attribute you are discretizing your data!

# Discretization: Advantages

---

- Hard to understand the optimal discretization
  - We should need the real data distribution
- Original values can be **continuous** and **sparse**
- Discretized data can be **simple** to be interpreted
- Data distribution after discretization can have a **Normal shape**
- Discretized data can be too much **sparse yet**
  - Elimination of the attribute



# Unsupervised Discretization

---

- Characteristics:
  - No label for the instances
  - The number of classes is unknown
  
- Techniques of *binning*:
  - **Natural binning** → Intervals with the same width
  - **Equal Frequency binning** → Intervals with the same frequency
  - **Statistical binning** → Use statistical information (Mean, variance, Quartile)

# Discretization of Quantitative Attributes

---

**Solution:** each value is replaced by the interval to which it belongs.

**height:** 0-150cm, 151-170cm, 171-180cm, >180c

**weight:** 0-40kg, 41-60kg, 60-80kg, >80kg

**income:** 0-10ML, 11-20ML, 20-25ML, 25-30ML, >30ML

CID	height	weight	income
1	151-171	60-80	>30
2	171-180	60-80	20-25
3	171-180	60-80	25-30
4	151-170	60-80	25-30

**Problem:** the discretization may be useless (see **weight**).

# How many groups (or classes)?

---

- If too few
  - ⇒ Loss of information on the distribution
- If too many
  - ⇒ Dispersion of values and does not show the form of distribution
- The optimal number of classes is function of  $N$  elements (Sturges, 1929)

$$C = 1 + \frac{10}{3} \log_{10}(N)$$

- The optimal width of the classes depends on the variance and the number of data (Scott, 1979)

$$h = \frac{3,5 \cdot s}{\sqrt{N}}$$

# Supervised Discretization

---

- **Characteristics:**

- The discretization has a quantifiable goal
- The number of classes is known

- **Techniques:**

- discretization based on Entropy
- discretization based on percentiles

# Binarization

---

- Binarization maps a continuous or categorical attribute into one or more binary variables
- Typically used for association analysis
- Often convert a continuous attribute to a categorical attribute and then convert a categorical attribute to a set of binary attributes
  - Association analysis needs asymmetric binary attributes
  - Examples: eye color and height measured as {low, medium, high}

# Binarization

$n = \log_2(m)$  binary digits are required to represent  $m$  integers.

It can generate some correlations

Table 2.5. Conversion of a categorical attribute to three binary attributes.

Categorical Value	Integer Value	$x_1$	$x_2$	$x_3$
<i>awful</i>	0	0	0	0
<i>poor</i>	1	0	0	1
<i>OK</i>	2	0	1	0
<i>good</i>	3	0	1	1
<i>great</i>	4	1	0	0

- **One variable for each possible value**
- Only presence or absence
- Association Rules requirements

Table 2.6. Conversion of a categorical attribute to five asymmetric binary attributes.

Categorical Value	Integer Value	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
<i>awful</i>	0	1	0	0	0	0
<i>poor</i>	1	0	1	0	0	0
<i>OK</i>	2	0	0	1	0	0
<i>good</i>	3	0	0	0	1	0
<i>great</i>	4	0	0	0	0	1

# Data Transformation: Motivations

---

- Data with errors and incomplete
- Data not adequately distributed
  - Strong asymmetry in the data
  - Many peaks
- Data transformation can reduce these issues

# Attribute Transformation

---

- An **attribute transform** is 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:  $x^k$ ,  $\log(x)$ ,  $e^x$ ,  $|x|$
  - **Normalization**
    - Refers to various techniques to adjust to differences among attributes in terms of frequency of occurrence, mean, variance, range
    - Take out unwanted, common signal, e.g., seasonality
  - In statistics, **standardization** refers to subtracting off the means and dividing by the standard deviation



# Properties of Transformation

---

- Define a transformation  $T$  on the attribute  $X$ :

$$Y = T(X)$$

such that :

- $Y$  preserve the **relevant** information of  $X$
- $Y$  eliminates at least one of the problems of  $X$
- $Y$  is more **useful** of  $X$

# Transformation Goals

---

- **Main goals:**

- stabilize the variances
- normalize the distributions
- Make linear relationships among variables

- **Secondary goals:**

- simplify the elaboration of data containing features you do not like
- represent data in a scale considered more suitable

# Why linear correlation, normal distributions, etc?

---

- Many statistical methods require
  - linear correlations
  - normal distributions
  - the absence of outliers
- Many data mining algorithms have the ability to automatically treat **non-linearity** and **non-normality**
  - The algorithms work still better if such problems are treated

# Normalizations

---

- min-max normalization

$$v' = \frac{v - \mathit{min}_A}{\mathit{max}_A - \mathit{min}_A} (\mathit{new\_max}_A - \mathit{new\_min}_A) + \mathit{new\_min}_A$$

- z-score normalization

$$v' = \frac{v - \mathit{mean}_A}{\mathit{stand\_dev}_A}$$

# Transformation Functions

---

- Exponential transformation

$$T_p(x) = \begin{cases} ax^p + b & (p \neq 0) \\ c \log x + d & (p = 0) \end{cases}$$

- with  $a, b, c, d$  and  $p$  real values
  - Preserve the order
  - Preserve some basic statistics
  - They are continuous functions
  - They are derivable
  - They are specified by simple functions

# Better Interpretation

---

- Linear Transformation

$$1\text{€} = 1936.27 \text{ Lit.}$$

- $p=1, a=1936.27, b=0$

$$T_p(x) = \begin{cases} ax^p + b & (p \neq 0) \\ c \log x + d & (p = 0) \end{cases}$$

$$^{\circ}\text{C} = 5/9(^{\circ}\text{F} - 32)$$

- $p = 1, a = 5/9, b = -160/9$

# Stabilizing the Variance

---

- **Logarithmic Transformation**

$$T(x) = c \log x + d$$

- Applicable to positive values
- Makes homogenous the variance in log-normal distributions
  - E.g.: normalize seasonal peaks

# Logarithmic Transformation: Example

<i>Bar</i>	<i>Birra</i>	<i>Ricavo</i>
A	Bud	20
A	Becks	10000
C	Bud	300
D	Bud	400
D	Becks	5
E	Becks	120
E	Bud	120
F	Bud	11000
G	Bud	1300
H	Bud	3200
H	Becks	1000
I	Bud	135

2300 Mean  
2883,3333 Scarto medio assoluto  
3939,8598 Standard Deviation  
5 Min  
120 1° Quartile  
350 Median  
1775 2° Quartile  
11000 Max

**Data are sparse!!!**



# Logarithmic Transformation: Example

<b><i>Bar</i></b>	<b><i>Birra</i></b>	<b><i>Ricavo (log)</i></b>
A	Bud	1,301029996
A	Becks	4
C	Bud	2,477121255
D	Bud	2,602059991
D	Becks	0,698970004
E	Becks	2,079181246
E	Bud	2,079181246
F	Bud	4,041392685
G	Bud	3,113943352
H	Bud	3,505149978
H	Becks	3
I	Bud	2,130333768

Media	2,585697
Scarto medio assoluto	0,791394
Deviazione standard	1,016144
Min	0,69897
Primo Quartile	2,079181
Mediana	2,539591
Secondo Quartile	3,211745
Max	4,041393

# Stabilizing the Variance

---

$$T(x) = ax^p + b$$

- **Square-root Transformation**

- $p = 1/c$ ,  $c$  integer number

- To make homogenous the variance of particular distributions e.g., Poisson Distribution

- **Reciprocal Transformation**

- $p < 0$

- Suitable for analyzing time series, when the variance increases too much wrt the mean