

Data Warehousing SoSe 2006

Dr. Jens-Peter Dittrich

jens.dittrich@inf

www.inf.ethz.ch/~jensdi

Institute of Information Systems



Data Mining

Based on Tutorial Slides by

Gregory Piatetsky-Shapiro

Kdnuggets.com



Outline

- Introduction
- Data Mining Tasks
- Classification & Evaluation
- Clustering
- Application Examples


Trends leading to Data Flood


- More data is generated:
 - Web, text, images ...
 - Business transactions, calls, ...
 - Scientific data: astronomy, biology, etc
- More data is captured:
 - Storage technology faster and cheaper
 - DBMS can handle bigger DB




Largest Databases in 2005

Winter Corp. 2005 Commercial
Database Survey:

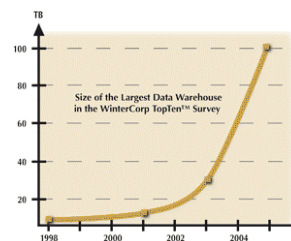
 Max Planck Inst. for
Meteorology , 222 TB

 Yahoo ~ 100 TB (Largest Data
Warehouse)

 AT&T ~ 94 TB

www.wintercorp.com/VLDB/2005_TopTen_Survey/TopTenWinners_2005.asp

Data Growth



In 2 years (2003 to 2005),
the size of the largest database TRIPLED!

Data Growth Rate

- Twice as much information was created in 2002 as in 1999 (~30% growth rate)
- Other growth rate estimates even higher
- Very little data will ever be looked at by a human

Knowledge Discovery is **NEEDED** to make sense and use of data.

© 2006 KDnuggets

7

Knowledge Discovery Definition

Knowledge Discovery in Data is the *non-trivial* process of identifying

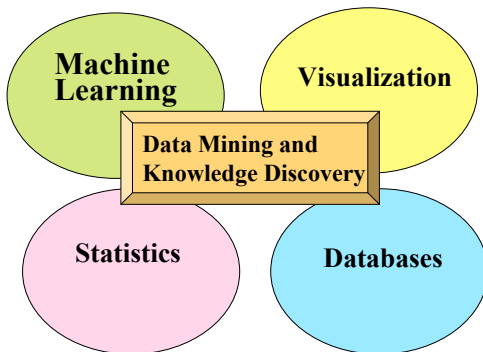
- *valid*
- *novel*
- potentially *useful*
- and ultimately *understandable patterns* in data.

from *Advances in Knowledge Discovery and Data Mining*, Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, (Chapter 1), AAAI/MIT Press 1996

© 2006 KDnuggets

8

Related Fields



© 2006 KDnuggets

9

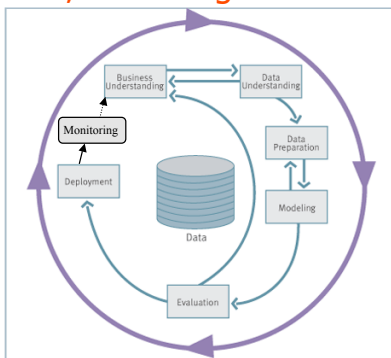
Statistics, Machine Learning and Data Mining

- Statistics:
 - more theory-based
 - more focused on testing hypotheses
- Machine learning
 - more heuristic
 - focused on improving performance of a learning agent
 - also looks at real-time learning and robotics – areas not part of data mining
- Data Mining and Knowledge Discovery
 - integrates theory and heuristics
 - focus on the entire process of knowledge discovery, including data cleaning, learning, and integration and visualization of results
- Distinctions are fuzzy

© 2006 KDnuggets

10

Knowledge Discovery Process flow, according to CRISP-DM



see www.crisp-dm.org for more information

Continuous monitoring and improvement is an addition to CRISP

© 2006 KDnuggets

11

Historical Note: Many Names of Data Mining

- Data Fishing, Data Dredging: 1960-
 - used by statisticians (as bad name)
- Data Mining :1990 --
 - used in DB community, business
- Knowledge Discovery in Databases (1989-)
 - used by AI, Machine Learning Community
- also Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction, ...



Currently: Data Mining and Knowledge Discovery are used interchangeably

© 2006 KDnuggets

12

Data Mining Tasks

Some Definitions

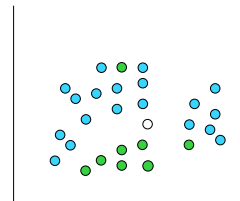
- Instance (also Item or Record):
 - an example, described by a number of attributes,
 - e.g. a day can be described by temperature, humidity and cloud status
- Attribute or Field
 - measuring aspects of the Instance, e.g. temperature
- Class (Label)
 - grouping of instances, e.g. days good for playing

Major Data Mining Tasks

- **Classification:** predicting an item class
- **Clustering:** finding clusters in data
- **Associations:** e.g. A & B & C occur frequently
- **Visualization:** to facilitate human discovery
- **Summarization:** describing a group
- Deviation Detection: finding changes
- Estimation: predicting a continuous value
- Link Analysis: finding relationships

Classification

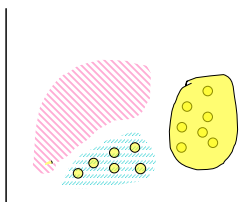
Learn a method for predicting the instance class from pre-labeled (classified) instances



Many approaches:
Statistics,
Decision Trees,
Neural Networks,
...

Clustering

Find "natural" grouping of instances given un-labeled data



Association Rules & Frequent Itemsets

Transactions

TID	Produce
1	MILK, BREAD, EGGS
2	BREAD, SUGAR
3	BREAD, CEREAL
4	MILK, BREAD, SUGAR
5	MILK, CEREAL
6	BREAD, CEREAL
7	MILK, CEREAL
8	MILK, BREAD, CEREAL, EGGS
9	MILK, BREAD, CEREAL



Frequent Itemsets:

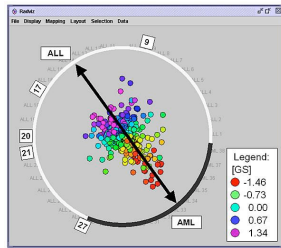
Milk, Bread (4)
Bread, Cereal (3)
Milk, Bread, Cereal (2)
...



Rules:
Milk => Bread (66%)

Visualization & Data Mining

- Visualizing the data to facilitate human discovery
- Presenting the discovered results in a visually "nice" way

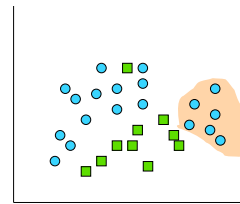


© 2006 KDnuggets

19

Summarization

- Describe features of the selected group
- Use natural language and graphics
- Usually in Combination with Deviation detection or other methods



Average length of stay in this study area rose 45.7 percent, from 4.3 days to 6.2 days, because ...

© 2006 KDnuggets

20

Data Mining Central Quest

Find true patterns and avoid *overfitting*

(finding seemingly significant but really random patterns due to searching too many possibilities)

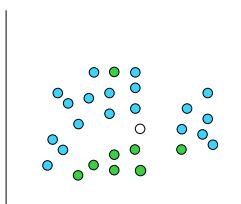
© 2006 KDnuggets

21

Classification Methods

Classification

Learn a method for predicting the instance class from pre-labeled (classified) instances



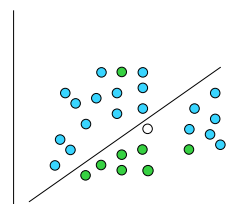
Many approaches:
Regression,
Decision Trees,
Bayesian,
Neural Networks,
...

Given a set of points from classes \bullet \circ
what is the class of new point \circ ?

© 2006 KDnuggets

23

Classification: Linear Regression



- Linear Regression
 $w_0 + w_1 x + w_2 y \geq 0$
- Regression computes w_i from data to minimize squared error to 'fit' the data
- Not flexible enough

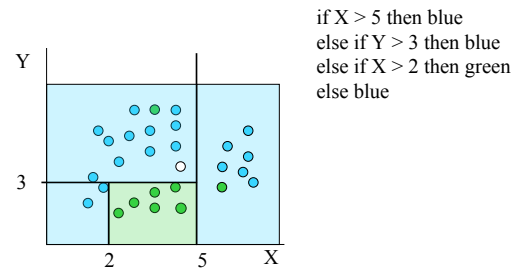
© 2006 KDnuggets

24

Regression for Classification

- Any regression technique can be used for classification
 - Training: perform a regression for each class, setting the output to 1 for training instances that belong to class, and 0 for those that don't
 - Prediction: predict class corresponding to model with largest output value (*membership value*)
- For linear regression this is known as *multi-response linear regression*

Classification: Decision Trees



DECISION TREE

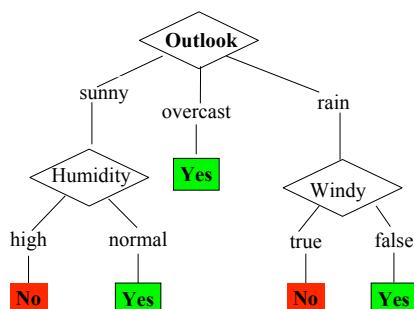
- An internal node is a test on an attribute.
- A branch represents an outcome of the test, e.g., Color=red.
- A leaf node represents a class label or class label distribution.
- At each node, one attribute is chosen to split training examples into distinct classes as much as possible
- A new instance is classified by following a matching path to a leaf node.

Weather Data: Play or not Play?

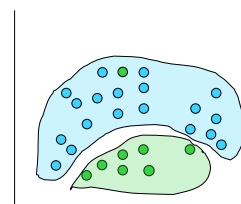
Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

*Note:
 Outlook is the
 Forecast,
 no relation to
 Microsoft
 email program*

Example Tree for "Play?"



Classification: Neural Nets



- Can select more complex regions
- Can be more accurate
- Also can overfit the data – find patterns in random noise

Classification: other approaches

- Naïve Bayes
- Rules
- Support Vector Machines
- Genetic Algorithms
- ...

See www.KDnuggets.com/software/

Evaluation

Evaluating which method works the best for classification

- No model is uniformly the best
- Dimensions for Comparison
 - speed of training
 - speed of model application
 - noise tolerance
 - explanation ability
- Best Results: Hybrid, Integrated models

Comparison of Major Classification Approaches

	Train time	Run Time	Noise Tolerance	Can Use Prior Knowledge	Accuracy on Customer Modelling	Under-standable
Decision Trees	fast	fast	poor	no	medium	medium
Rules	med	fast	poor	no	medium	good
Neural Networks	slow	fast	good	no	good	poor
Bayesian	slow	fast	good	yes	good	good

A hybrid method will have higher accuracy

Evaluation of Classification Models

- How predictive is the model we learned?
- Error on the training data is *not* a good indicator of performance on future data
 - The new data will probably not be **exactly** the same as the training data!
- Overfitting – fitting the training data too precisely - usually leads to poor results on new data

Evaluation issues

- Possible evaluation measures:
 - Classification Accuracy
 - Total cost/benefit – when different errors involve different costs
 - Lift and ROC (Receiver operating characteristic) curves
 - Error in numeric predictions
- How reliable are the predicted results ?

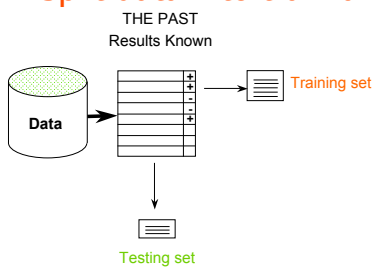
Classifier error rate

- Natural performance measure for classification problems: *error rate*
 - *Success*: instance's class is predicted correctly
 - *Error*: instance's class is predicted incorrectly
 - Error rate: proportion of errors made over the whole set of instances
- *Training set error rate*: is way too optimistic!
 - you can find patterns even in random data

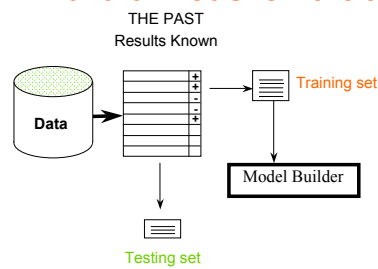
Evaluation on "LARGE" data

- If many (>1000) examples are available, including >100 examples from each class
- A simple evaluation will give useful results
 - Randomly split data into training and test sets (usually 2/3 for train, 1/3 for test)
 - Build a classifier using the *train* set and evaluate it using the *test* set

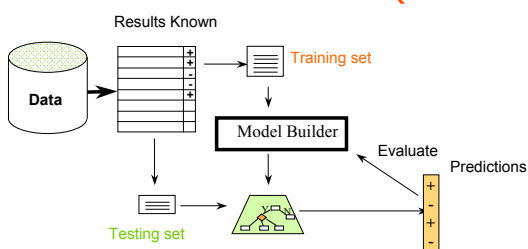
Classification Step 1: Split data into train and test sets



Classification Step 2: Build a model on a training set



Classification Step 3: Evaluate on test set (Re-train?)



Unbalanced data

- Sometimes, classes have very unequal frequency
 - Attrition prediction: 97% stay, 3% attrite (in a month)
 - medical diagnosis: 90% healthy, 10% disease
 - eCommerce: 99% don't buy, 1% buy
 - Security: >99.99% of Americans are not terrorists
- Similar situation with multiple classes
- Majority class classifier can be 97% correct, but useless

Handling unbalanced data – how?

If we have two classes that are very unbalanced, then how can we evaluate our classifier method?

Balancing unbalanced data, 1

- With two classes, a good approach is to build **BALANCED** train and test sets, and train model on a balanced set
 - randomly select desired number of minority class instances
 - add equal number of randomly selected majority class
- How do we generalize “balancing” to multiple classes?

Balancing unbalanced data, 2

- Generalize “balancing” to multiple classes
 - Ensure that each class is represented with approximately equal proportions in train and test

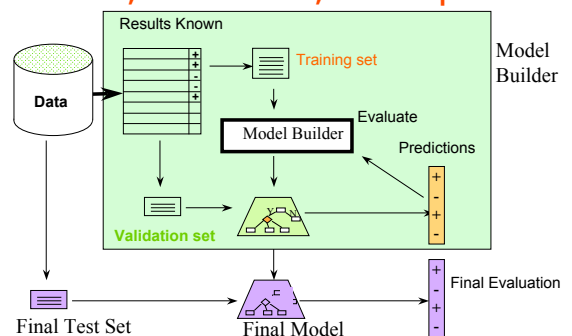
A note on parameter tuning

- It is important that the test data is not used *in any way* to create the classifier
- Some learning schemes operate in two stages:
 - Stage 1: builds the basic structure
 - Stage 2: optimizes parameter settings
- The test data can't be used for parameter tuning!
- Proper procedure uses three sets: **training data, validation data, and test data**
 - Validation data is used to optimize parameters

Making the most of the data

- Once evaluation is complete, *all the data* can be used to build the final classifier
- Generally, the larger the training data the better the classifier
- The larger the test data the more accurate the error estimate

Classification: Train, Validation, Test split



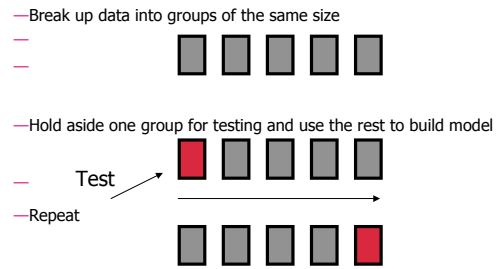
Cross-validation

- *Cross-validation* avoids overlapping test sets
 - First step: data is split into k subsets of equal size
 - Second step: each subset in turn is used for testing and the remainder for training
- This is called *k-fold cross-validation*
- Often the subsets are stratified before the cross-validation is performed
- The error estimates are averaged to yield an overall error estimate

© 2006 KDnuggets

49

Cross-validation example:



© 2006 KDnuggets

50

More on cross-validation

- Standard method for evaluation: stratified ten-fold cross-validation
- Why ten? Extensive experiments have shown that this is the best choice to get an accurate estimate
- Stratification reduces the estimate's variance
- Even better: repeated stratified cross-validation
 - E.g. ten-fold cross-validation is repeated ten times and results are averaged (reduces the variance)

© 2006 KDnuggets

51

Direct Marketing Paradigm

- Find most likely prospects to contact
- Not everybody needs to be contacted
- Number of targets is usually much smaller than number of prospects
- Typical Applications
 - retailers, catalogues, direct mail (and e-mail)
 - customer acquisition, cross-sell, attrition prediction
 - ...

© 2006 KDnuggets

52

Direct Marketing Evaluation

- **Accuracy on the entire dataset is not the right measure**
- Approach
 - develop a target model
 - score all prospects and rank them by decreasing score
 - select top P% of prospects for action
- How do we decide what is the best subset of prospects ?

© 2006 KDnuggets

53

Model-Sorted List

Use a model to assign score to each customer
Sort customers by decreasing score
Expect more targets (hits) near the top of the list

No	Score	Target	CustID	Age
1	0.97	Y	1746	...
2	0.95	N	1024	...
3	0.94	Y	2478	...
4	0.93	Y	3820	...
5	0.92	N	4897	...
...
99	0.11	N	2734	...
100	0.06	N	2422	...

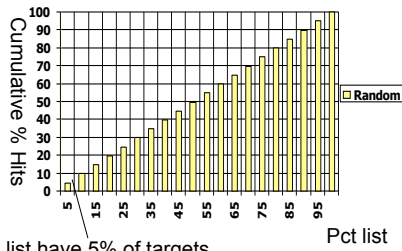
3 hits in top 5% of the list
If there 15 targets overall, then top 5 has $3/15=20\%$ of targets

© 2006 KDnuggets

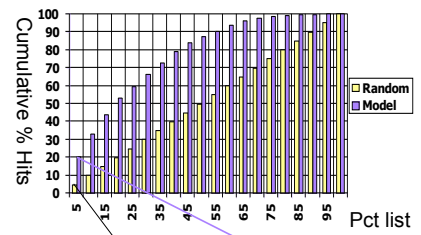
54

CPH (Cumulative Pct Hits)

Definition:
CPH(P,M)
 = % of all targets
 in the first P%
 of the list scored
 by model M
 CPH frequently
 called Gains



CPH: Random List vs Model-ranked list

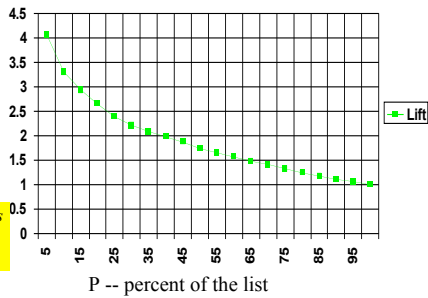


5% of random list have 5% of targets,
 but 5% of model ranked list have 21% of targets
 $CPH(5\%,model)=21\%$.

Lift

$$Lift(P,M) = CPH(P,M) / P$$

Lift (at 5%)
 = $21\% / 5\%$
 = 4.2
 better
 than random



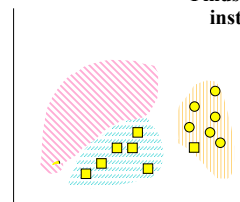
Note: Some authors
 use "Lift" for what
 we call CPH.

Lift – a measure of model quality

- Lift helps us decide which models are better
- If cost/benefit values are not available or changing, we can use Lift to select a better model.
- Model with the higher Lift curve will generally be better

Clustering

Unsupervised learning:
 Finds "natural" grouping of
 instances given un-labeled data



Clustering Methods

- Many different method and algorithms:
 - For numeric and/or symbolic data
 - Deterministic vs. probabilistic
 - Exclusive vs. overlapping
 - Hierarchical vs. flat
 - Top-down vs. bottom-up

Clustering Evaluation

- Manual inspection
- Benchmarking on existing labels
- Cluster quality measures
 - distance measures
 - high similarity within a cluster, low across clusters

The distance function

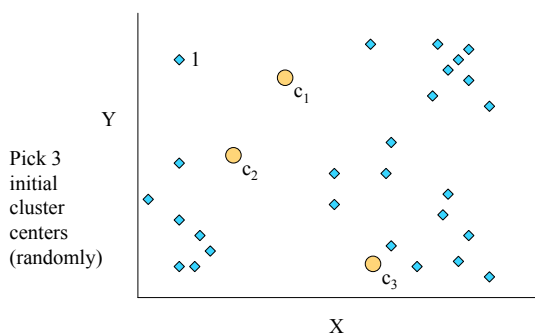
- Simplest case: one numeric attribute A
 - $\text{Distance}(X,Y) = A(X) - A(Y)$
- Several numeric attributes:
 - $\text{Distance}(X,Y) = \text{Euclidean distance between } X,Y$
- Nominal attributes: distance is set to 1 if values are different, 0 if they are equal
- Are all attributes equally important?
 - Weighting the attributes might be necessary

Simple Clustering: K-means

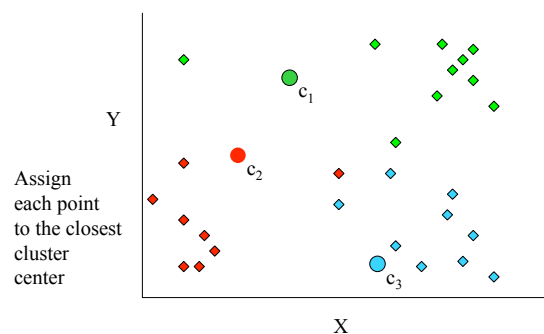
Works with numeric data only

- 1) Pick a number (K) of cluster centers (at random)
- 2) Assign every item to its nearest cluster center (e.g. using Euclidean distance)
- 3) Move each cluster center to the mean of its assigned items
- 4) Repeat steps 2,3 until convergence (change in cluster assignments less than a threshold)

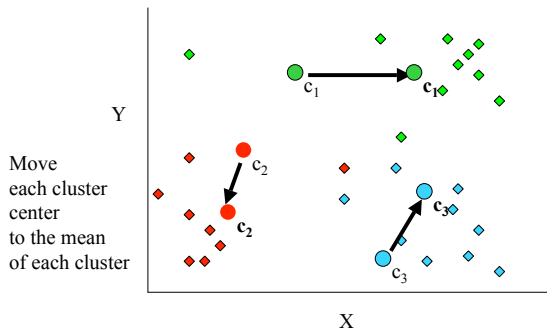
K-means example, step 1



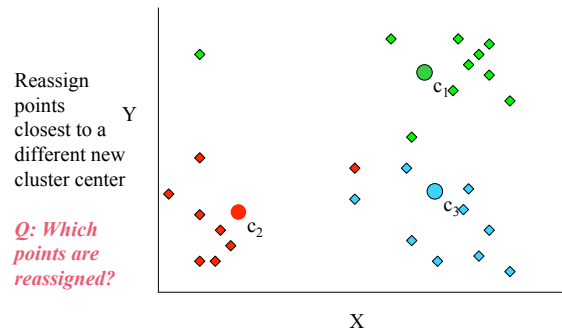
K-means example, step 2



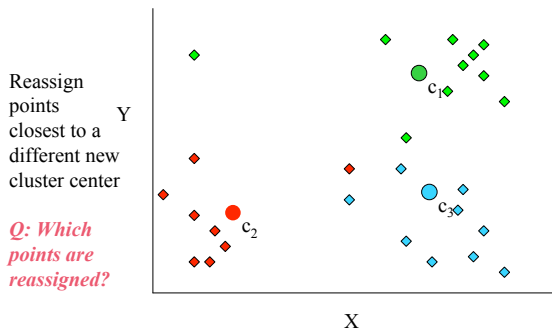
K-means example, step 3



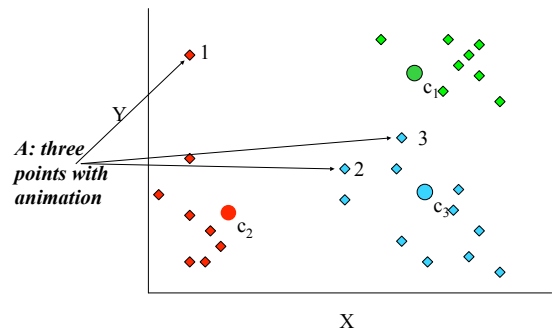
K-means example, step 4a



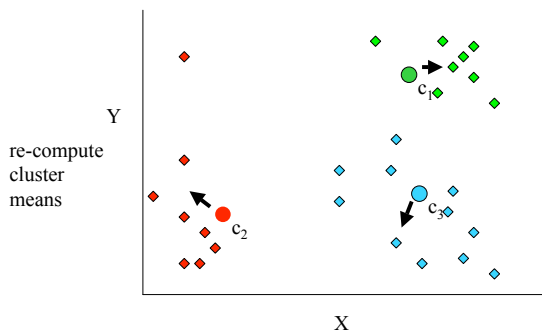
K-means example, step 4b



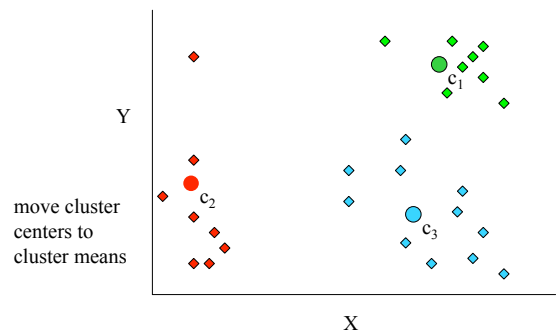
K-means example, step 4c



K-means example, step 4d



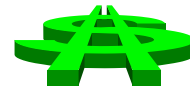
K-means example, step 5



Data Mining Applications

Problems Suitable for Data-Mining

- require knowledge-based decisions
- have a changing environment
- have sub-optimal current methods
- have accessible, sufficient, and relevant data
- provides high payoff for the right decisions!



Major Application Areas for Data Mining Solutions

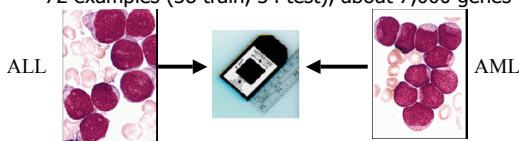
- Advertising
- Bioinformatics
- Customer Relationship Management (CRM)
- Database Marketing
- Fraud Detection
- eCommerce
- Health Care
- Investment/Securities
- Manufacturing, Process Control
- Sports and Entertainment
- Telecommunications
- Web

Application: Search Engines

- Before Google, web search engines used mainly keywords on a page – results were easily subject to manipulation
- Google's early success was partly due to its algorithm which uses mainly links to the page
- Google founders Sergey Brin and Larry Page were students at Stanford in 1990s
- Their research in databases and data mining led to Google

Microarrays: Classifying Leukemia

- Leukemia: Acute Lymphoblastic (ALL) vs Acute Myeloid (AML), Golub et al, Science, v.286, 1999
- 72 examples (38 train, 34 test), about 7,000 genes



Visually similar, but genetically very different

Best Model: 97% accuracy,
1 error (sample suspected mislabelled)

Microarray Potential Applications

- New and better molecular diagnostics
 - Jan 11, 2005: FDA approved Roche Diagnostic AmpliChip, based on Affymetrix technology
- New molecular targets for therapy
 - few new drugs, large pipeline, ...
- Improved treatment outcome
 - Partially depends on genetic signature
- Fundamental Biological Discovery
 - finding and refining biological pathways
- Personalized medicine ?!

Application: Direct Marketing and CRM

- Most major direct marketing companies are using modeling and data mining
- Most financial companies are using customer modeling
- Modeling is easier than changing customer behaviour
- Example
 - Verizon Wireless reduced customer attrition rate from 2% to 1.5%, saving many millions of \$

© 2006 KDnuggets

79

Application: e-Commerce

- Amazon.com recommendations
 - if you bought (viewed) X, you are likely to buy Y
- Netflix
 - If you liked "Monty Python and the Holy Grail", you get a recommendation for "This is Spinal Tap"
- Comparison shopping
 - Froogle, mySimon, Yahoo Shopping, ...



© 2006 KDnuggets

80

Application: Security and Fraud Detection

- Credit Card Fraud Detection
 - over 20 Million credit cards protected by Neural networks (Fair, Isaac)
- Securities Fraud Detection
 - NASDAQ KDD system
- Phone fraud detection
 - AT&T, Bell Atlantic, British Telecom/MCI



© 2006 KDnuggets

81

Data Mining, Privacy, and Security

- TIA: Terrorism (formerly Total) Information Awareness Program –
 - TIA program closed by Congress in 2003 because of privacy concerns
- However, in 2006 we learn that NSA is analyzing US domestic call info to find potential terrorists
 - Invasion of Privacy or Needed Intelligence?

© 2006 KDnuggets

82

Criticism of Analytic Approaches to Threat Detection:

Data Mining will

- be ineffective - generate millions of false positives
- and invade privacy

First, can data mining be effective?

© 2006 KDnuggets

83

Can Data Mining and Statistics be Effective for Threat Detection?

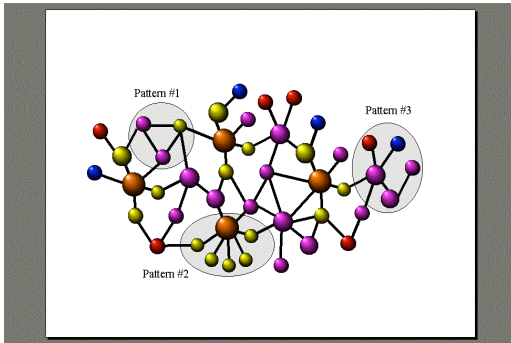
- Criticism: Databases have 5% errors, so analyzing 100 million suspects will generate 5 million false positives
- Reality: Analytical models correlate many items of information to reduce false positives.
- Example: Identify one biased coin from 1,000.
 - After one throw of each coin, we cannot
 - After 30 throws, one biased coin will stand out with high probability.
 - Can identify 19 biased coins out of 100 million with sufficient number of throws



© 2006 KDnuggets

84

Another Approach: Link Analysis



Can find unusual patterns in the network structure

© 2006 KDnuggets

85

Analytic technology can be effective

- Data Mining is just one additional tool to help analysts
- Combining multiple models and link analysis can reduce false positives
- Today there are millions of false positives with manual analysis
- Analytic technology has the potential to reduce the current high rate of false positives

© 2006 KDnuggets

86

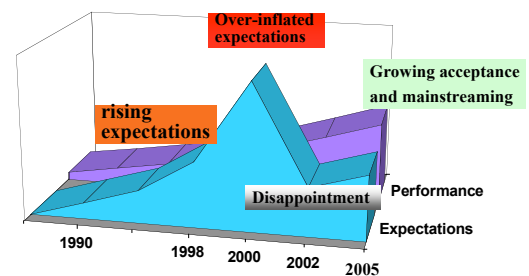
Data Mining with Privacy

- Data Mining looks for patterns, not people!
- Technical solutions can limit privacy invasion
 - Replacing sensitive personal data with anon. ID
 - Give randomized outputs
 - Multi-party computation – distributed data
 - ...
- Bayardo & Srikant, Technological Solutions for Protecting Privacy, IEEE Computer, Sep 2003

© 2006 KDnuggets

87

The Hype Curve for Data Mining and Knowledge Discovery



© 2006 KDnuggets

88

Summary

- Data Mining and Knowledge Discovery are needed to deal with the flood of data
- Knowledge Discovery is a process !
- Avoid overfitting (finding random patterns by searching too many possibilities)

© 2006 KDnuggets

89

Additional Resources

www.KDnuggets.com

data mining software, jobs, courses, etc

www.acm.org/sigkdd

ACM SIGKDD – the professional society for data mining

© 2006 KDnuggets

90