

Data Mining

(Mining Knowledge from Data)

Data Preprocessing

Marcel Jiřina, Pavel Kordík



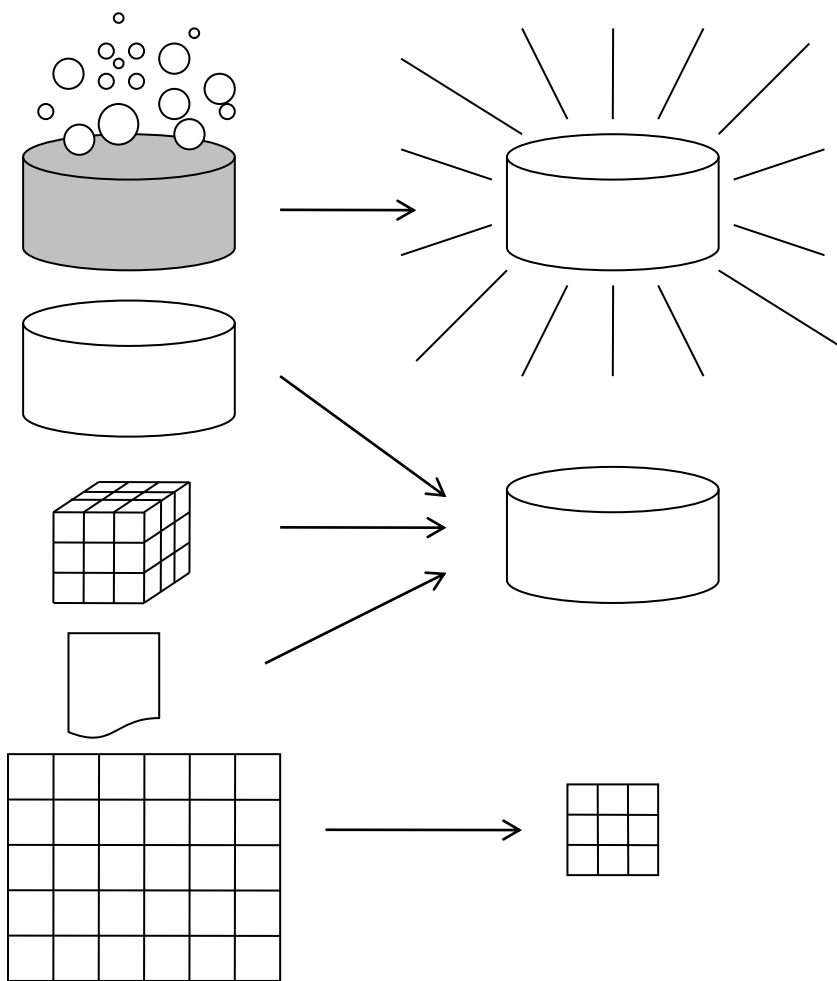
ČESKÉ
VYSOKÉ
UČENÍ
TECHNICKÉ
V PRAZE

FIT

Motivation

- Data in the form in which they are stored in databases are usually not suitable for analysis and modeling
- They may be incomplete, inconsistent, contain erroneous data...
- The data needed for analysis may be stored in different places and in different formats.
- "Garbage in, Garbage out" - if we use data of poor quality for modeling, the result will also be poor.

Parts of data preprocessing



Data cleansing

Data integration

Data reduction

$-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

Data transformation

Data preprocessing

- The input data for preprocessing are raw data in databases, data warehouses, files...
- The output data are in a form suitable for modeling - e.g. training set matrix in the form:

instances

attributes

ID	Name	address	City	Zip	Phone
1	Alan	1800 Bon Ave.	Elk Grove	95758	916-333-4444
2	Tom	600 Bender Rd.	Sacramento	95412	916-112-2345
3	Sam	300 Tent St.	San Jose	95112	408-345-2134

- Different methods require different handling data

Data validation

- The aim is to determine the quality of the data and find incorrect values
- To validate data we use any external information about values of the attributes

Methods of data validation

- Checking data types
- Checking the range of attributes
- Comparing the values with values of other instances
- Checking consistency
 - e.g. If pregnant = yes then sex \neq male

Missing values

- Missing value may indicate that none of possible values is appropriate or it has not be recorded during data collecting
- Recognition of these cases is usually not easy

ID	Name	Sandwitch	Sauce
1	Alan	Turkey	Sweet Union
2	Tom	Ham	
3	Sam	Beef	Thousand Island

- Tom had a sandwich without sauce or the salesman forgot to record it?

Dealing with missing data

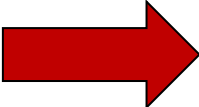
- Removing instances with missing values
- Replacing by zero
- Replacing by mean/modus
- Replacing by mean/modus of the nearest K instances
- Regression or classification model for predicting the value of the attribute

Converting attribute type

- We have different types of attributes, see previous lecture
 - Binomial
 - Nominal
 - Ordinal
 - Numerical
- Many methods support all types of attributes => for specific methods attributes must be converted to the correct type

Nominal -> binomial

- For each different value of a nominal attribute we create a new attribute.
- Coding “1 of N”

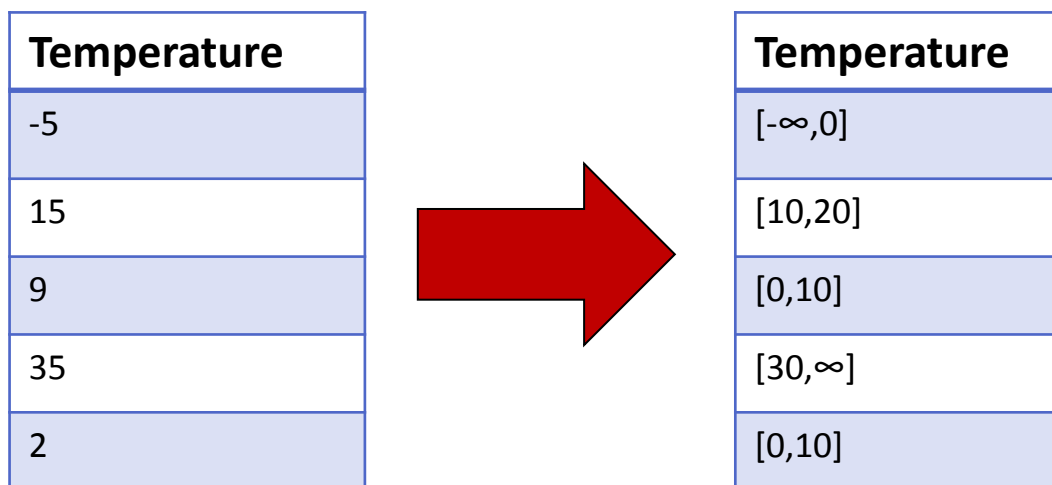


Weather		Weather = Rainy	Weather = Sunny	Weather = Cloudy
Rainy		1	0	0
Sunny		0	1	0
Cloudy		0	0	1

- It can not be used if the attribute has many different values. In that case, we divide values into groups and use only group (binning).

Discretization

- Transfer Numerical -> Nominal
- Otherwise also “binning”
- Dividing/distribution of continuous variables into multiple intervals
- The number of intervals must be chosen in advance



Equal-width discretization

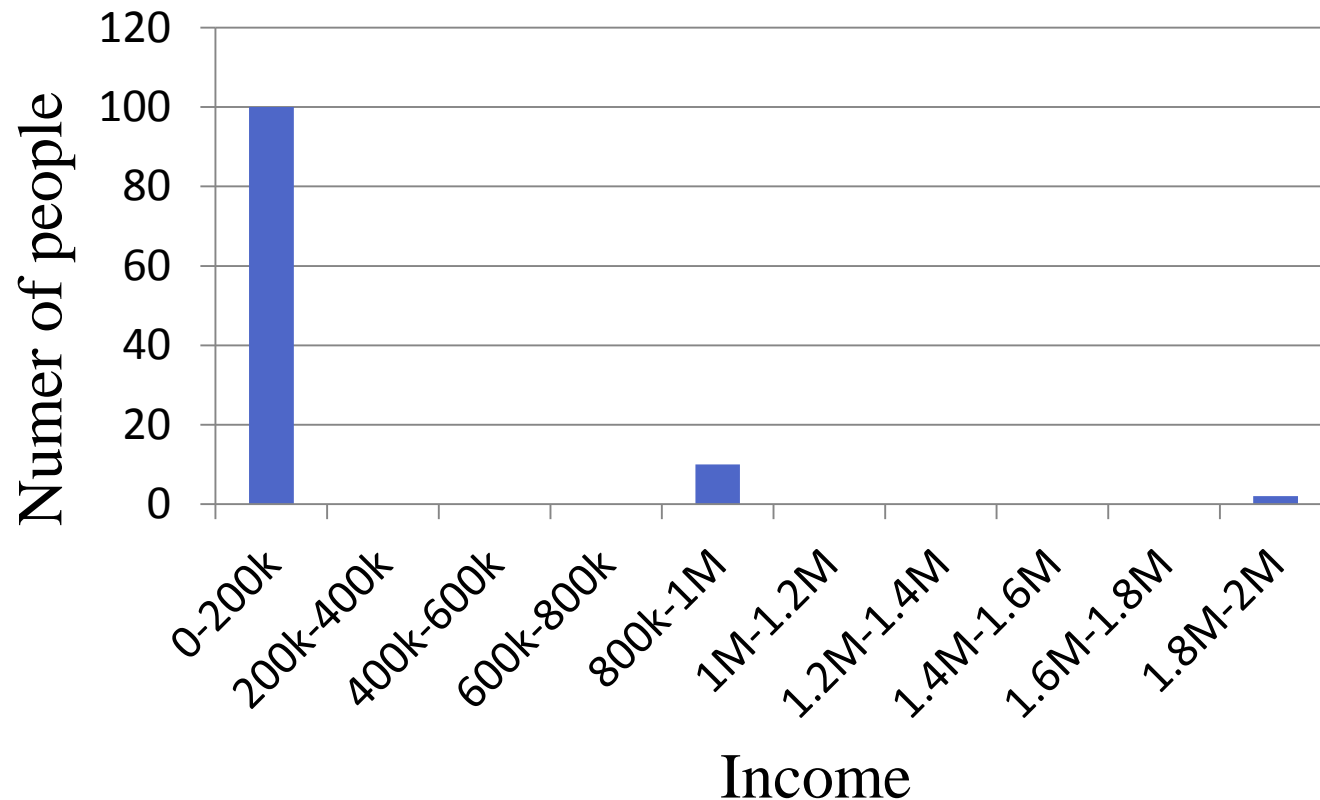
- Dividing/distribution of values into equal intervals
- The size of the interval is $(\text{Max}-\text{Min}) / \text{bins}$



- Some intervals may not contain any value
- The simplest method
- Outliers can cause uneven distribution
- Not suitable for askew distributed data

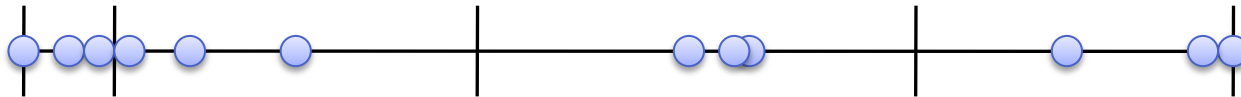
Equal-width discretization

- Problems in the presence of outliers:



Equal-height discretization

- Also binning by the frequency
- Intervals are selected so that nearly the same number of values would appear in each interval



- Askew distributions are not the problem
- Usually better results than Equal-width

Other methods

- Supervised discretization
 - Boundaries of intervals are chosen so that the data are divided the best way into classes
 - Discretization by entropy, information gain...
 - More in the lecture about decision trees

Normalization

- Some methods work only with a limited range of values ([0,1], [-1, + 1])
 - For example, some types of neural networks
- Methods based on distances badly manage different ranges of values of attributes
 - Which instance is the most similar to the last instance?

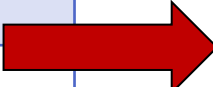
Number of children	Annual salary
0	200000
5	205000
5	201000

- The Euclidean distance is smaller for the first row
- If the range of values of an attribute is significantly lower, then the Euclidean distance does not depend on it

Normalization

- The solution if the normalization

Number of children	Annual salary
0	200000
5	205000
5	201000



Number of children	Annual salary
0	0
1	1
1	0,2

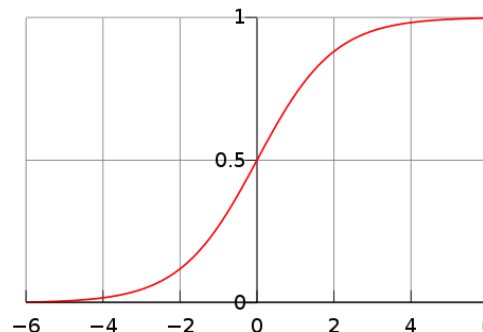
Min-max normalization

- Values are linearly transformed into a new range of values, usually $[0,1]$ or $[1,1]$
- $$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$
- Distribution remains the same
- We need to know the minimum and maximum value
- Problems with outliers

Soft-max normalization

- Nonlinear transformation using the sigmoid function

- $$v' = \frac{1}{1 + e^{-\left(\frac{v - \mu}{\sigma}\right)}}$$



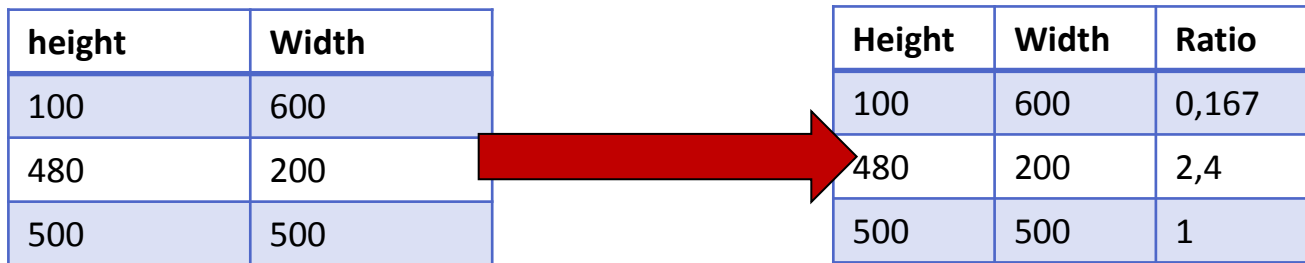
- Transforming into the interval $[0,1]$ (it can be adjusted-shifted/scaled for another interval)
- Within a distance of standard deviations from the mean the data are transformed almost linearly
- Data further from the mean are transformed nonlinearly
- We need not know the minimum and maximum values
- Outliers not reduce the resolution for other values

Other types of normalization

- Z-Score
 - Normalization with zero mean and one standard deviation
 - $v' = \frac{v - \mu}{\sigma}$
- Decimal scaling
 - Decimal point shift
 - $v' = \frac{v}{10^j}$
- Logarithmic scale
 - Used when the values of attributes differ by more orders
 $v' = \log_a(v)$

Creating new attributes

- Sometimes we can create derived attributes, which will help to improve the model



The diagram illustrates the process of creating a derived attribute. A red arrow points from a table with two columns to a table with three columns. The first table has columns 'height' and 'Width'. The second table has columns 'Height', 'Width', and 'Ratio'. The data in the second table is derived from the first table by calculating the ratio of height to width.

height	Width
100	600
480	200
500	500

Height	Width	Ratio
100	600	0,167
480	200	2,4
500	500	1

- For example, we can create a new attribute *ratio* from the attributes *width* and *height* of an image

Reduction of the number of instances

- Random selection of representative samples
- Sampling with replacement
 - Always choose a random element from all options
 - One instance can be selected more times
- Sampling without replacement
 - Once selected element can not be selected again
- Stratified sampling
 - Sampling with replacement when the resulting class ratio (number of samples in each class) is selected in advance

Reduce the number of attributes

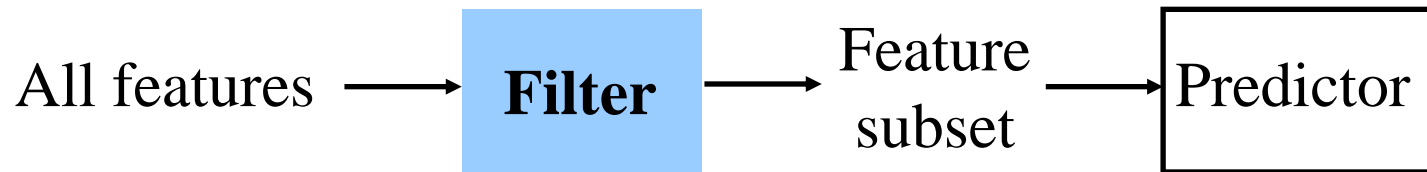
- Raw data can contain irrelevant, noisy, redundant attributes
- The training matrix should have at least about an order of magnitude more instances than attributes
- There are two approaches to the solution:
 - Selecting the most relevant subset of attributes
 - Transformation of the data into the less-dimensional space

Attribute (feature) selection

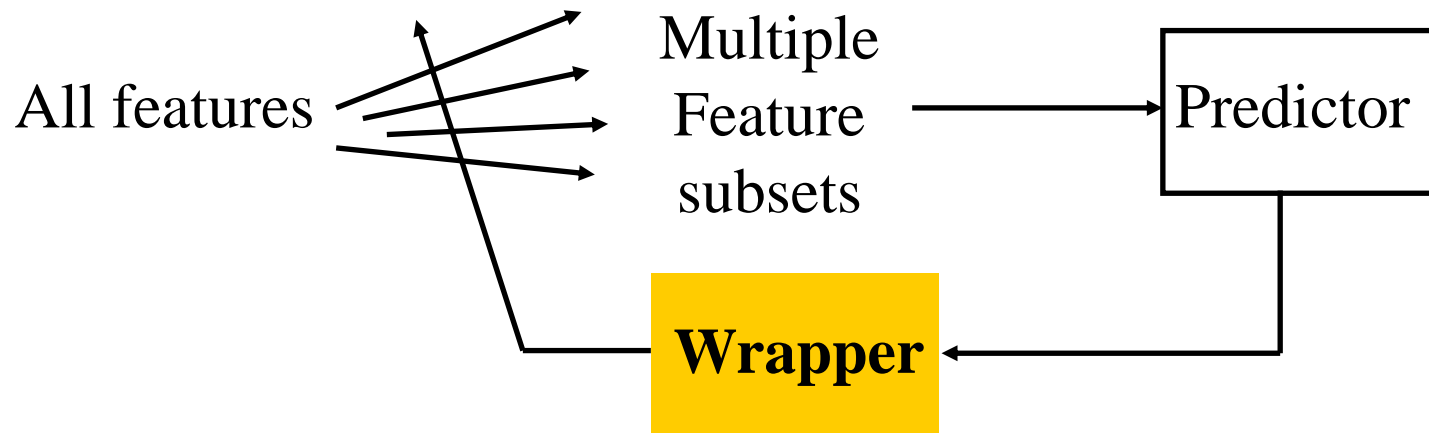
- Univariate / Multivariate
 - Univariate methods always evaluate the contribution of one attribute
 - Multivariate methods evaluate a chosen subset of attributes as a whole
- Filter/Wrapper/Embedded
 - Filter methods evaluate attributes independently the model used
 - Wrapper methods use the error of the model as a criterion for selecting attributes
 - Embedded methods - feature selection is a part of the learning/training process of the model

Filter / Wrapper

- Filter:



- Wrapper:



Attribute (feature) selection

- Various methods of attribute/feature selection consist of three basic steps:
 - Searching: method of selecting subsets of attributes
 - Rating: method of assessing the relevance of attributes or a group of attributes, statistical indicators for the filter method or the error of a model for the wrapper methods
 - Stopping rule for the search - adding new attribute does not improve the result, or the maximum number of attributes...

The search methods

- Possible attributes of n attributes is 2^n
- Complete search is time consuming
- Heuristic methods:
 - They do not guarantee finding of the optimal solution
 - Forward selection
 - Getting started with an empty subset of attributes
 - In each step the best attribute is added to the subset until the stopping condition is fulfilled
 - Backward elimination
 - Getting started with all the attributes and the least relevant attribute is removed in every step
- Random selection, evolutionary methods, ...

Methods for assessing the relevance of attributes

- Correlation coefficient
- Mutual information
- T-test
- Misclassification error
- ...

Conclusion

- Preprocessing of data:
 - Validation
 - Missing data (data imputation)
 - Conversion, encoding
 - Outliers
 - Discretization
 - Normalization
 - Reduction of attributes
 - Reduction of instances
 - ...