Database and Information Systems

11. Deductive Databases
12. Data Warehouses and OLAP
13. Data Mining
14. Index Structures for Similarity Queries
15. Semi-Structured Data
16. Document Retrieval
17. Web Mining
18. Content Extraction
19. Multimedia Data

What is Data Mining?

- What is Data Mining?
- Data-Mining as a Process
- Data Preprocessing
- Data Mining Tasks

Data Mining

Readings:

Dunham, Margaret H.: Data Mining - Introductory and Advanced Topics. Pearson Education, 2003, Chapter 3-6.

Kantardzic, Mehmet: Data Mining - Concepts, Models, Methods, and Algorithms. Wiley-Interscience 2003, Chapter 2-9.

Larose, Daniel T.: Data mining methods and models, Wiley-Interscience, 2006.

Tan, Pang-Ning ; Steinbach, Michael ; Kumar, Vipin: Introduction to data mining, Pearson/Addison-Wesley, 2006.

What is Data Mining?

- What is Data Mining?
- Data-Mining as a Process
- Data Preprocessing
- Data Mining Tasks

What is Data Mining?

- mining as a metaphor:
    - the search for the precious thing amongst the overburden
    - goal is not always clearly specified
        - what's precious?
        - successful even if looking for gold but found diamonds
    - mining is an explorative activity
        - finding cues
        - making hypotheses
        - evaluating hypotheses
        - getting the precious stuff

What is Data Mining?

- ”data mining” is a misleading metaphor!

	mining	data mining
target	coal, gold, ore, ...	trends, associations
overburden	dirt, rock	”useless” data

- alternative notion:
  - KDD: Knowledge Discovery in Databases
- alternative view:
  - KDD is a complex process, DM only one step in it

Example Tasks

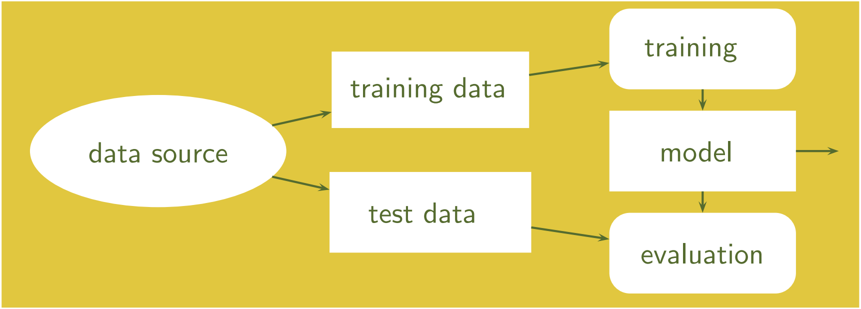
- customer relationship management
  - grouping of customer populations (tailored marketing)
  - prediction of customer behaviour (individualized marketing)
  - risk assessment (risky credits, fraudulent credit card use)
- fault analysis
  - interdependencies between faults
  - interdependencies between maintenance procedures and faults
- time-series analysis
  - stock market development
  - event prediction (stock market crashes, bankruptcies)

Why is Data Mining special?

- There is no silver bullet for data mining!
  - many different techniques
  - extremely laborious parameter tuning
  - very few clues for performance predictions
- data mining aims at partially correct solutions
  - vague task descriptions
  - no perfect solution methods
  - issues: quality, portability to similar application domains
- contrast: quering a data base
  - precise semantics
  - 100% correct results must be guaranteed
  - issues: performance, maintenance, ...

Evaluation

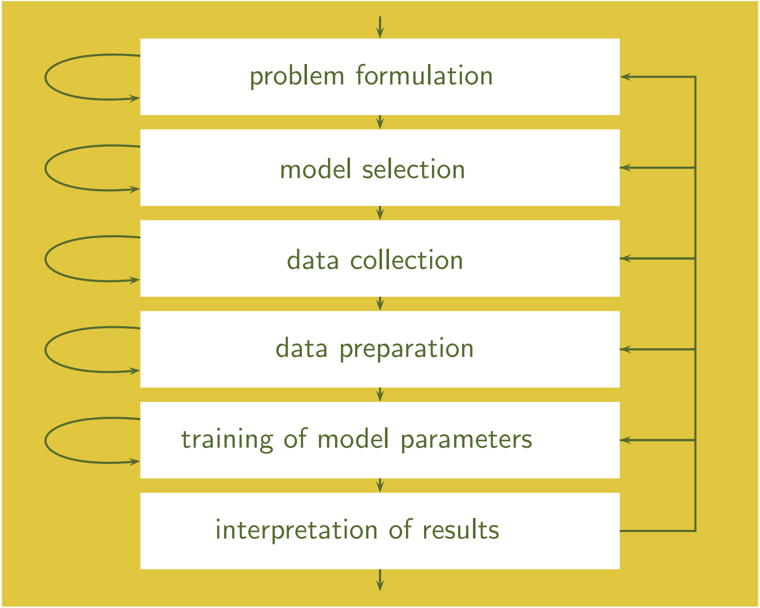
- models in data-mining are usually developed and evaluated on data collections
- for development purposes a clear separation between training and test data is necessary



Data Mining as a Process

- What is Data-Mining?
- Data-Mining as a Process
- Data Preprocessing
- Data Mining Tasks

Data Mining as a Process



problem formulation

model selection

data collection

data preparation

training of model parameters

interpretation of results

Data-Mining as a Process

1. problem formulation
   - based on domain-specific knowledge and experience
   - Which $\left\{ \begin{array}{c} \text{categories} \\ \text{dependencies} \end{array} \right\}$ are of particular interest?
2. selection of a suitable model class
   - based on the available and feasible data mining techniques
   - Which kind of model seems most promising given the available data?

Data-Mining as a Process

3. data collection
    - designed experiment:
        - data can be generated on demand
        - sampling distribution is known
    - observational approach:
        - random data collection
        - sampling distribution is unknown or implicit in the data collection procedure
    - data collection affects the outcome

Data-Mining as a Process

4. data preparation
    - data cleansing
        - outliers: measurement, coding or recording errors
        - outliers: abnormal but natural values
        - missing values
    - outlier treatment
        - only if outlier detection is not the goal
        - either removal as part of preprocessing
        - or application of techniques insensitive to outliers
    - data transformation
        - rescaling: making features comparable
        - dimensionality reduction: gaining efficiency
        - binning: map numerical values to qualitative classes

Data-Mining as a Process

5. training of model parameters
    - estimation of stochastic parameters
    - adjustment of threshold values
    - adjustment of weights
    - ...

    - training is usually an optimization problem
6. evaluation and interpretation of results
    - defining metrics for quality assessment
    - measuring the quality of results on held out test data
    - summarization and visualization of results

    - usually simple models ...

        ... are better trainable

        ... are better interpretable

        ... but less accurate

CRISP-DM

- Cross industry standard process for data mining
- life-cycle model

1. Business understanding phase
    - analysis of objectives and requirements
    - problem definition
    - initial strategy development
2. Data understanding phase
    - data collection
    - exploratory data analysis
    - assessment of data quality
3. Data preparation phase
    - cleansing, transformation etc.

CRISP-DM

4. Modelling phase
   - selection of modelling techniques and tools
   - parameter tuning / optimization
   - data analysis
5. Evaluation Phase
   - evaluation of the model
   - comparison of the outcome to the initial objectives
   - deployment decision
6. Deployment phase
   - reporting
   - transfer to other application cases
   - if applicable: introduction into day-to-day business

Data Preprocessing

- What is Data-Mining?
- Data-Mining as a Process
- Data Preprocessing
- Data Mining Tasks

Data Preprocessing

- Data Types
- Metrics
- High-Dimensional Data Spaces
- Missing Data
- Outlier Detection
- Time-Dependent Data
- Dimensionality Reduction
- Number of Values Reduction
- Sample Size Reduction

Data Types

Dienstag 24. April 2007, 09:00 Uhr

Berlin (AP) Mit seiner rekordverdächtigen Trockenheit fällt der Frühling 2007 auch unter den Aspekten der globalen Erwärmung aus der Rolle. ”Wir gehen eher davon aus, dass im Frühjahr etwas mehr Niederschlag ist”, sagte Wolfgang Kusch, der Präsident des Deutschen Wetterdienstes, am Dienstag im ZDF-Morgenmagazin. Ein signifikanter Rückgang der Niederschläge werde dagegen für den Sommer erwartet. ”Das zeigen auch unsere Statistiken jetzt schon.”

Der Anstieg der Temperaturen lasse sich vor allem bei der langfristigen Wetterbeobachtung feststellen, sagte Kusch. Seit Beginn der Wetteraufzeichnungen seien die Temperaturen stets angestiegen. Seit 1901 habe sich die durchschnittliche Temperatur in Deutschland von knapp acht auf knapp neun Grad plus um 0,9 Grad erwärmt. ”Das sind immerhin mehr als zehn Prozent”, sagte Kusch. ”Das sind ganz gravierende Veränderungen.”

Von den zehn wärmsten Jahren dieser Periode lägen neun in der Zeit zwischen 1989 und jetzt. Der frühere Ausreißer war 1934. Der Anstieg habe sich beschleunigt. ”Von 1990 bis 2000 war das absolut wärmste Jahrzehnt”, sagte der Chef des Wetterdiensts. ”Und dieses Jahrzehnt ist wieder auf Rekordkurs.” Nicht nur die Mitteltemperatur nehme zu, auch die Extreme würden zunehmen, sagte er voraus. Gerade der Nordosten Deutschlands werde deutlich trockener.

Data Types

Tuesday, April 24th, 2007, 09:00 am

Berlin (AP) With its possibly record-breaking period of dryness spring time 2007 must be considered exceptional even from the perspective of global warming. "In the long run we expect a higher level of precipitation in spring", says Wolfgang Kusch, president of the German Weather Forcasting Service, tuesday morning in the ZDF-Morgenmagazin. A significant drop in the amount of rain is expected for summer time. "Our statistics do confirm this already."

The rise in temperature becomes most obvious in longterm weather observations, says Kusch. From the earliest recordings till now temperature increased continuously. Since 1901 the average temperature in Germany rose from almost eight to almost nine degree centigrade by 0.9 degree. "That's already more than 10 percent ", said Kusch. "We are witnessing really serious changes."

Among the ten warmest years of this period nine occurred between 1989 and now. The earlier outlier was the year 1934. Also, the speed of increase has accelerated. "The period from 1990 to 2000 was the absolutely warmest decade", said the head of the Wheather Forcasting Service. "And the current decade is again on a record-marking track." Not only the average temperature increases, also extreme conditions become more frequent. In particular North-East Germany will become significantly more dry.

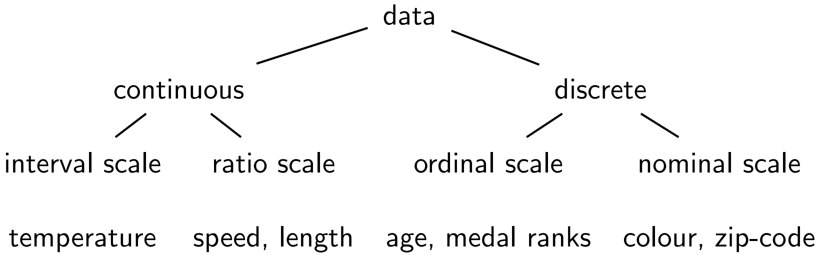
Data Types

- continuous (quantitative, metric)
    - measurement with (theoretically) infinite precision
    - interval scale
        - no problem-specific zero point
        - example: temperature
    - ratio scale
        - problem specific zero point (absence of the feature)
        - example: speed, length

Data Types

- discrete
  - ordinal scale (discretized)
    - problem-specific order relation
    - no distance relation
    - example: age, income classes, medal ranks
    - scale need not be linear
  - nominal scale (qualitative, categorical)
    - no problem-specific order relation
    - no distance relation
    - example: color, zip-code

Data Types



data

continuous

discrete

interval scale

ratio scale

ordinal scale

nominal scale

temperature

speed, length

age, medal ranks

colour, zip-code

Data Types

- temporal characteristics
  - static
    - no change over time
  - dynamic (temporal)
    - change over time
  - periodic
    - no problem-specific order relation
    - distance relation
    - example: days of the week

Data Preprocessing

- Data Types
- Metrics
- High-Dimensional Data Spaces
- Missing data
- Outlier Detection
- Time-Dependent Data
- Dimensionality Reduction
- Number of Values Reduction
- Sample Size Reduction

Metrics

- many data mining techniques are based on some notion of similarity/dissimilarity
- i.e. MINKOVSKIJ distance

$$d(\vec{x_1}, \vec{x_2}) = \left(\sum_{i=1}^{n} |x_{1i} - x_{2i}|^m \right)^{1/m}$$

- $m = 1$: Manhattan distance, city block distance
- $m = 2$: EUCLIDIAN distance
- $m = \infty$: max-distance, TCHEBYCHEV distance

Metrics

- special case: binary data → Hamming distance

$$|x_{1i} - x_{2i}| = \begin{cases} 0 & \text{if } x_{1i} = x_{2i} \\ 1 & \text{if } x_{1i} \neq x_{2i} \end{cases}$$

$$d(\vec{x_1}, \vec{x_2}) = \left(\sum_{i=1}^{n} |x_{1i} - x_{2i}|^m \right)^{1/m} = \left(\sum_{i=1}^{n} |x_{1i} - x_{2i}| \right)^{1/m}$$

$$= \left(n_{0/1} + n_{1/0} \right)^{1/m}$$

- root function is monotonic, m can be chosen as 1

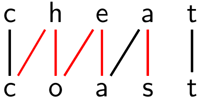
$$d(\vec{x_1}, \vec{x_2}) = n_{0/1} + n_{1/0}$$

Metrics

- self identity: $\forall \vec{x} \ . \ d(\vec{x}, \vec{x}) = 0$
- positivity: $\forall \vec{x_1} \neq \vec{x_2} \ . \ d(\vec{x_1}, \vec{x_2}) > 0$
- symmetry: $\forall \vec{x_1}, \vec{x_2} \ . \ d(\vec{x_1}, \vec{x_2}) = d(\vec{x_2}, \vec{x_1})$
- triangle inequation: $\forall \vec{x_1}, \vec{x_2}, \vec{x_3} \ . \ d(\vec{x_1}, \vec{x_3}) \leq d(\vec{x_1}, \vec{x_2}) + d(\vec{x_2}, \vec{x_3})$

Metrics

- metrics for complex objects, i.e. strings
- string edit distance, LEVENSHTEIN metric
- minimal effort to transform a sequence into another one
- basic operations
  - substitution
  - insertion
  - deletion
- alignment: pairwise, order preserving mapping between the elements of the two strings
- alternative alignments with same distance possible



Metrics

- string edit distance is a non-deterministic, recursive function

$$d(x_{0:0}, y_{0:0}) = 0$$

$$d(x_{1:m}, y_{0:0}) = d(x_{0:0}, y_{1:m}) = m$$

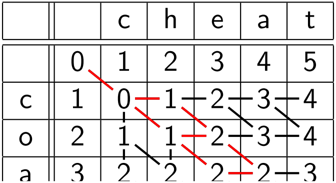
$$d(x_{1:m}, y_{1:n}) = \min \begin{cases} d(x_{2:m}, y_{2:n}) + d(x_1, y_1) \\ d(x_{1:m}, y_{2:n}) + 1 \\ d(x_{2:m}, y_{1:n}) + 1 \end{cases}$$

Metrics

- finding the minimum distance is an optimization problem
  → dynamic programming

local distances

		c	h	e	a	t
	0	1	1	1	1	1
c	1	0	1	1	1	1
o	1	1	1	1	1	1
a	1	1	1	1	0	1
s	1	1	1	1	1	1
t	1	1	1	1	1	0



		c	h	e	a	t
	0	1	2	3	4	5
c	1	0	1	2	3	4
o	2	1	1	2	3	4
a	3	2	2	2	2	3

global distances

		c	h	e	a	t
	0	1	2	3	4	5
c	1					
o	2					
a	3					
s	4					
t	5					

Data Preprocessing

- Data Types
- Metrics
- High-Dimensional Data Spaces
- Missing data
- Outlier Detection
- Time-Dependent Data
- Dimensionality Reduction
- Number of Values Reduction
- Sample Size Reduction

High-Dimensional Data Spaces

- all high-dimensional data spaces are sparse
  - keeping the same data-point density in a space with more dimensions requires exponentially more data points
  - to enclose a prespecified portion of data points, an increasingly large part of the hypercube needs to be "encircled":

$$e(p) = p^{\frac{1}{n}}$$

portion	dimensionality	edge length
p	n	$e(p)$
0.1	1	0.100
0.1	2	0.316
0.1	3	0.464
	...	
0.1	10	0.794

High-Dimensional Data Spaces

- paradoxes (cont.)
    - almost every point is closer to an edge of the cube than to another sample point
    - almost every point is an outlier

Data Preprocessing

- Data Types
- Metrics
- High-Dimensional Data Spaces
- Missing Data
- Outlier Detection
- Time-Dependent Data
- Dimensionality Reduction
- Number of Values Reduction
- Sample Size Reduction

Missing Data

- some data mining tools are insensitive to missing data
- ignore all incomplete tuples
    - might result in the loss of a substantial amount of data
- manual completion
- automatic completion
    - using a global constant
    - using the global mean value
    - using a class-dependent mean value
    - use a predictive model, e.g. based on a correlation with other features

Data Preprocessing

- Data Types
- Metrics
- High-Dimensional Data Spaces
- Missing Data
- Outlier Detection
- Time-Dependent Data
- Dimensionality Reduction
- Number of Values Reduction
- Sample Size Reduction

Outlier Detection

- not applicable if the application is aimed at outlier detection e.g. fraudulent credit card transactions
- the task: find k out of n tuples, which are
    - considerably dissimilar
    - exceptional
    - inconsistent with the remaining data

Outlier Detection

1. manual detection supported by visualization tools
   - only for low dimensional data
2. statistical methods
   - threshold for the variance, e.g. two times variance
   - only applicable if the distribution is known
3. using domain knowledge
   - value restrictions, e.g. $0 \leq \text{age} < 150$
4. distance-based detection
   - applicable also for multi-dimensional data
   - a sample is an outlier if it has not enough neighbors

Example

- sample data

$$S = [s_1, ..., s_7] = [(2,4), (3,2), (1,1), (4,3), (1,6), (5,3), (4,2)]$$

- distance matrix

	s_1	s_2	s_3	s_4	s_5	s_6	s_7
s_1		2.236	3.162	2.236	2.236	3.162	2.828
s_2	2.236		2.236	1.414	4.472	2.236	1.000
s_3	3.162	2.236		3.605	5.000	4.472	3.162
s_4	2.236	1.414	3.605		4.242	1.000	1.000
s_5	2.236	4.472	5.000	4.242		5.000	5.000
s_6	3.162	2.236	4.472	1.000	5.000		1.414
s_7	2.828	1.000	3.162	1.000	5.000	1.414	

Example

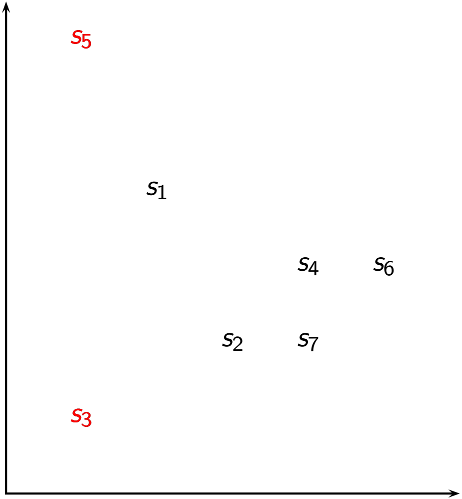
- neighborhood: $d \leq \theta = 3$

	s_1	s_2	s_3	s_4	s_5	s_6	s_7
s_1		2.236	3.162	2.236	2.236	3.162	2.828
s_2	2.236		2.236	1.414	4.472	2.236	1.000
s_3	3.162	2.236		3.605	5.000	4.472	3.162
s_4	2.236	1.414	3.605		4.242	1.000	1.000
s_5	2.236	4.472	5.000	4.242		5.000	5.000
s_6	3.162	2.236	4.472	1.000	5.000		1.414
s_7	2.828	1.000	3.162	1.000	5.000	1.414	

- number of points in the neighborhood

sample	s_1	s_2	s_3	s_4	s_5	s_6	s_7
	4	5	1	4	1	3	4

Example



Outlier Detection

5. deviation-based methods
   - measure the dissimilarity of a data set (e.g. variance)
   - determine the smallest subset of data that if removed results in the largest reduction of dissimilarity
   - combinatorics of subset selection → extremely expensive

Data Preprocessing

- Data Types
- Metrics
- High-Dimensional Data Spaces
- Missing data
- Outlier Detection
- Time-Dependent Data
- Dimensionality Reduction
- Number of Values Reduction
- Sample Size Reduction

Time-Dependent Data

- time series: $[v_1, v_2, v_3, ..., v_n]$
- typical task: predict the next value v_{n+1}
- predictions are based on a window of k preceding values $v_{n+1} = f([v_{n-k+1}, v_{n-k+2}, ..., v_n])$
- data can be rearranged (here $k = 5$):

sample	window					next value
1	v_1	v_2	v_3	v_4	v_5	v_6
2	v_2	v_3	v_4	v_5	v_6	v_7
3	v_3	v_4	v_5	v_6	v_7	v_8
4	v_4	v_5	v_6	v_7	v_8	v_9

Time-Dependent Data

- postponed prediction: predict a value further in the future
- offset o (here $k = 5$, $o = 3$):

sample	window					next value
1	v_1	v_2	v_3	v_4	v_5	v_8
2	v_2	v_3	v_4	v_5	v_6	v_9
3	v_3	v_4	v_5	v_6	v_7	v_{10}
4	v_4	v_5	v_6	v_7	v_8	v_{11}

Time-Dependent Data

- smoothing time-series data: moving average

$$MA(i,k) = \frac{1}{k} \sum_{j=i-k+1}^{i} v_j$$

- different weighting schemas
  - exponential moving average: putting more emphasis on the most recent values

$$EMA(i,k) = \begin{cases} v_i & \text{if } k = 1 \\ p \cdot v_i + (1-p) \cdot EMA(i-1, k-1) & \text{else} \end{cases}$$

  - p controls the influence of more recent data ($p = 0...1$)

Time-Dependent Data

- comparative features: emphasize changes in the data, e.g.

$$v_i - MA(i,k)$$

$$MA(i,k) - MA(i-l,k)$$

$$\frac{v_i}{MA(i,k)}$$

Time-Dependent Data

- survival data: How long does in take until an event occurs?
- applications
  - medicine: outbreak of a desease, death
  - industry: component failure
- problems
  - censored observation:
    the event does not always occur until the end of a (limited) observation period
  - input values are time-dependent:
    the object under observation changes

Data Preprocessing

- Data Types
- Metrics
- High-Dimensional Data Spaces
- Missing Data
- Outlier Detection
- Time-Dependent Data
- Dimensionality Reduction
- Number of Values Reduction
- Sample Size Reduction

Dimensionality Reduction

- Which data can be discarded without sacrificing the quality of the data mining results?
- too many dimensions
    - mining results degrade (insufficient data)
    - resulting model is incomprehensible
    - problem becomes untractable
- too few dimensions
    - data dependencies are lost
    - mining results degrade (limited expressiveness)

Dimensionality reduction

- feature selection approaches
  - feature ranking
  - minimum subset selection
- feature composition approaches
  - principal component analysis

Feature Ranking

- needed: evaluation measure based on
    - accuracy of data
    - consistency
    - information content
    - distances between samples
    - statistical dependencies between features

Feature Ranking

- idea: discard features which contribute least to the overall entropy of the sample set $S = \{\vec{x_1}, \dots \vec{x_n}\}$
- entropy of a sample set is the sum over the entropy of all sample pairs

$$E(S) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E(\vec{x_i}, \vec{x_j})$$

- sample pair is a coin-tossing experiment
  similar – dissimilar: $p_{ij} = p(\vec{x_i} \sim \vec{x_j})$

$$E = -\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} p_{ij} \cdot \log p_{ij} + (1 - p_{ij}) \cdot \log(1 - p_{ij})$$

Feature Ranking

- numerical data:
  probability of being similar is estimated by the inverse of the distance D_{ij} between the two samples

$$p_{ij} = e^{-\alpha \cdot D_{ij}/D}$$

D: average distance among samples in the data set
D_{ij}: distance, e.g. normalized Euclidian distance

$$D_{ij} = \sqrt{\sum_{k=1}^{n} \left(\frac{v_{ki} - v_{kj}}{\max_m(v_{km}) - \min_m(v_{km})} \right)^2}$$

$\rightarrow$ avoids the dominance of dimensions
with great variance over low-variance dimensions

Feature Ranking

- nominal data: Hamming distance

$$p_{ij} = \frac{1}{n} \sum_{k=1}^{n} |x_{ik} = x_{jk}|$$

$$|x_{ik} = x_{jk}| = \begin{cases} 1 & \text{if } x_{ik} = x_{jk} \\ 0 & \text{else} \end{cases}$$

Feature Ranking

- algorithm
  - start with the full set of features F
  - S_F is the sample S using the feature set F
  - $\forall f \in F \; . \; F' = F - \{f\}$

    $$f_{min} = \arg\min_f E(S_F) - E(S_{F'})$$

    rank f_{min} lowest
  - continue with $F' = F - \{f_{min}\}$ until $F = \emptyset$
- features with lower ranks can be discarded

Number of Values Reduction

- one dimensional: feature discretization (binning)
  → mapping values to intervals
  - value exchange techniques
  - merging techniques
- multi-dimensional: clustering of feature vectors
  - splitting techniques

Value Exchange

- given k, the number of bins, distribute the values to minimize the average distance to the mean of a bin
- algorithm
    - sort all values for a given feature
    - assign approximately equal values to each bin
    - move a border element v_i to the neighboring bin if that reduces the global average distance
    - continue until no further improvement can be achieved
- gradient descent search: optimum not guaranteed

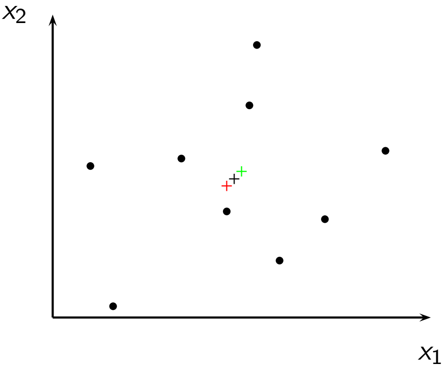
Merging Techniques

- merging of bins: χ^2 merger
  - merge adjacent intervals if they are similar with respect to a classification task
  - independence of the class assignment from the intervals is measured by the χ^2 test
  - adjacent intervals can be merged if their χ^2 value is below a threshold

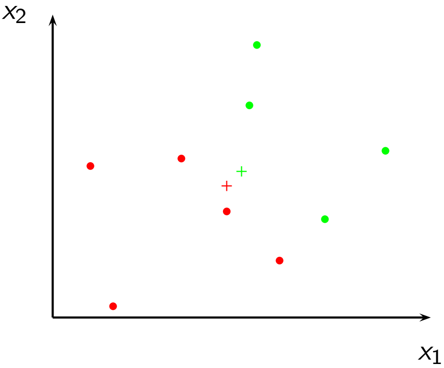
Splitting Techniques

- splitting of bins: vector quantization, k-means clustering
    - compute the mean of all data (centroid)
    - split the centroid(s)
    - assign data points to the nearest centroid (bin)
    - continue until enough bins have been generated

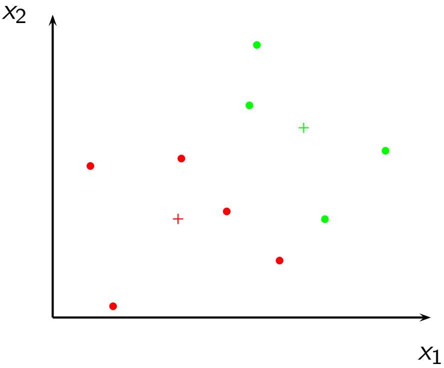
Number of Values Reduction



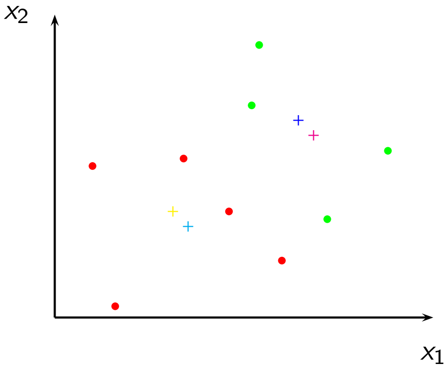
Number of Values Reduction



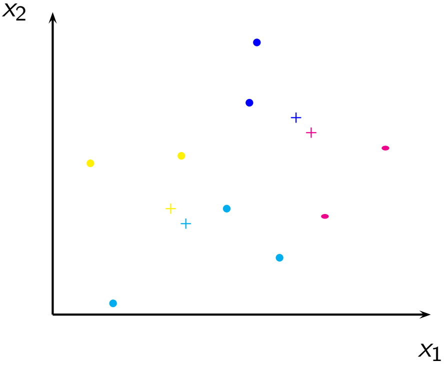
Number of Values Reduction



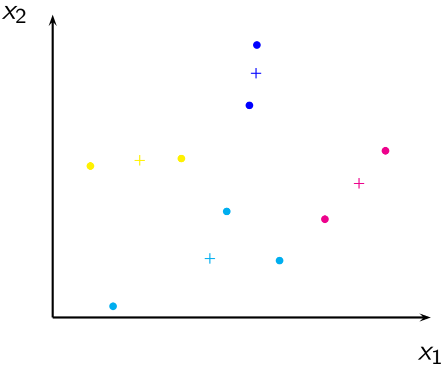
Number of Values Reduction



Number of Values Reduction



Number of Values Reduction



Number of Values Reduction

- global clustering criterion: minimizing the mean square error
  - mean vector as centroid

$$\vec{c_k} = \frac{1}{n_k} \sum_{i=1}^{n_k} \vec{x_{ik}}$$

  - error for one cluster (within-cluster variation)

$$e_k^2 = \sum_{i=1}^{n_k} (\vec{x_{ik}} - \vec{c_k})^2$$

  - global error

$$e = \sum_{k=1}^{K} e_k^2$$

Number of Values Reduction

- "one-dimensional" vector quantisation:
  - mapping continuous values to discrete ones
- multi-dimensional vector quantisation:
  - dimensionality + number of values reduction
- vector quantisation is a special case of unsupervised data clustering:
  - learning without teacher