Data Mining

University of Szeged

Requirements and grading

- Starting with week 3 there will be a short quiz on every practice, 3 points each
- The sum of the 10 best scoring quizzes are taken
- To get a passing mark at least 7 of the quizzes has to score above 0 and the total has to be ≥ 15
- Retake can be taken at the end of the semester from the entire material

Final mark	
[0 – 15)	fail
[15 – 20)	pass
[20 – 24)	fair
[24 – 27)	good
[27 – 30)	excellent

Requirements and grading

- Prerequisite: passing grade from practice
- Oral exam

Summary of the semester

- Introduction, basic concepts
- Data description and preprocessing (data clearing and dimensionality reduction)
- Unsupervised learning
- Supervised learning (classification)
- Outlier detection
- Frequent pattern mining and association rules
- Graph based data mining techniques
- Web-scale data mining and text mining

Recommended literature

- The official notes for the lecture
- H. Witten, E. Frank, M. A. Hall: Data Mining: Practical Machine Learning Tools and Techniques
- P. Tan, M. Steinbach, V. Kumar: Introduction to Data Mining
- B. Liu: Web Data Mining
- **J. Leskovec, A. Rajaraman, J. Ullman: Mining Massive Datasets**
- C. Manning, P. Raghavan, H. Schütze: Introduction to Information Retrieval
- R. Duda, P. Hart, D. Stork: Pattern Recognition
- C. Bishop: Pattern Recognition and Machine Learning
- ...

Motivation

- 1,000,000$ NetFlix Prize
- 20,000$ Kaggle StackOverflow Challange
- Warren Buffet's $10^9$$ offer related to March Madness
- Big Data to fight against terrorism

Main fields of data mining

- Commercial applications
  - Classification of debt inquiries
  - Segmentation of customer groups
  - Churn analysis
- Scientific applications
  - Astronomy
  - Medicine research
  - Medical diagnostics

What is data mining then?

- The recognition of useful, (sometimes) unexpected patterns from a vast amount of data (e.g. from the Web)
- Technological development made it prevalent
  - Both hardware (HDD/RAM/CPU) & software (e.g. MapReduce)
- The knowledge obtained should be easily understandable, valid, useful and novel
- $\approx$ Knowledge Discovery

What is *not* data mining?

- Database queries
- "Simple" statistics (but can be used as a tool)
- Bonferroni Principle: having a massive dataset, uninteresting patterns can emerge just by chance

Total Information Avareness programme

- introducing The Big Brother in USA
- Thought experiment: we are willing to find evil-doers based on hotel reservations log
- 10^9 people each go to any of the 100-bedded hotel with chance 0.01 $\rightarrow$ there are 100,000 (10^9*0.01/100)
- How many suspicious pairs of people (sleeping in the same hotel two times) will be detected over 1,000 days if there are no evil-doers (i.e. all the people are just behaving randomly)?
- P(x and y stay at the very same hotel at some day) = 0.01*0.01*0.00001=10^{-9}
- P(x and y stay at the very same hotel at two days) = $(0.01*0.01*0.00001)^2 = 10^{-18}$
- possible people-night pairings = $\binom{10^9}{2} * \binom{10^3}{2} \approx 2.5 * 10^{23}$
- suspicious pairs all together $\approx 2.5 * 10^{23} * 10^{-18} = 250,000$

Rhine-paradox

- David Rhine's research on parapsychology
- students had to predict the color of 10 cards (blue/red)
- Rhine's results: almost 0.1 % of the subjects were extra-sensory geniuses (2^{-10})
- when 'paraphenomen' were called back they produced average results

Rhine's conclusion?

Paraphenomens loose their special skills once they are told about them.

Simpson's paradox

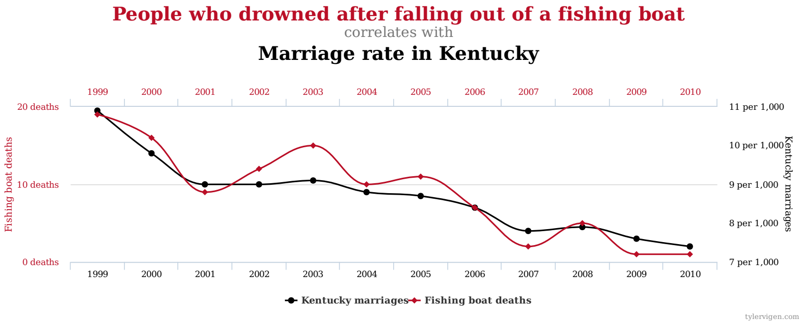
- Think twice before coming to a conclusion!

	Admitted/Applied	
	Female	male
Major A	7/100	3/50
Major B	91/100	172/200
Total	98/200	175/250

- Should we conclude that females are positively discriminated upon admittance?
- Not really, based on the aggregated data (cf. 49% vs. 70%)

Correlation vs. causality

- Number of marriages in Kentucky positively correlate (r>0.95) to the number of people drowned [1]



[1] Source

After all, are they data mining?

- Determining the mean of the ages of the customers of a shop from its database
- Determining the average shoe size from previous data of the army and ruling out outlier data at the same time
- Searching for mutations of organs in tomographic records
- Determining the distribution of electives by sex
- Predicting who would take part in an election

Related areas

- Mathematics: probability theory, statistics, graph theory, algebra, analysis
- Algorithm and computational theory
- Databases (SQL and/or Solr, Elasticsearch, Kibana, . . .)
- Machine learning, pattern recognition, artificial intelligence

Tools, software

- Commercial products (e.g. SAS)
- Machine learning APIs, numeric mathematic libs
  - Weka, MALLET
  - Clementine (SPSS Inc.), Intelligent Miner (IBM), DBMiner (Simon Fraser Univ.)
  - Octave, Matlab, Maple, R
  - Python (numpy, scipy, scikit-learn, pandas)
  - ...

Object of data mining

- (Massive) **data sets** made up of **data objects** that are described by (usually) high-dimensional **feature sets**

Data object	Data attributes
record	field
data point	dimension
sample/measurement	variable
instance/sample	attribute, feature

- Curse of dimensionality: as the number of dimensions grow we need exponentially large number of data points (in order the performance not to drop dramatically)
- Distances often lose their importances in high dimensional spaces $\rightarrow$ dimensionality reduction procedures (to be covered later)

Forms of data sets

- Records
- Lists of transactions (shopping carts)
- Data matrix
- Occurrence (e.g. document-term) matrix

Types of variables based on their scale of measurement

Type of attribute		Description	Examples	Statistics
Category	Nominal	Variables can be checked for equality only	names of cities, hair color	mode, entropy, correlation, χ^2-test
	Ordinal	$>$ relation can be interpreted among variables	grades, {fail, pass, excellent}	median, percentiles
Numerical	Interval	The difference of two variables can be formed and interpreted	shoe sizes, dates, °C	mean, deviation, significance (e.g. F,t-) tests
	Ratio	Ratios can be formed from values of the variables of this kind	age, length, temperature in Kelvin	percent, harmonic mean

Discrete and continuous variables

- Discrete variable: finite or countably infinite number of possible values
- Continuous variable: can take up any real value
- Measurement scales vs. range of variables
  - Variables measured at nominal or ordinal scale are discrete most of the times
  - Variables measured at interval or ratio scale are continuous most of the times
  - Might there be such as continuous binary attribute?
  - What about the measurement scale of discrete count variables?

Further characterization of variables - Symmetry vs. anti-symmetry

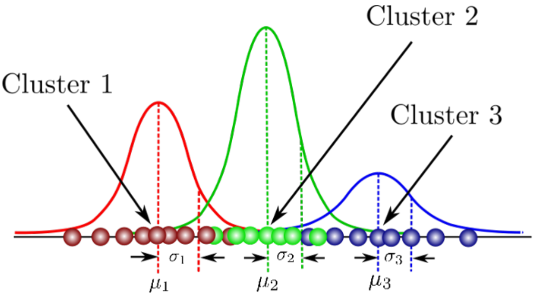
- The absence of an attribute does not necessarily indicate the same amount of similarity of two points compared to the presence of it
- e.g. sparse document vectors

Manipulating features

- Feature discretization
  - Categorical features instead of numerical ones can be more beneficial for certain algorithms both in terms of speed and accuracy
- Feature selection
  - Intuitively, more features should help to obtain better performance, however, it is not always the case
  - We try to select the best subset of features (beware that there are exponentially many subsets! $\rightarrow$ use heuristics)

Unsupervised discretization

- Disregard information about the class label of the data points
  - Form bins of fix-sized intervals
    - We might end up with (near) empty bins
    - Dense regions of feature values might be split
  - Form bins which hold the same amount of observations (we get a "flat" histogram across the bins)
  - Density-based discretization (pl. Gaussian Mixture Model)



Supervised discretization

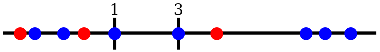
- If observations fall into different classes, we can take into consideration that information as well upon discretization
- Based on mutual information, information gain, χ^2, ... criteria

Mutual information

The mutual information between two random variables X and Y is
$$MI(X;Y) = H(X) - H(X|Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$
It tells us how much our uncertainty decreases about the possible value of X once we become aware of the true value of variable Y.

Example for discretization based on mutual information

Suppose we have the below 10 observations belonging to either of the positive (P) and negative (N) classes.
Does performing discretization for $X = 1$ or $X = 3$ seems to be a better choice?



	P	N
$X \leq 1$	2	3
$X > 1$	1	4

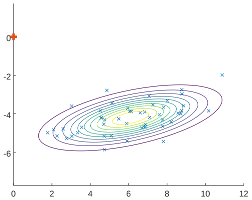
	P	N
$X \leq 3$	2	4
$X > 3$	1	3

$$\frac{2}{10}\log_2\frac{4}{3} + \frac{3}{10}\log_2\frac{6}{7} + \frac{1}{10}\log_2\frac{2}{3} + \frac{4}{10}\log_2\frac{8}{7} \approx 0.035 > \frac{2}{10}\log_2\frac{10}{9} + \ldots \approx 0.006$$

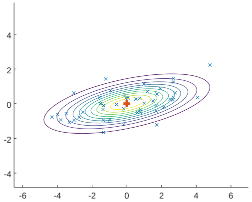
Performing discretization at $X = 1$ should be preferred over discretization at $X = 3$ based on the mutual information

Preprocessing continuous variables – mean-centering

- Subtract the mean observation from every single observation
- Transformed variables show the extent to which the original observations differ from average behavior (its mean will be 0)



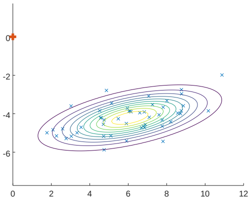
(a) Original data



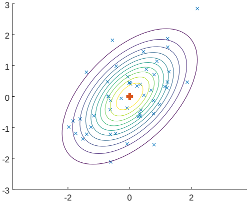
(b) Centralized data

Preprocessing continuous variables – standardizing

- Express the variables in terms of z-scores (from statistics)
  - To what extent does an observation differs from its expected value expressed in terms of its standard deviation



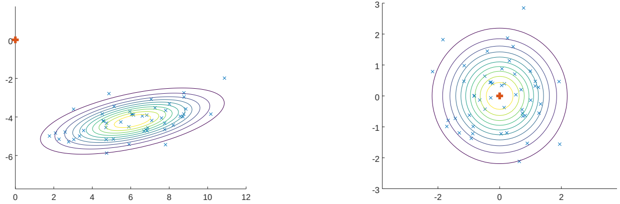
(a) Original data



(c) Standardized data

Preprocessing continuous variables – whitening

- Remove correlation between the variables
  - Convert the (mean-centered) data X with covariance matrix Σ by applying the linear transformation L on it, i.e. XL
    - L such be such that $\Sigma^{-1} = LL^\intercal$ holds

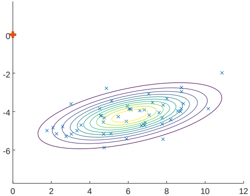


(a) Original data

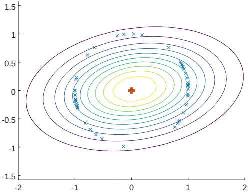
(d) Whitened data

Preprocessing continuous observations – unit normalizing

- Entire observations (rows) can be normalized not just the random variables (columns)
  - For any $x \neq 0$, the transformed vector $x_u = \frac{1}{\sqrt{x^\intercal x}}x$ will point into the direction of x with $\|x_u\|_2 = 1$

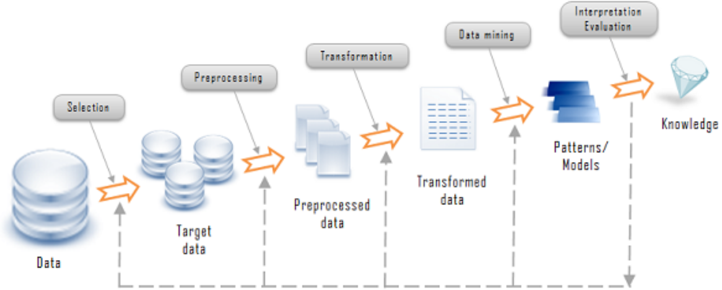


(a) Original data



(e) Unit normalized (and mean centered) data

The process of knowledge discovery



Data quality

- There are errors and inconsistencies in basically every database
  - data collection
  - data digitalization
  - measurement error

Cleaning data

- Missing values, e.g. $x_i = (0, 0, 4, 2, ?, ?, 1, 6, 'True')$
  - approximation (e.g. assign the mode/mean/median attribute value of the k-most-similar data entry)
  - throwing away the data point
  - throwing away the attribute
- Noisy data (pl. age=280)
  - Similarly to missing values
- Filtering out duplicate data entries

Random variables

- Characterize outcomes of experiments
- Results of n experiment/measurement: $x_1, x_2, \dots, x_n$
- Expected value $\mu_X = \mathbb{E}[X] = \sum_{x \in X} P(X = x) * x$
- Variance: expected value of the squared difference from the μ_X
- $Var(X) = \sum_{x \in X} P(X = x) * (x - \mu_X)^2 = \mathbb{E}[X^2] - \mathbb{E}^2[X]$

Example

X=[1,4,7]

$\mathbb{E}[X] = (1 + 4 + 7)/3 = 4$ $\mathbb{E}[X^2] = (1 + 16 + 49)/3 = 22$

$Var(X) = 22 - 4^2 = 6$

Some refreshment on algebra (which might come handy later on)

- Euclidean distance: $\|\mathbf{a}\|_2 = \sqrt{\sum_{i=1}^{d} a_i^2}$
- Inner/scalar product: $\mathbf{a}^\intercal \mathbf{a} = \sum_{i=1}^{d} a_i^2$
- Eigenvalues, eigenvectors
  Right-side eigenvalues: $A\mathbf{x} = \lambda\mathbf{x} \Leftrightarrow \det(A - \lambda I) = 0$
  e.g. $A = \begin{bmatrix} 3 & \sqrt{20} \\ \sqrt{20} & 4 \end{bmatrix}$ -> $\lambda^2 - 7\lambda - 8 = 0$
  Left-side eigenvalues: $\mathbf{y}A = \lambda\mathbf{y}$