

2. Data preprocessing

Road Map



- Data types
- Measuring data
- Data cleaning
- Data integration
- Data transformation
- Data reduction
- Data discretization
- Summary

Data types



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

Categorical vs. Numerical



- ❑ **Categorical data**, consisting in names representing some categories, meaning that they belong to a definable category. Example: color (with categories red, green, blue and white) or gender (male, female).
- ❑ The values of this type are **not ordered**, the usual operations that may be performed being equality and set inclusion.
- ❑ **Numerical data**, consisting in numbers from a **continuous** or **discrete** set of values.
- ❑ Values are **ordered**, so testing this order is possible (<, >, etc).
- ❑ Sometimes we must or may convert categorical data in numerical data by assigning a numeric value (or code) for each label.

Scale types



Stanley Smith Stevens, director of the Psycho-Acoustic Laboratory, Harvard University, proposed in a 1946 Science article that all measurement in science are using four different types of scales:

- Nominal
- Ordinal
- Interval
- Ratio

Nominal



- ❑ Values belonging to a nominal scale are characterized by **labels**.
- ❑ Values are **unordered** and **equally weighted**.
- ❑ We cannot compute the mean or the median from a set of such values
- ❑ Instead, we can determine the **mode**, meaning the value that occurs most frequently.
- ❑ Nominal data **are categorical** but may be treated sometimes as numerical by assigning numbers to labels.

Ordinal



- ❑ Values of this type are **ordered** but the difference or distance between two values cannot be determined.
- ❑ The values only determine **the rank order** /position in the set.
- ❑ Examples: the military rank set or the order of marathoners at the Olympic Games (without the times)
- ❑ For these values we can compute the mode or the median (the value placed in the middle of the ordered set) 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

Interval



- ❑ These are **numerical** values.
- ❑ For interval scaled attributes the **difference** between two values is meaningful.
- ❑ Example: the temperature using Celsius scale is an interval scaled attribute because the difference between 10 and 20 degrees is the same as the difference between 40 and 50 degrees.
- ❑ **Zero does not mean 'nothing'** but is somehow arbitrarily fixed. For that reason **negative** values are also allowed.
- ❑ We can compute the mean, the standard deviation or we can use regression to predict new values.

Ratio



- Ratio scaled attributes are like interval scaled attributes but zero means 'nothing'.
- Negative** values are not allowed.
- The ratio between two values is meaningful.
- Example: age - a 10 years child is two times older than a 5 years child.
- Other examples: temperature in Kelvin, mass in kilograms, length in meters, etc.
- All mathematical operations can be performed, for example logarithms, geometric and harmonic means, coefficient of variation

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 than '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.

Road Map



- Data types
- Measuring data**
- Data cleaning
- Data integration
- Data transformation
- Data reduction
- Data discretization
- Summary

Measuring data



Measuring central tendency:

- Mean
- Median
- Mode
- Midrange

Measuring dispersion:

- Range
- Kth percentile
- IQR
- Five-number summary
- Standard deviation and variance

Central tendency (1)



□ Consider a set of n values of an attribute: x_1, x_2, \dots, x_n .

□ **Mean:** The **arithmetic mean** or average value is:

$$\mu = (x_1 + x_2 + \dots + x_n) / n$$

□ If the values x have different weights, w_1, \dots, w_n , then the **weighted arithmetic mean** or weighted average is:

$$\mu = (w_1x_1 + w_2x_2 + \dots + w_nx_n) / (w_1 + w_2 + \dots + w_n)$$

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

Central tendency (2)



- Median:** The median value of an ordered set is **the middle value** in the set.
- Example: Median for {1, 3, 5, 7, 1001, 2002, 9999} is 7.
- If n is even the median is the mean of the middle values: the median of {1, 3, 5, 7, 1001, 2002} is 6 (arithmetic mean of 5 and 7).

- Mode:** The mode 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:
$$\text{mean} - \text{mode} = 3 \times (\text{mean} - \text{median})$$

Central tendency (3)



- ❑ **Midrange.** The midrange of a set of values is the arithmetic mean of the largest and the smallest value.
- ❑ For example the midrange of $\{1, 3, 5, 7, 1001, 2002, 9999\}$ is 5000 (the mean of 1 and 9999).

Dispersion (1)



- ❑ **Range.** The range is the difference between the largest and smallest values.
- ❑ Example: for $\{1, 3, 5, 7, 1001, 2002, 9999\}$ range is $9999 - 1 = 9998$.
- ❑ **k^{th} percentile.** The k^{th} percentile is a value x_j belonging of that dataset and having the property that k percent of the values are less or equal than x_j .
- ❑ Example: the median is the 50^{th} percentile.
- ❑ The most used percents are the median and the 25^{th} and 75^{th} percentiles, called also **quartiles** (notation: Q1 for 25% and Q3 for 75%).

Dispersion (2)



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

$$\text{IQR} = \text{Q3} - \text{Q1}$$

- ❑ Potential outliers are values more than $1.5 \times \text{IQR}$ below Q1 or above Q3.
- ❑ **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 (3)



- **Standard deviation.** The standard deviation of n values (observations) is:

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

- The square of standard deviation is called **variance**.
- The standard deviation measures the spread of the values around the mean value.
- A value of 0 is obtained only when all values are identical.

Road Map



- Data types
- Measuring data
- Data cleaning
- Data integration
- Data transformation
- Data reduction
- Data discretization
- Summary

Objectives



- The main objectives of data cleaning are:
 - Replace (or remove) missing values,
 - Smooth noisy data,
 - Remove or just identify outliers
- Some attributes are allowed to contain a NULL value.
- In these cases the value stored in the database (or the attribute value in the dataset) must be something like 'Not applicable' and not a NULL value.

Missing values (1)



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

Missing values (2)



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 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, it 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 ([Han, Kamber 06]).
- ❑ 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:
 1. Regression (was presented in first course)
 2. 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.

Example



- Consider the following ordered data for some attribute: 1, 2, 4, 6, 9, 12, 16, 17, 18, 23, 34, 56, 78, 79, 81

Initial bins	Use mean for binning	Use median for binning	Use bin boundaries for binning
1, 2, 4, 6, 9	4, 4, 4, 4, 4	4, 4, 4, 4, 4	1, 1, 1, 9, 9
12, 16, 17, 18, 23	17, 17, 17, 17, 17	17, 17, 17, 17, 17	12, 12, 12, 23, 23
34, 56, 78, 79, 81	66, 66, 66, 66, 66	78, 78, 78, 78	34, 34, 81, 81, 81

Result



So the smoothing result is:

- ❑ Initial: 1, 2, 4, 6, 9, 12, 16, 17, 18, 23, 34, 56, 78, 79, 81
- ❑ Using the mean: 4, 4, 4, 4, 4, 17, 17, 17, 17, 17, 66, 66, 66, 66, 66
- ❑ Using the median: 4, 4, 4, 4, 4, 17, 17, 17, 17, 17, 78, 78, 78, 78, 78
- ❑ Using the bin boundaries: 1, 1, 1, 9, 9, 12, 12, 12, 23, 23, 34, 34, 81, 81, 81

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 than 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 (one of them is k-means) may produce erroneous results because the mean is very sensitive to outliers.

Identifying outliers



- ❑ Use of **IQR**: values more than $1.5 \times \text{IQR}$ 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).
- ❑ Use of **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).

Road Map



- Data types
- Measuring data
- Data cleaning
- Data integration**
- Data transformation
- Data reduction
- Data discretization
- Summary

Objectives



Data integration means merging data from different data sources into a coherent dataset.

The main activities are:

- Schema integration
- Remove duplicates and redundancy
- Handle inconsistencies

Schema integration



- ❑ Must **identify the translation** of every source scheme to the final scheme (entity identification problem)
- ❑ Subproblems:
 - ❑ The same thing is **called differently** in every data source. Example: the customer id may be called Cust-ID, Cust#, CustID, CID in different sources.
 - ❑ Different things are **called with the same name** in different sources. 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 twice in the final result) or
 - ❑ as duplicate instance (same object is found twice in the final database).
- ❑ These duplicates must be identified and removed.

Redundancy



- **Redundancy**: Some information may be deduced / computed from others.
- 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 also be removed from the dataset before running the data mining algorithm
- Note that in existing data warehouses some redundancy is sometimes 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.

Road Map



- Data types
- Measuring data
- Data cleaning
- Data integration
- Data transformation
- Data reduction
- Data discretization
- Summary

Objectives



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

Normalization



- ❑ All attribute data are scaled to fit a specified range:
 - ❑ 0 to 1,
 - ❑ -1 to 1 or generally
 - ❑ $|v| \leq r$ 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



We can achieve normalization using:

□ **Min-max normalization:**

$$v_{\text{new}} = (v - v_{\text{min}}) / (v_{\text{max}} - v_{\text{min}})$$

□ For positive values the formula is:

$$v_{\text{new}} = v / v_{\text{max}}$$

□ **z-score** normalization (σ is the standard deviation):

$$v_{\text{new}} = (v - v_{\text{mean}}) / \sigma$$

□ **Decimal scaling:** $v_{\text{new}} = v / 10^n$

where n is the smallest integer for that all numbers become (as absolute value) less than the range r (for $r = 1$, all new values of v are ≤ 1) then

Feature construction



- New attribute construction is called also *feature construction*.
- 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 annually sales, counts and averages for number of customers or transactions, and so on.
- All these summaries are used for the 'slice and dice' process when data is stored in a data warehouse.
- The result is a data cube and each summary information is attached to a level of granularity.

Road Map



- Data types
- Measuring data
- Data cleaning
- Data integration
- Data transformation
- Data reduction
- Data discretization
- Summary

Objectives



- 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.

Reduction methods (1)



Methods that may be used for data reduction (see [Han, Kamber 06]) :

- Data cube aggregation**, already discussed.
- Attribute selection**: keep only relevant attributes. This can be made by:
 - 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.

Reduction methods (2)



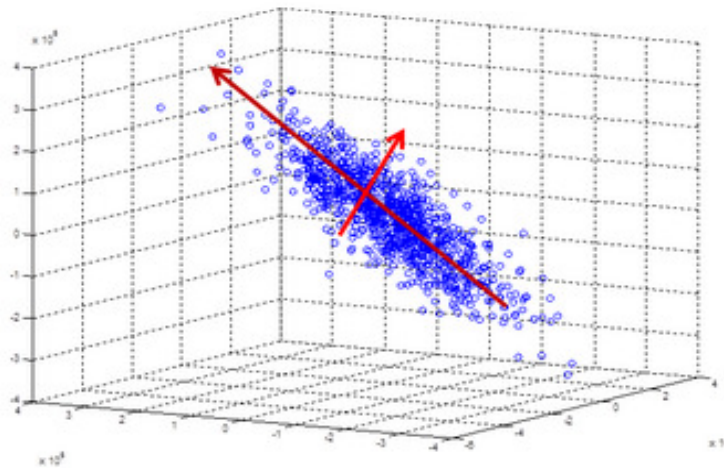
- ❑ **Dimensionality reduction**: encoding mechanisms are used to reduce the data set size or compress data.
- ❑ A popular method is Principal Component Analysis (PCA): given N data vectors having n dimensions, find $k \leq n$ orthogonal vectors (called principal components) that can be used for representing data.
- ❑ A PCA example is presented on the following slide, for a multivariate Gaussian distribution centered at 1, 3 (source: wikipedia)

PCA example



PCA for a multivariate Gaussian distribution (source:
<http://2011.igem.org/Team:USTC-Software/parameter>)

*An extreme example of data(over the cutoff)
scattering*



Reduction methods (3)



- **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** and concept hierarchy generation, discussed in the following paragraph.

Road Map



- Data types
- Measuring data
- Data cleaning
- Data integration
- Data transformation
- Data reduction
- Data discretization
- Summary

Objectives



- ❑ There are many data mining algorithms that cannot use continuous attributes. Replacing these continuous values with discrete ones is called **discretization**.
- ❑ 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 (1)



- ❑ 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.
- ❑ Some of the most popular methods to perform discretization are:
 1. **Binning**: equi-width bins or equi-frequency bins may be used. Values in the same bin receive the same label.

Discretization (1)



- Popular methods to perform discretization - cont:
 2. **Histograms**: histograms partition values for an attribute in buckets. Each bucket has a different label and labels replace values.
 3. **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.
 4. **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, $\text{Street} \subseteq \text{City} \subseteq \text{Department} \subseteq \text{Country}$. In that case we can construct an attribute 'Localization' at any level of this hierarchy, by using the n rightmost attributes ($n = 1 \dots 4$).
 - ❑ Specify (manually) high level concepts for value sets of low level attribute values associated with. For example $\{\text{Muntenia, Oltenia, Dobrogea}\} \subseteq \text{Tara_Romaneasca}$.
 - ❑ 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.

Summary



This second course presented:

- 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) and dispersion (range, IQR, five-number summary, standard deviation and variance).
- A description of every preprocessing step:
 - cleaning,
 - integration,
 - transformation,
 - reduction and
 - Discretization

- Next week: Association rules and sequential patterns

References



- ➡ [Han, Kamber 06] Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition, Morgan Kaufmann Publishers, 2006, 47-101
- ➡ [Stevens 46] Stevens, S.S, On the Theory of Scales of Measurement. Science June 1946, 103 (2684): 677–680.
- ➡ [Wikipedia] Wikipedia, the free encyclopedia, en.wikipedia.org
- ➡ [Liu 11] Bing Liu, 2011. CS 583 Data mining and text mining course notes, <http://www.cs.uic.edu/~liub/teach/cs583-fall-11/cs583.html>