University
Politehnica
of Bucharest

Faculty of
Automatic
Control and
Computers

Computer
Science and
Engineering
Department

Introduction to Data Mining

Ciprian-Octavian Truică
ciprian.truica@cs.pub.ro

Overview

- What is data mining?
- Data mining steps
- Data mining methods and sub-domains
- Data Preprocessing
- Summary

Overview

- What is data mining?
- Data mining steps
- Data mining methods and sub-domains
- Data Preprocessing
- Summary

Definition

- Data Mining is commonly defined as the process of discovering useful patterns or knowledge from data sources [1], e.g., databases, texts, images, the web, etc.
- The patterns must be valid, potentially useful and understandable.
- It is the analysis step in Knowledge Discovery in Databases (KDD) [2].

Definition [3]

- Data mining is the discovery of “models” from data:
  - Statistical modeling
  - Machine learning
  - Computational approaches to modeling
  - Summarization
  - Feature extraction

Definition (Wikipedia)

- Data mining is the computational process of discovering patterns in large data sets (Big Data) involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.
- The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.
- Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

Definition [4]

- A class of undirected queries, often against the most atomic data, that seek to find unexpected patterns in the data.
- The most valuable results from data mining are clustering, classifying, estimating, predicting, and finding things that occur together.
- There are many kinds of tools that play a role in data mining, including decision trees, neural networks, memory- and case-based reasoning tools, visualization tools, genetic algorithms, fuzzy logic, and classical statistics.
- Generally, data mining is a client of the data warehouse.

What is not data mining

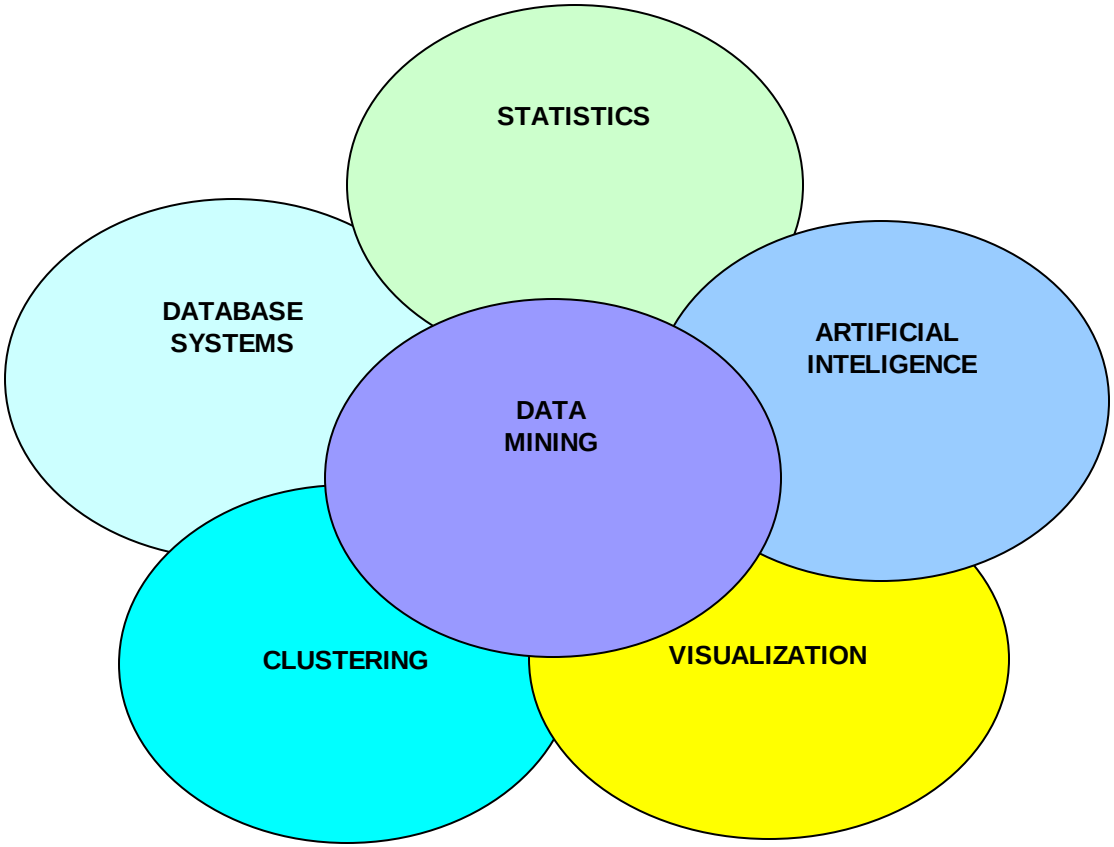
- Find a certain person in an employee database
- Compute the minimum, maximum, sum, count or average values based on table/tables columns
- Use a search engine to find your name occurrences on the web

Conclusions

- The data mining process converts data into valuable knowledge that can be used for decision support
- Data mining is a collection of data analysis methodologies, techniques and algorithms for discovering patterns and models
- Data mining is used on large data sets
- Data mining process is automated (no need for human intervention)

Conclusions

The most important communities involved in data mining



Overview

- What is data mining?
- Data mining steps
- Data mining methods and sub-domains
- Data Preprocessing
- Summary

Data mining steps

- Step1: Data collection
  - Data gathering from existing databases or (for Internet documents) from Web crawling.

Data mining steps

- Step 2. Data preprocessing :
  - Data cleaning: replace (or remove) missing values, smooth noisy data, remove or just identify outliers, remove inconsistencies.
  - Data integration: integration of data from multiple sources, with possible different data types and structures and also handling of duplicate or inconsistent data.
  - Data transformation: data normalization (or standardization), summarizations, generalization, new attributes construction, etc.

Data mining steps

- Step 2 (cont.). Data preprocessing :
  - Data reduction (called also feature extraction): not all the attributes are necessary for the particular Data Mining process we want to perform. Only relevant attributes are selected for further processing reducing the total size of the dataset (and the time needed for running the algorithm).

Data mining steps

- Step 2 (cont.). Data preprocessing :
  - Discretization: some algorithms work only on discrete data. For that reason the values for continuous attributes must be replaced with discrete ones from a limited set. One example is replacing age (number) with an attribute having only three values: Young, Middle-age and Old

Data mining steps

- Step 3. Pattern extraction and discovery:
  - This is the stage where the data mining algorithm is used to obtain the result.
  - Some authors consider that Data Mining is reduced only at this step, the whole process being called KDD.

Data mining steps

- Step 4. Visualization:
  - Because data mining extracts hidden properties/information from data it is necessary to visualize the results for a better understanding and evaluation.
  - Also needed for the input data.

Data mining steps

- Step 5. Evaluation of results:
  - Not everything that outputs from a data mining algorithm is a valuable fact or information.
  - Some of them are statistic truths and others are not interesting/useful for our activity.
  - Expert judgment is necessary in evaluating the results

Bonferroni principle

- A true information discovered by a 'data mining' process can be a statistical truth. Example (from [Ullman 03]):
  - In 1950's David Rhine, a parapsychologist, tested students in order to find if they have or not extra-sensorial perception (ESP).
  - He asked them to guess the color of 10 successive cards – red or black. The result was that 1/1000 of them guessed all 10 cards (he declared they have ESP).
  - Re-testing only these students he found that they have lost ESP after knowing they have this feature
  - David Rhine did not realize that the probability of guessing 10 successive cards is $1/1024 = 1/2^{10}$, because the probability for each of these 10 cards is ½ (red or black).

Bonferroni principle

- This kind of results may be included in the output of a data mining algorithm but must be recognized as a statistical truth and not a real data mining output.
- This fact is also the object of the Bonferroni principle. This can be synthesized as:
  - If there are too many possible conclusions to draw, some will be true for purely statistical reasons, with no physical validity.
  - If you look in more places for interesting patterns than your amount of data will support, you are bound to find meaningless patterns.

Overview

- What is data mining?
- Data mining steps
- Data mining methods and sub-domains
- Data Preprocessing
- Summary

Method types

- Prediction methods (supervised machine leaning):
  - These methods use some variables to predict the values of other variables.
  - E.g. classification algorithms which build models that can be used for classifying new, unseen data, using known, labeled data.

Method types

- Description methods (unsupervised machine leaning):
  - Algorithms in this category find patterns that can describe the inner structure of the dataset.
  - E.g. clustering algorithms that find groups of similar objects in a dataset (called clusters) and possible isolated objects, far away from any cluster, called outliers.

Algorithms

- Supervised learning algorithms:
  - Classification
  - Regression
  - Deviation detection
- Unsupervised leaning algorithms:
  - Clustering
  - Association rule discovery
  - Sequential pattern discovery

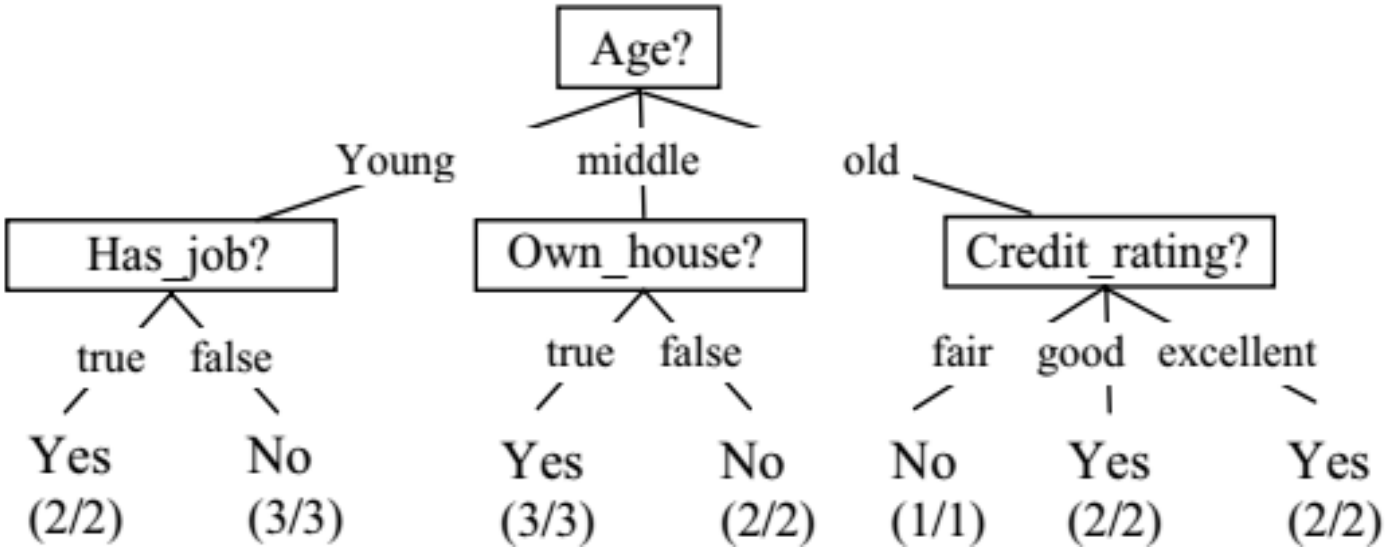
Classification

- **Input:**
  - A set of $\boldsymbol{k}$ classes $C = \{c_1, c_2, \ldots, c_k\}$
  - A set of $\boldsymbol{n}$ labeled items $D = \{(d_1, c_{i1}), (d_2, c_{i2}), \ldots, (d_n, c_{in})\}$.
  - The items are $d_1, \ldots, d_n$, each item d_j being labeled with class $c_j \in C$. D is called the **training set**.
  - For calibration of some algorithms a **validation set** is required. This validation set contains also labeled items not included in the training set.
- **Output:**
  - A **model** or **method** for classifying new items (a classifier).
  - The set of new items that will be classified using the model/method is called the **test set**

Classification Example [1]
- Using a decision tree classifier -

Age	Has_job	Own_house	Credit_rating	Class
young	false	false	fair	**No**
young	false	false	good	**No**
young	true	false	good	**Yes**
young	true	true	fair	**Yes**
young	false	false	fair	**No**
middle	false	false	fair	**No**
middle	false	false	good	**No**
middle	true	true	good	**Yes**
middle	false	true	excellent	**Yes**
middle	false	true	excellent	**Yes**
old	false	true	excellent	**Yes**
old	false	true	good	**Yes**
old	true	false	good	**Yes**
old	true	false	excellent	**Yes**
old	false	false	fair	**No**

Classification Example [1]
- Using a decision tree classifier -



- E.g. John is a old man, has a job but doesn't have a house and is credit is good.
- Will he get a new credit from the bank? Yes

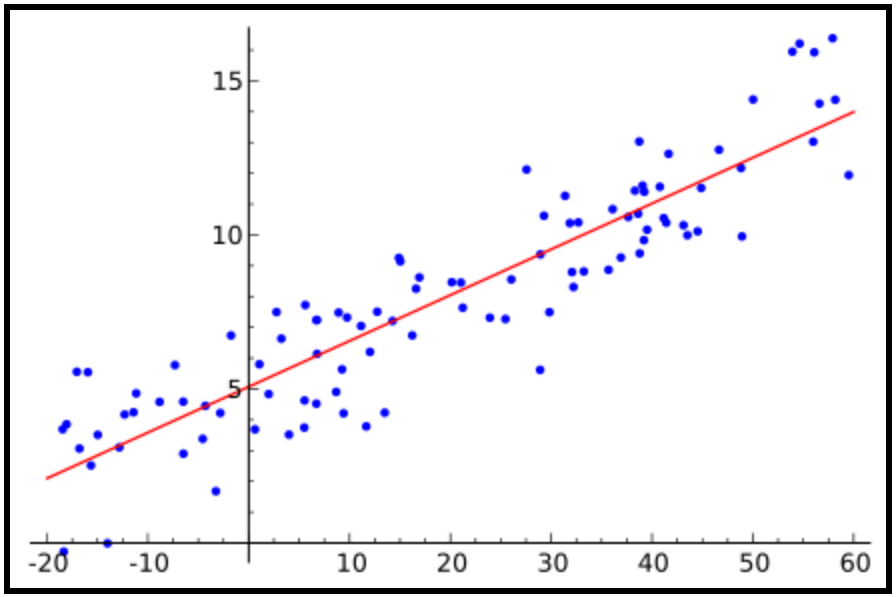
Regression

- Regression is a statistical model.
- Is used to predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency [5].
- Used in prediction and forecasting - its use overlaps machine learning.
- Regression analysis is also used to understand relationship between independent variables and dependent variable and can be used to infer causal relationships between them.

Regression

- There are many types of regression. For example, Wikipedia lists:
  - Linear regression model
  - Simple linear regression
  - Logistic regression
  - Nonlinear regression
  - Nonparametric regression
  - Robust regression
  - Stepwise regression

Linear Regression



Source:
http://en.wikipedia.org/wiki/File:Linear_regression.svg

Deviation detection

- Deviation detection or anomaly detection deals with discovering significant deviation from the normal behavior.
- Outliers are a significant category of abnormal data.
- Deviation detection can be used in many circumstances:
  - Data mining algorithm running stage: often such information may be important for business decisions and scientific discovery.
  - Auditing: such information can reveal problems or mal-practices.
  - Fraud detection in a credit card system: fraudulent claims often carry inconsistent information that can reveal fraud cases.
  - Intrusion detection in a computer network may rely on abnormal data.
  - Data cleaning (part of data preprocessing): such information can be detected and possible mistakes may be corrected in this stage.

Deviation detection

- Distance based techniques (example: k-nearest neighbor).
- One Class Support Vector Machines.
- Predictive methods (decision trees, neural networks).
- Cluster analysis based outlier detection.
- Pointing at records that deviate from association rules
- Hotspot analysis

Algorithms

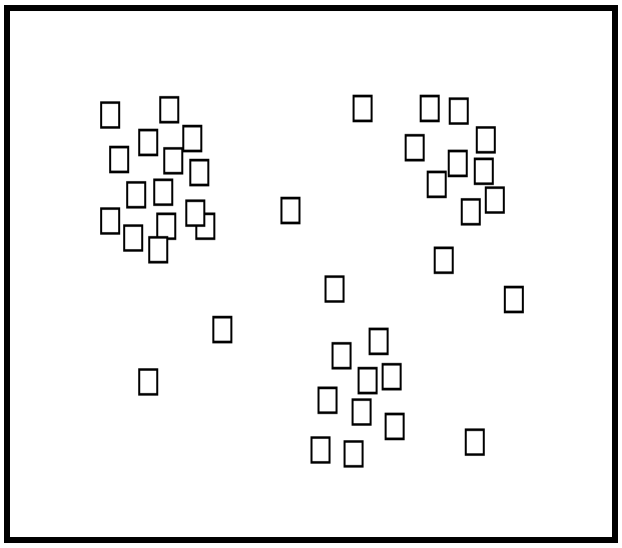
- Supervised learning algorithms:
  - Classification
  - Regression
  - Deviation detection
- Unsupervised leaning algorithms:
  - Clustering
  - Association rule discovery
  - Sequential pattern discovery

Clustering

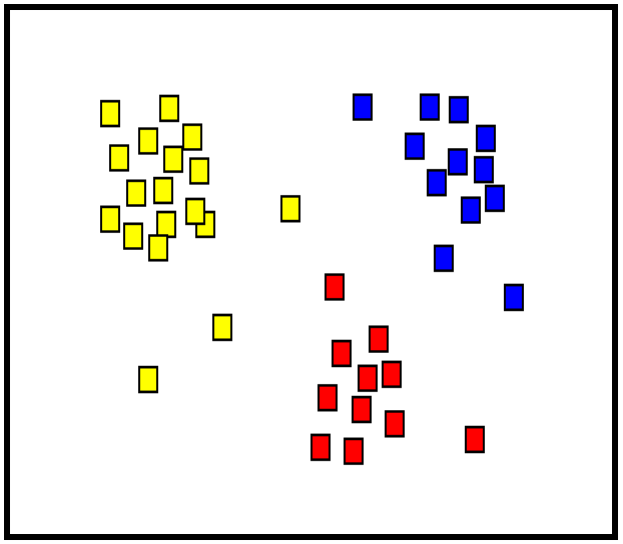
- **Input:**
  - A set of n objects $D = \{d_1, d_2, \dots, d_n\}$ (called usually points).
  - The objects are not labeled and there is no set of class labels defined.
  - A distance function (dissimilarity measure) that can be used to compute the distance between any two points. Low valued distance means 'near', high valued distance means 'far'.
  - Some algorithms also need a predefined value for the number of clusters in the produced result.
- **Output:**
  - A set of object (point) groups called clusters where points in the same cluster are near one to another and points from different clusters are far one from another, considering the distance function.

Clustering

- Having a set of points in a 2 dimensional space, find the natural clusters formed by these points.



INITIAL



AFTER CLUSTERING

Sources:
http://en.wikipedia.org/wiki/File:Cluster-1.svg
http://en.wikipedia.org/wiki/File:Cluster-2.svg

Association rule learning

- Let us consider:
  - A set of $\boldsymbol{m}$ items $I = \{i_1, i_2, \dots, i_m\}$.
  - A set of $\boldsymbol{n}$ transactions $T = \{t_1, t_2, \dots, t_n\}$, each transaction containing a subset of items from I, so if $t_k \in T$ then $t_k = \{i_{k1}, i_{k2}, \dots, i_{kj}\}$ where j depends on k.
- Then:
  - A **rule** is an implication with the following form:

$$X \rightarrow Y \text{ where } X, Y \subseteq I.$$

Association rule learning

- The **support** of a rule is the number/proportion of transactions containing the union between the left and the right part of the rule (and is equal with the support of this union as an itemset):

$$support(X) = \frac{|t \in T | X \subseteq t|}{|T|}$$

$$support(X \longrightarrow Y) = support(X \cup Y)$$

- The **confidence** of a rule is the proportion of transactions containing Y in the set of transactions containing X:

$$confidence(X \longrightarrow Y) = \frac{support(X \cup Y)}{support(X)}$$

- We accept a rule as a valid one if the support and the confidence of the rule are at least equal with some given **thresholds**.

Association rule learning

- **Input:**
  - A set of $\boldsymbol{m}$ items $I = \{i_1, i_2, \ldots, i_m\}$.
  - A set of $\boldsymbol{n}$ transactions $T = \{t_1, t_2, \ldots, t_n\}$, each transaction containing a subset of I, so if $t_k \in T$ then $t_k = \{i_{k1}, i_{k2}, \ldots, i_{kj}\}$ where j depends on k.
  - A threshold s for the support, given either as a percent or in absolute value.
  - If an itemset $X \in I$ is part of w transactions then w is the support of X.
  - If $w \geq s$ then X is called frequent itemset
  - A second threshold c for rule confidence.
- **Output:**
  - The set of frequent itemsets in T, having support $\geq s$
  - The set of rules derived from T, having support $\geq s$ and confidence $\geq c$

Association rule learning

- Consider the following set of transactions:

Transaction ID	Items
1	Bread, Milk, Butter, Orange Juice, Onion, Beer
2	Bread, Milk, Butter, Onion, Garlic, Beer, Orange Juice, Shirt, Pen, Ink, Baby diapers
3	Milk, Butter, Onion, Garlic, Beer
4	Orange Juice, Shirt, Shoes, Bread, Milk
5	Butter, Onion, Garlic, Beer, Orange Juice

- If $s = 0.6$ then $\{Bread, Milk, Orange\ Juice\}$ or $\{Onion, Garlic, Beer\}$ are frequent itemsets.
- Also, if $s = 0.6$ and $c = 0.7$ then the rule $\{Onion, Beer\} \rightarrow \{Garlic\}$ is a valid one because its support is 0.6 and the confidence is 0.75.

Sequential Pattern Discovery

- The model**:**
  - **Itemset**: a set of n distinct items
    $$I = \{i_1, i_2, \ldots, i_n\}$$
  - **Event**: a non-empty collection of items; we can assume that items are in a given order (e.g. lexicographic): $(i_1, i_2 \ \ldots, i_k)$
  - **Sequence**: an ordered list of events:
    $$\langle e_1, e_2 , \ldots, e_m \rangle$$

Sequential Pattern Discovery

- **Input:**
  - A set of sequences S (or a sequence database).
  - A Boolean function that can test if a sequence S_1 is included (or is a subsequence) of a sequence S_2. In that case S_2 is called a super sequence of S_1.
  - A threshold s (percent or absolute value) needed for finding frequent sequences.
- **Output:**
  - The set of frequent sequences, i.e. the set of sequences that are included in at least s sequences from S.
  - Sometimes a set of rules can be derived from the set of frequent sequences, each rule having the following form $S_1 \rightarrow S_2$ where S_1 and S_2 are sequences.

Sequential Pattern Discovery

- In a bookstore we can find frequent sequences like:

$$\{(\boldsymbol{BookOnC}, \boldsymbol{BookOnC++}), (\boldsymbol{BookOnC\#})\}$$

- From this sequence we can derive a rule like that: after buying books about C and C++, a customers buys books on C#:

$$\boldsymbol{BookOnC}, \boldsymbol{BookOnC++} \rightarrow \boldsymbol{BookOnC\#}$$

Overview

- What is data mining?
- Data mining steps
- Data mining methods and sub-domains
- Data Preprocessing
- Summary

Overview

- Data Preprocessing
  - Data Types
  - Measuring Data
  - Data cleaning
  - Data integration
  - Data transformation
  - Data reduction
  - Data discretization

Data Types

- Categorical vs. Numerical
- Scale types:
  - Nominal
  - Ordinal
  - Interval
  - Ratio

Categorical Data

- Categorical data consists in names representing some categories.
- This type of data belongs to a definable category.
- The values of this type of data are not ordered.
- Operations that can be performed on this type of data is equality and set inclusion.

Numerical Data

- Numerical data consists in numbers from a continuous or discrete set of values
- The values of this type of data are ordered
- Operations that can be performed on this data include $<, \leq, =, \geq, >$
- Conversions between categorical and numerical data are needed for data mining algorithms.

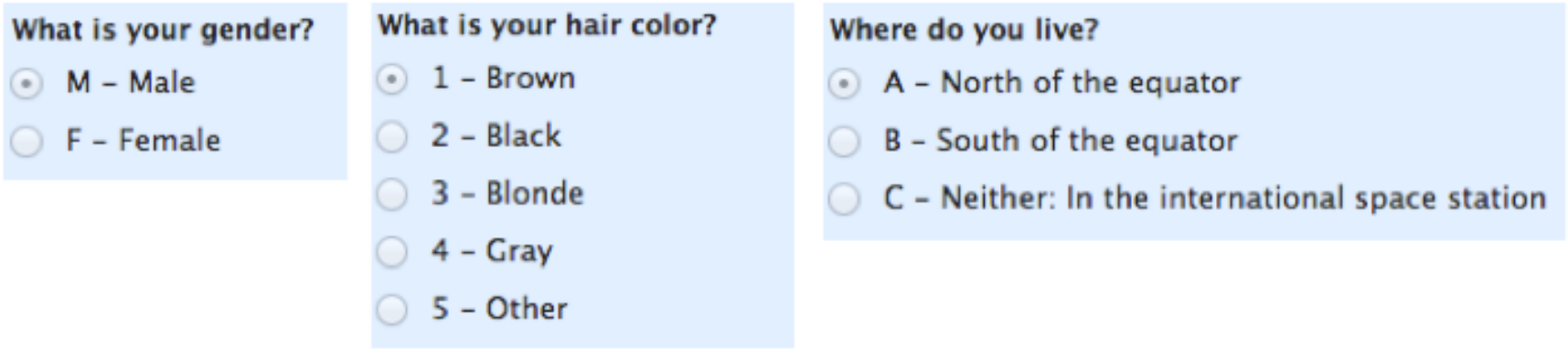
Scale Types

- Scale type (scales of measurement) is a taxonomy for the measurement levels
- The levels are:
  - Nominal
  - Ordinal
  - Interval
  - Ratio

Nominal

- Values belonging to a nominal scale are characterized by labels.
- Values are unordered and equally weighted.
- The mean or the median cannot be computed from a set of such values.
- The mode can be determined (frequent values)
- Nominal data are categorical but may be treated sometimes as numerical by assigning numbers to labels.

Nominal

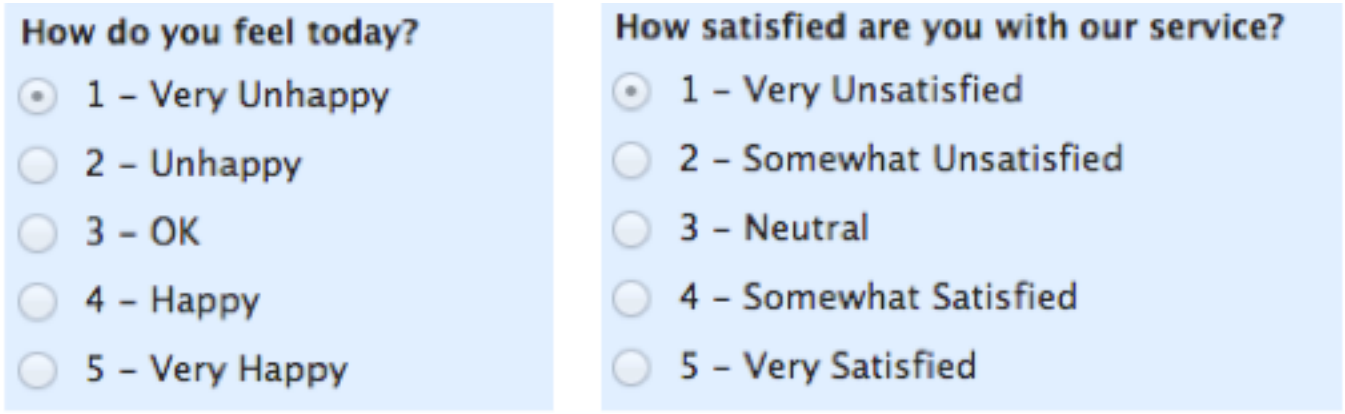


Source: http://www.mymarketresearchmethods.com/wp-content/uploads/2012/11/nominal-scales.png

Ordinal

- Values of this type are ordered but the difference or distance between two values cannot be determined, e.g. the military rank .
- The values only determine the rank order/position in the set.
- The mode and the median can be computed, but not the mean.
- These values are categorical in essence but can be treated as numerical because of the assignment of numbers (position in set) to the values

Ordinal



Source: http://www.mymarketresearchmethods.com/wp-content/uploads/2012/11/ordinal-scales.png

Interval

- These are numerical values.
- The difference between two values is meaningful for interval scaled data.
- The increments between values are known, consistent, and measurable.
- Zero does not mean ‘nothing’ but is somehow arbitrarily fixed. It is not a ‘true zero’
- Negative values are also allowed.
- The mean and the standard deviation can be computed.
- Regression can be used to predict new values.

Interval – Celsius example

- The increments between values are known, consistent, and measurable.
- E.g. the difference between two values is measurable
  - The difference between $60^o C$ and $50^o C$ is a measurable $10^o C$ as is the difference between $100^o C$ and $90^o C$ degrees.

Interval – Celsius example

- Zero does not mean 'nothing' but is somehow arbitrarily fixed.
- 0^oC is the melting point of water, why not the boiling point of water?
- There is no such thing as 'no temperature'.
- Without a true zero, it is impossible to compute ratios.
- With interval data, we can add and subtract, but cannot multiply or divide.
- E.g. $10^oC + 10^oC = 20^oC$ degrees, but 20^oC is not twice as hot as 10^oC degrees because there is no such thing as 'no temperature' when it comes to the Celsius scale.

Ratio

- Ratio scaled data are like interval scaled date but zero means ‘nothing’.
- Negative values are not allowed.
- The ratio between two values is meaningful.
- All mathematical operations can be performed, e.g. logarithms, geometric and harmonic means, coefficient of variation
- E.g.: age, temperature in Kelvin, mass in kilograms, length in meters, etc.

In conclusion

	Nominal	Ordinal	Interval	Ratio
Frequency	Yes	Yes	Yes	Yes
Median	No	Yes	Yes	Yes
Add or subtract	No	No	Yes	Yes
Mean, standard deviation, standard error of the mean.	No	No	Yes	Yes
Ratio, or coefficient of variation.	No	No	No	Yes

Binary Data

- Sometimes an attribute may have only two values, as the gender in a previous example. In that case the attribute is called binary.
  - **Symmetric binary:** when the two values are of the same weight and have equal importance (as in the gender case)
  - **Asymmetric binary:** one of the values is more important than the other. Example: a medical bulletin containing blood tests for identifying the presence of some substances, evaluated by 'Present' or 'Absent' for each substance. In that case 'Present' is more important that 'Absent'.
- Binary attributes can be treated as interval or ratio scaled but in most of the cases these attributes must be treated as nominal (binary symmetric) or ordinal (binary asymmetric)
- There are a set of similarity and dissimilarity (distance) functions specific to binary attributes

Overview

- Data Preprocessing
  - Data Types
  - Measuring Data
  - Data cleaning
  - Data integration
  - Data transformation
  - Data reduction
  - Data discretization

Measuring Data

- Measuring central tendency:
  - Mean
  - Median
  - Mode
  - Midrange (mid-extreme)
- Measuring dispersion:
  - Range
  - k^{th} percentile
  - Interquartile range (IQR)
  - Five-number summary
  - Standard deviation and variance

Central tendency

- **Central tendency** (or **measure of central tendency**) is a central or typical value for a probability distribution.

Central Tendency – Mean

- Given a set of n values $X = \{x_1, x_2 \ \ldots, x_n\}$.
- **Mean (μ):** The **arithmetic mean** or average value is:

$$\mu = \frac{x_1 \ + \ x_2 \ + \ \ldots \ + \ x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- If the values x have different weights, $w_1, w_2 \ \ldots, w_n$, then the **weighted arithmetic mean** or weighted average is computed instead of the arithmetic mean.
- Given a set of n values $X_w = \{(x_1, w_1), (x_2, w_2) \ldots, (x_n, w_n)\}$ the weighted arithmetic mean is:

$$\mu = \frac{x_1 w_1 \ + \ x_2 w_2 \ + \ \ldots \ + \ x_n w_n}{w_1 \ + \ w_2 \ + \ \ldots + w_n} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

- If the extreme values are eliminated from the set (smallest 1% and biggest 1%) a **trimmed mean** is obtained.

Central Tendency – Median

- The **median (m)** value of an ordered set is the middle value in the set.
- If n is odd then the median is the middle value of the set, e.g.: for $\{1,3,5,7,9\}$ the median is $m = 5$.
- If n is even then the median is the mean of the middle values, e.g. for $\{1,3,5,7,9,11\}$ the median is $m = 6 (= \frac{5+7}{2})$.

Central tendency – Mode

- The **mode (f)** of a dataset is the most frequent value.
- A dataset may have more than a single mode. For 1, 2 and 3 modes the dataset is called unimodal, bimodal and trimodal.
- When each value is present only once there is no mode in the dataset.
- For a unimodal dataset the mode is a measure of the central tendency of data. For these datasets we have the empirical relation:

$$\mu - f = 3 \cdot (\mu - m)$$

Central tendency – Midrange

- The **midrange (M)** of a set of values is the arithmetic mean of the largest and the smallest value.
- Given a set $X = \{x_1, x_2, \dots, x_n\}$ then the midrange is:

$$M = \frac{max_{x \in X}(x) + min_{x \in X}(x)}{2}$$

Dispersion

- **Dispersion** is a quantifiable variation of measurements of differing members of a population

Dispersion – Range

- The **range** is the difference between the largest and smallest values.
- Given a set $X = \{x_1, x_2, \dots, x_n\}$, the range is:

$$M = max_{x \in X}(x) - min_{x \in X}(x)$$

Dispersion – k^{th} percentile

- The k^{th} percentile is a value of x_i having the property that k percent of the values in the set are less or equal to it.
- Examples:
  - the median is the 50^{th} percentile.
  - The most used percentages are the median and the 25^{th} and 75^{th} percentiles, also named **quartiles**
  - Notation: $Q1$ for 25%, $Q2$ for 50% and $Q3$ for 75%.

Dispersion – k^{th} percentile

- **Computing method**: There are more than one different methods for computing $Q1$, $Q2$ and $Q3$.
- The most obvious method is the following:
  - Put the values of the data set in ascending order
  - Compute the median using its definition. It divides the ordered dataset into two halves (lower and upper), neither one including the median.
  - The median value is $Q2$
  - The median of the lower half is $Q1$ (or the lower quartile)
  - The median of the upper half is $Q3$ (or the upper quartile)

Dispersion – Interquartile Range

- **Interquartile range** (IQR) is the difference between $Q3$ and $Q1$:

$$IQR = Q3 - Q1$$

- Potential outliers are values more than $1.5\ x\ IQR$ below $Q1$ or above $Q3$

Dispersion – Five-number summary

- **Five-number summary**. Sometimes the median and the quartiles are not enough for representing the spread of the values
- The smallest and biggest values must be considered also.
- $(Min, Q1, Median, Q3, Max)$ is called the five-number summary.

Dispersion – Standard Deviation

- **Standard deviation** is a measure that is used to quantify the amount of variation or dispersion of a set of data values.
- The standard deviation measures the spread of the values around the mean value.
- Given a set $X = \{x_1, x_2, \dots, x_n\}$, the standard deviation is:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \text{ where } \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- $\sigma = 0$ only when all values are identical.
- The square of standard deviation is called **variance**(σ^2).
- **Variance** measures how far the values from a set are spread out from their mean

Overview

- Data Preprocessing
  - Data Types
  - Measuring Data
  - Data cleaning
  - Data integration
  - Data transformation
  - Data reduction
  - Data discretization

Data Cleaning

- The main objectives of data cleaning are:
  - Replace (or remove) missing values,
  - Smooth noisy data,
  - Remove or just identify outliers

NULL Values

- When a NULL value is present in data it may be:
  - ***Legal NULL value:*** Some attributes are allowed to contain a NULL value. In such a case the value must be replaced by something like ‘Not applicable’ and **not** a NULL value.
  - ***Missing value:*** The value existed at measurement time but was not collected.

Missing values

- May appear from various reasons:
  - human/hardware/software problems,
  - data not collected (considered unimportant at collection time),
  - deleted data due to inconsistencies, etc.

Dealing With Missing Data

- There are two solutions in handling missing data:

1. Ignore the data point / example with missing attribute values. If the number of errors is limited and these errors are not for sensitive data, removing them may be a solution.

Dealing With Missing Data

2. Fill in the missing value. This may be done in several ways:
   - Fill in manually. This option is not feasible in most of the cases due to the huge volume of the datasets that must be cleaned.
   - Fill in with a (distinct from others) value ‘not available’ or ‘unknown’.
   - Fill in with a value measuring the central tendency, for example attribute mean, median or mode.
   - Fill in with a value measuring the central tendency but only on a subset (for example, for labeled datasets, only for examples belonging to the same class).
   - The most probable value, if that value may be determined, for example by decision trees, expectation maximization (EM), Bayes, etc.

Smooth noisy data

- The noise can be defined as a random error or variance in a measured variable
- Wikipedia define noise as a colloquialism for recognized amounts of unexplained variation in a sample.
- For removing the noise, some smoothing techniques may be used:
  - Regression (was presented in first course)
  - Binning

Binning

- **Data binning** or **bucketing** is a data pre-processing technique used to reduce the effects of minor observation errors.
- The original data values which fall in a given small interval, a bin, are replaced by a value representative of that interval, often the central value.
- It is a form of quantization.

Source: https://en.wikipedia.org/wiki/Data_binning

Binning

- Binning can be used for smoothing an ordered set of values.
- Smoothing is made based on neighbor values.
- There are two steps:
  - Partitioning ordered data in several bins. Each bin contains the same number of examples (data points).
  - Smoothing for each bin: values in a bin are modified based on some bin characteristics: mean, median, boundaries.

Binning

- Consider the following ordered set of values for some attribute:

  {1, 3, 4, 8, 9, 13, 16, 17, 20, 24, 34, 56, 78, 80, 82}

Initial bins	Binning using mean	Binning using median	Binning using boundaries
1, 3, 4, 8, 9	5, 5, 5, 5, 5	4, 4, 4, 4, 4	1, 1, 1, 9, 9
13, 16, 17, 20, 24	18, 18, 18, 18, 18	17, 17, 17, 17, 17	13, 13, 13, 24, 24
34, 56, 78, 80, 82	66, 66, 66, 66, 66	78, 78, 78, 78, 78	34, 34, 82, 82, 82

- The new set can be one of the following:
- Binning using mean: {5, 5, 5, 5, 5, 18, 18, 18, 18, 18, 66, 66, 66, 66, 66}
- Binning using median: {4, 4, 4, 4, 4, 17, 17, 17, 17, 17, 78, 78, 78, 78, 78}
- Benning using boundaries: {1, 1, 1, 9, 9, 13, 13, 13, 24, 24, 34, 34, 82, 82, 82}

Outliers

- An outlier is an attribute value numerically distant from the rest of the data.
- Outliers may be sometimes correct values: for example, the salary of the CEO of a company may be much bigger that all other salaries. But in most of the cases outliers are and must be handled as noise.
- Outliers must be identified and then removed (or replaced, as any other noisy value) because many data mining algorithms are sensitive to outliers.
- For example any algorithm using the arithmetic mean (e.g. k-means) may produce erroneous results because the mean is very sensitive to outliers.

Identifying Outliers

- **IQR**
  - Values more than 1.5 x IQR below Q1 or above Q3 are potential outliers.
  - Boxplots may be used to identify these outliers (boxplots are a method for graphical representation of data dispersion).
- **Standard deviation**
  - Values that are more than two standard deviations away from the mean for a given attribute are also potential outliers.
- **Clustering**
  - After clustering a certain dataset some points are outside any cluster or far away from any cluster center.

Overview

- Data Preprocessing
  - Data Types
  - Measuring Data
  - Data cleaning
  - Data integration
  - Data transformation
  - Data reduction
  - Data discretization

Data Integration

- Data integration means merging data from different data sources into a coherent dataset.
- The main steps are:
  - Schema integration
  - Remove duplicates
  - Remove redundant information (redundancy)
  - Handle inconsistencies

Schema Integration

- Must identify the translation of every source scheme to the final scheme (entity identification problem)
- Subproblems:
  - The same attributes can have different names depending on the data source. Example: the customer id may be called Cust-ID, Cust#, CustID, CID in different sources.
  - Attributes which represent different information have the same names. Example: for employees data, the attribute 'City' means city where resides in a source and city of birth in another source.

Duplicates

- Duplicates: The same information may be stored in many data sources.
- Merging them can cause sometimes duplicates of that information:
  - as duplicate attribute (same attribute with different names is found multiple times in the final result) or
  - as duplicate instance (same object/entity is found multiple times in the final database).
- These duplicates must be identified and removed.

Redundancy

- Redundancy: Some information may be deduced / computed.
- For example
  - age may be deduced from birthdate,
  - annual salary may be computed from monthly salary and other bonuses recorded for each employee.
- Redundancy must be removed from the dataset before running the data mining algorithm
- Note that in existing data warehouses some redundancy is allowed.

Inconsistencies

- Inconsistencies are conflicting values for a set of attributes.
- Example: Birthdate = January 1, 1980, Age = 12 represents an obvious inconsistency but we may find other inconsistencies that are not so obvious.
- For detecting inconsistencies extra knowledge about data is necessary: for example, the functional dependencies attached to a table scheme can be used.
- Available metadata describing the content of the dataset may help in removing inconsistencies

Overview

- Data Preprocessing
  - Data Types
  - Measuring Data
  - Data cleaning
  - Data integration
  - Data transformation
  - Data reduction
  - Data discretization

Data Transformation

- Data is transformed and summarized in a better form for the data mining process:
  - Normalization
  - New attribute construction
  - Summarization using aggregate functions

Normalization

- All attributes are scaled to fit a specified range:
  - 0 to 1,
  - -1 to 1 or generally
  - |v| <= r where r is a given positive value.
- Needed when the importance of some attributes is bigger only because the range of the values of that attributes is bigger.
- Example: Euclidian distance between A(0.5, 101) and B(0.01, 2111) is ≈ 2010, determined almost exclusively by the second dimension.

Normalization

- Given a set $X = \{x_1, x_2, \ldots, x_n\}$
- **Min-max normalization**:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- **Min-max normalization** for positive numbers

$$x' = \frac{x}{x_{max}}$$

- **Z-score** normalization (σ is the standard deviation, μ is the mean):

$$x' = \frac{x - \mu}{\sigma}$$

Normalization

- **Decimal scaling** (apply only if $\max(|x|) \geq 1$):

$$\boldsymbol{x' = \frac{x}{10^c}}$$

where c is the smallest integer such that $\max(|x'|) < 1$

$$\max(|x'|) = \max\left(\frac{|x|}{10^c}\right) = \frac{\max(|x|)}{10^c}$$

$$\frac{\max(|x|)}{10^c} < 1 \Leftrightarrow \max(|x|) < 10^c \Leftrightarrow$$

$$c > \log_{10} \max(|x|) \;\&\; c \; is \; the \; smallest \; integer$$

$$\Longrightarrow c = [\log_{10} \max(|x|)] + 1$$

Normalization

- **L1-norm normalization**

$$x' = \frac{x}{\sum_{i=1}^{n} x_i}$$

- **L2-norm normalization**

$$x' = \frac{x}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

Feature construction

- New attribute construction is called also ***feature construction***.
- It means: building new attributes based on the values of existing ones.
- Example: if the dataset contains an attribute ‘Color’ with only three distinct values {Red, Green, Blue} then three attributes may be constructed: ‘Red’, ‘Green’ and ‘Blue’ where only one of them equals 1 (based on the value of ‘Color’) and the other two 0.
- Another example: use a set of rules, decision trees or other tools to build new attribute values from existing ones. New attributes will contain the class labels attached by the rules / decision tree used / labeling tool

Summarization

- At this step aggregate functions may be used to add summaries to the data.
- Examples: adding sums for daily, monthly and annual sales, counts and averages for a number of customers or transactions, and so on.
- All these summaries are used for the OLAP operations in multidimensional modeling (usually in data warehousing).
- The result is a data cube and each summary information is attached to a level of granularity

Overview

- Data Preprocessing
  - Data Types
  - Measuring Data
  - Data cleaning
  - Data integration
  - Data transformation
  - Data reduction
  - Data discretization

Data Reduction

- Not all information produced by the previous steps is needed for a certain data mining process.
- Reducing the data volume by keeping only the necessary attributes leads to a better representation of data and reduces the time for data analysis.

Data Reduction Methods

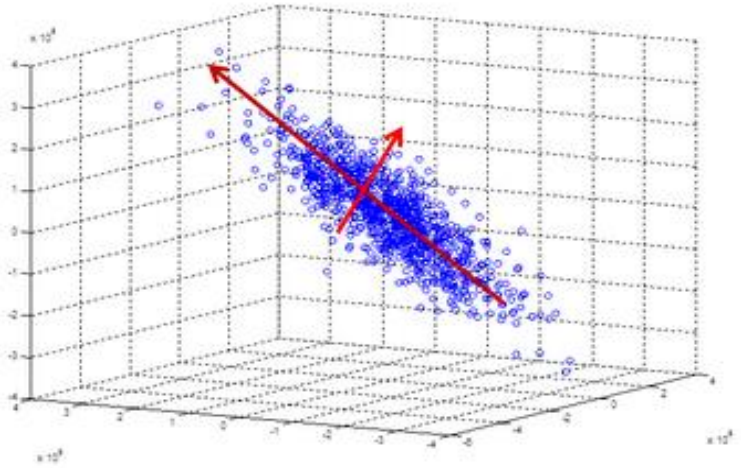
- **Data cube aggregation**
- **Attribute selection** by keeping only relevant attributes:
  - stepwise forward selection (start with an empty set and add attributes),
  - stepwise backward elimination (start with all attributes and remove some of them one by one)
  - a combination of forward selection and backward elimination.
  - decision tree induction: after building the decision tree, only attributes used for decision nodes are kept.

Data Reduction Methods

- **Dimensionality reduction**: encoding mechanisms are used to reduce the data set size or compress data.
- A popular method is Principal Component Analysis (PCA): given N data vectors having n dimensions, find $K \leq N$ orthogonal vectors (called principal components) that can be used for representing data.

PCA Example

An extreme example of data(over the cutoff) scattering



source: http://2011.igem.org/Team:USTC-Software/parameter

Data Reduction Methods

- **Numerosity reduction**: the data are replaced by smaller data representations such as parametric models (only the model parameters are stored in this case) or nonparametric methods: clustering, sampling, histograms.
- **Discretization**

Overview

- Data Preprocessing
  - Data Types
  - Measuring Data
  - Data cleaning
  - Data integration
  - Data transformation
  - Data reduction
  - Data discretization

Data Discretization

- Discretization transferring continuous functions, models, and equations into discrete counterparts.
- There are many data mining algorithms that cannot use continuous attributes.
- Using discretization these values can be replaced with discrete ones.
- Even for discrete attributes, is better to have a reduced number of values leading to a reduced representation of data.
- This may be performed by concept hierarchies
- Discretization means reducing the number of values for a given continuous attribute by dividing its values in intervals.
- Each interval is labeled and each attribute value will be replaced with the interval label.

Data Discretization Methods

- Binning: equi-width bins or equi-frequency bins may be used. Values in the same bin receive the same label.
- Histograms: like binning, histograms partition values for an attribute in buckets. Each bucket has a different label and labels replace values.
- Entropy based intervals: each attribute value is considered a potential split point (between two intervals) and an information gain is computed for it (reduction of entropy by splitting at that point). Then the value with the greatest information gain is picked. In this way intervals may be constructed in a top-down manner.
- Cluster analysis: after clustering, all values in the same cluster are replaced with the same label (the cluster-id for example)

Concept hierarchies

- Usage of a concept hierarchy to perform discretization means replacing low-level concepts (or values) with higher level concepts.
- Example: replace the numerical value for age with young, middle-aged or old.
- For numerical values, discretization and concept hierarchies are the same.

Concept hierarchies

- For categorical data the goal is to replace a bigger set of values with a smaller one (categorical data are discrete by definition):
  - Manually define a partial order for a set of attributes. For example the set {Street, City, Department, Country} is partially ordered, Street $\subseteq$ City $\subseteq$ Department $\subseteq$ Country. In that case we can construct an attribute 'Localization' at any level of this hierarchy, by using the n rightmost attributes (n = 1 .. 4).
  - Specify (manually) high level concepts for value sets of low level attribute values associated with. For example {Muntenia, Oltenia, Dobrogea} $\subseteq$ Tara_Romaneasca.
  - Automatically identify a partial order between attributes, based on the fact that high level concepts are represented by attributes containing a smaller number of values compared with low level ones.

Overview

- What is data mining?
- Data mining steps
- Data mining methods and sub-domains
- Data Preprocessing
- Summary

Summary

- This first course presented:
  - A list of alternative definitions of Data Mining and some examples of what is Data Mining and what is not Data Mining
  - A discussion about the researchers communities involved in Data Mining and about the fact that Data Mining is a cluster of subdomains
  - The steps of the Data Mining process from collecting data located in existing repositories (data warehouses, archives or operational systems) to the final evaluation step.
  - A brief description of the main subdomains of Data Mining with some examples for each of them.

Summary

- Data types: categorical vs. numerical, the four scales (nominal, ordinal, interval and ratio) and binary data.
- A short presentation of data preprocessing steps and some ways to extract important characteristics of data:
  - central tendency (mean, mode, median, etc.)
  - dispersion (range, IQR, five-number summary, standard deviation and variance).
- A description of every preprocessing step:
  - Cleaning
  - Integration
  - Transformation
  - Reduction
  - Discretization

References

[1] Liu, Bing. *Web data mining: exploring hyperlinks, contents, and usage data, second edition*. Springer Science & Business Media, 2011.

[2] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." *AI magazine* 17.3 (1996): 37.

[3] Leskovec, Jure, Anand Rajaraman, and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2014.

[4] Kimball, Ralph, and Margy Ross. *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons, 2011.

[5] Pang-Ning Tan, Michael Steinbach, Vipin Kumar. *Introduction to Data Mining*, Adisson-Wesley, 2006