Applications

- Banking: loan/credit card approval
  - predict good customers based on old customers
- Customer relationship management:
  - identify those who are likely to leave for a competitor.
- Targeted marketing:
  - identify likely responders to promotions
- Fraud detection: telecommunications, financial transactions
  - from an online stream of event identify fraudulent events
- Manufacturing and production:
  - automatically adjust knobs when process parameter changes

Applications (continued)

- Medicine: disease outcome, effectiveness of treatments
  - analyze patient disease history: find relationship between diseases
- Molecular/Pharmaceutical: identify new drugs
- Scientific data analysis:
  - identify new galaxies by searching for sub clusters
- Web site/store design and promotion:
  - find affinity of visitor to pages and modify layout

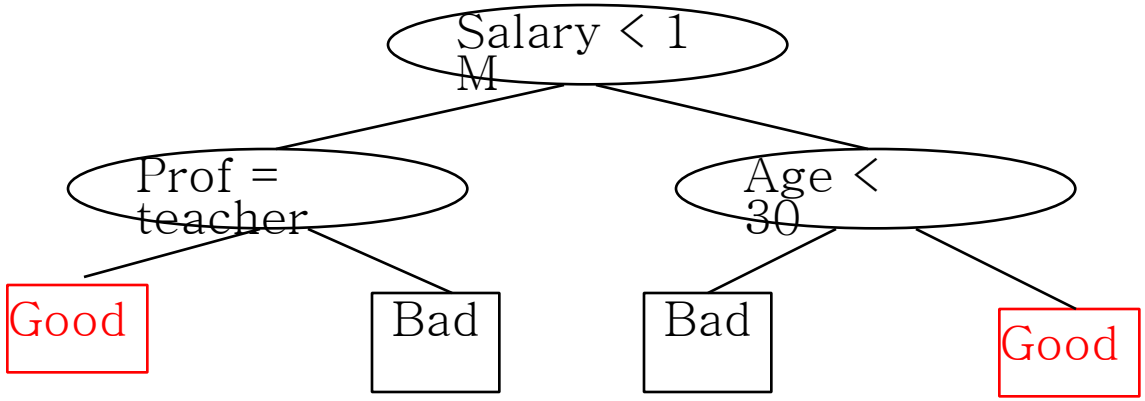
Some basic Techineques

- Predictive:
  - Decision Tree
  - Neural network
  - Memory-based Reasoning
- Descriptive:
  - Clustering / similarity matching
  - Association rules and variants / link analysis
  - Genetic Algorithms

Predictive Learning

1.Decision trees

- Tree where internal nodes are simple decision rules on one or more attributes and leaf nodes are predicted class labels.



Decision tree classifiers

- Widely used learning method
- Easy to interpret: can be re-represented as if-then-else rules
- Approximates function by piece wise constant regions
- Does not require any prior knowledge of data distribution, works well on noisy data.
- Has been applied to:
  - classify medical patients based on the disease,
  - equipment malfunction by cause,
  - loan applicant by likelihood of payment.

Pros and Cons of decision trees

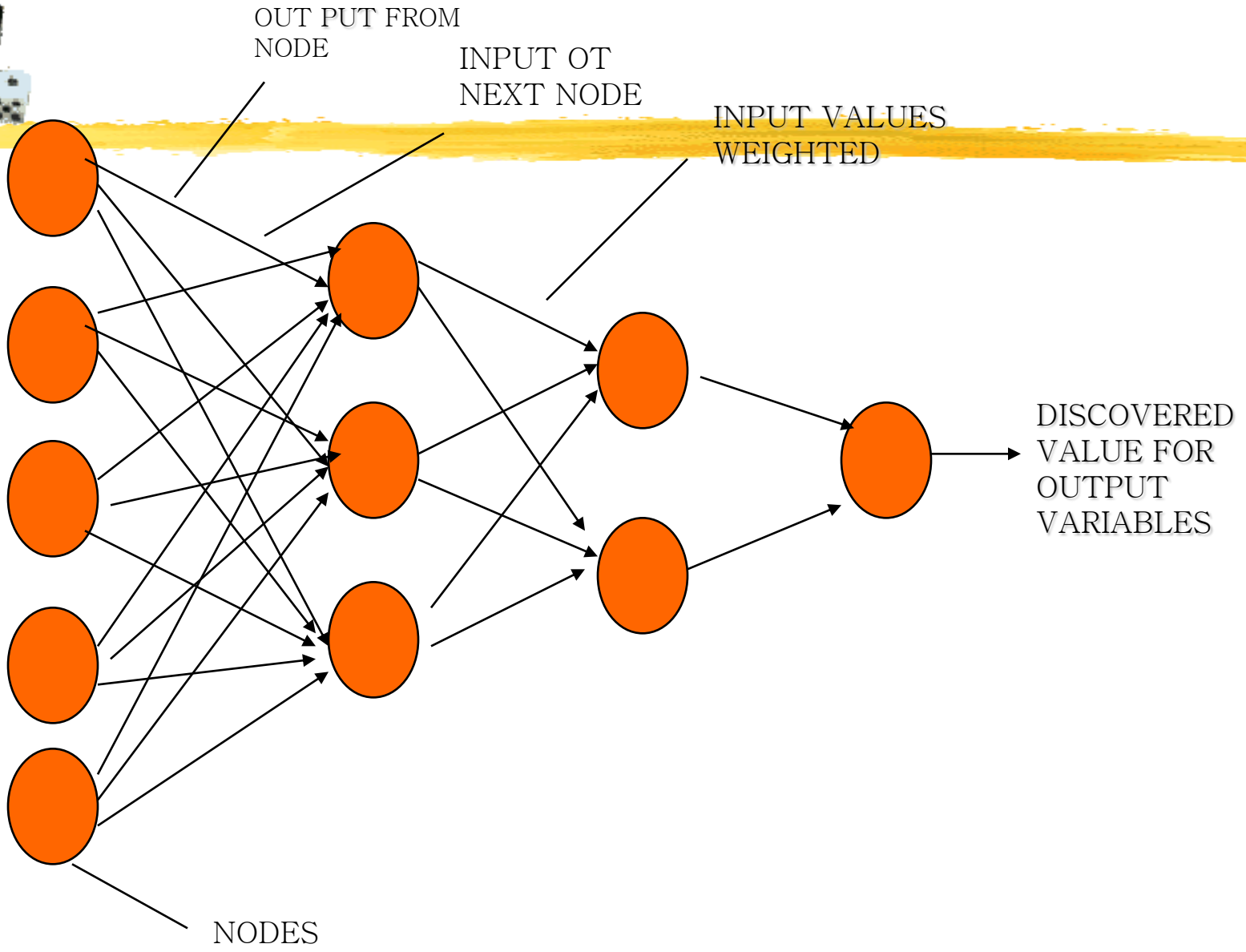
- **Pros**
  - \+ Reasonable training time
  - \+ Fast application
  - \+ Easy to interpret
  - \+ Easy to implement
  - \+ Can handle large number of features
- **Cons**
  - Cannot handle complicated relationship between features
  - simple decision boundaries
  - problems with lots of missing data

More information:
http://www.stat.wisc.edu/~limt/treeprogs.html

2.NEURAL NETWORKS

- SIMILAR TO HUMAN BRAIN
- TRAINED DATASET AND PATTERNS FOR CLASSIFICATION AND PREDICTION
- ALGORITHMS ARE EFFECTIVE WHEN DATA IS SHAPELESS AND LACKS PATTERN

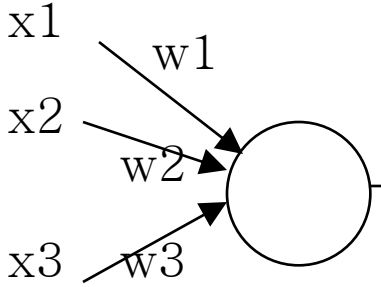
VALUES FOR INPUT VARIABLES



Neural network

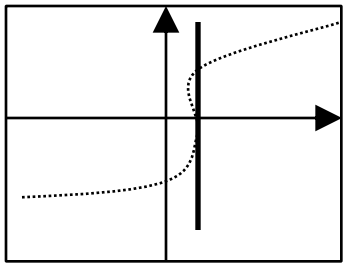
- Set of nodes connected by directed weighted edges

Basic NN unit

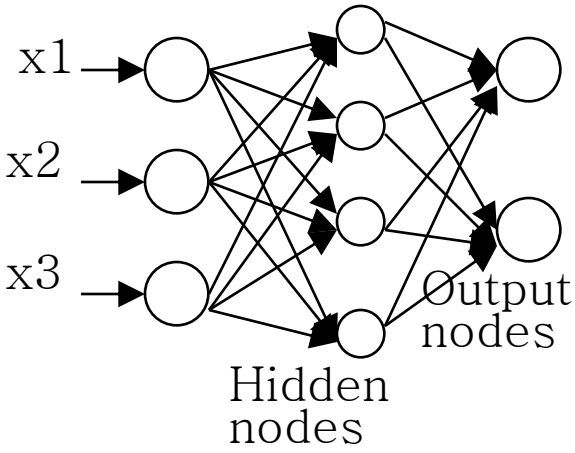


$$o = \sigma\left(\sum_{i=1}^{n} w_i x_i\right)$$

$$\sigma(y) = \frac{1}{1+e^{-y}}$$



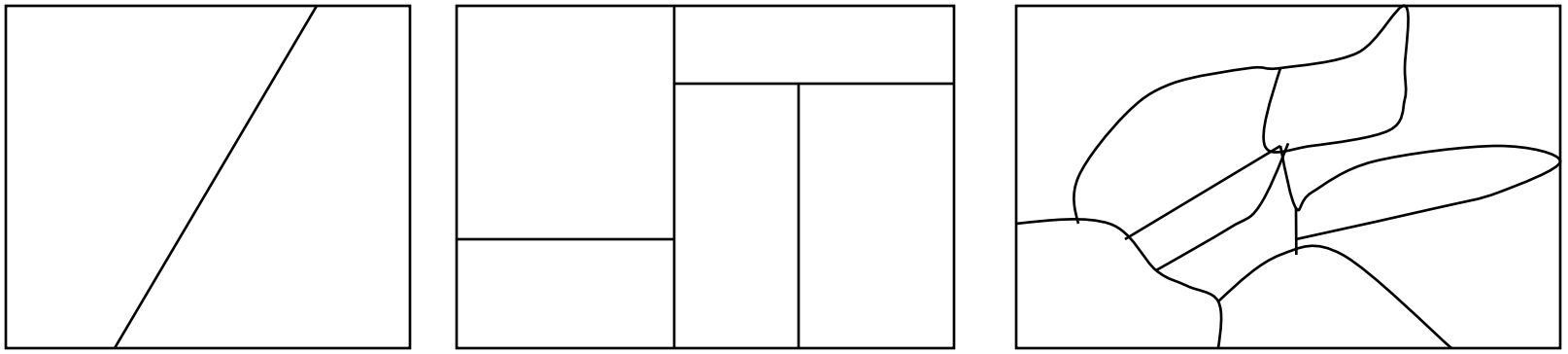
A more typical NN



Neural networks

- Useful for learning complex data like handwriting, speech and image recognition

Decision boundaries:



Linear regression Classification tree Neural network

Pros and Cons of Neural Network

· **Pros**

+ Can learn more complicated class boundaries
+ Fast application
+ Can handle large number of features

· **Cons**

- Slow training time
- Hard to interpret
- Hard to implement: trial and error for choosing number of nodes

Conclusion: Use neural nets only if decision-trees/NN fail.

3.Memory-Based Reasoning (MBR)

- Based on the concept of similarity
  - Memory-Based Reasoning (MBR) – results are based on analogous situations in the past
- Our ability to reason from experience depends on our ability to recognize appropriate examples from the past…
  - Traffic patterns/routes
  - Movies
  - Food
- We identify similar example(s) and apply what we know/learned to current situation
- These similar examples in MBR are referred to as *neighbors*

MBR Applications

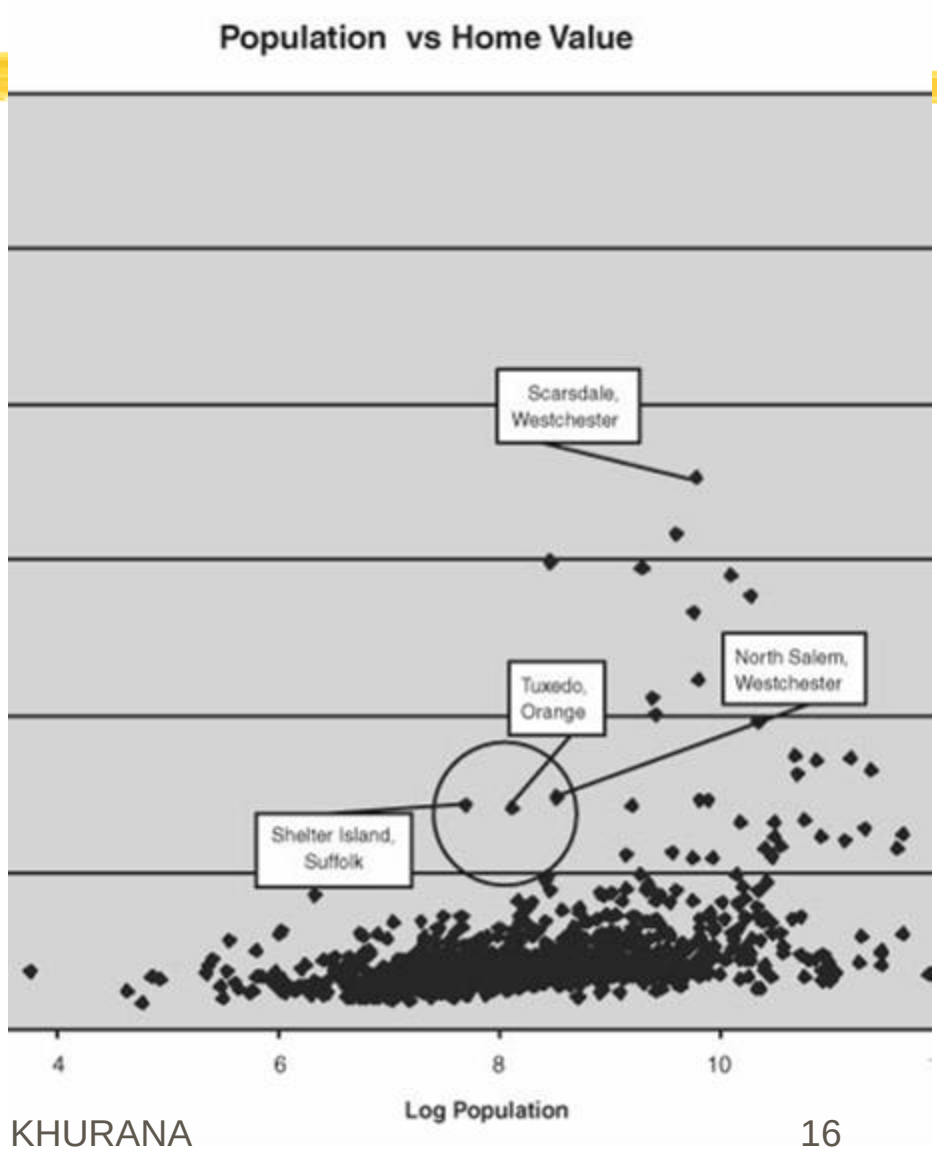
- Fraud detection
- Customer response prediction
- Medical treatments
- Classifying responses – MBR can process free-text responses and assign codes

MBR Strengths

- Ability to use data "as is" – utilizes both a ***distance function*** and a ***combination function*** between data records to help determine how "neighborly" they are
- Ability to adapt – adding new data makes it possible for MBR to learn new things
- Good results without lengthy training

MBR Example – Rents in Tuxedo, NY

- Classify nearest neighbors based on descriptive variables – population & median home prices (not geography in this example)
- Range midpoint in 2 neighbors is $1,000 & $1,250 so Tuxedo rent should be $1,125; 2nd method yields rent of $977
- Actual midpoint rent in Tuxedo turns out to be $1,250 (one method) and $907 in another.



MBR Challenges

1. Choosing appropriate historical data for use in training
2. Choosing the most efficient way to represent the training data
3. Choosing the distance function, combination function, and the number of neighbors

Memory-Based Reasoning Exercise

- Work in teams of 3 or 4
- Time Limit = 10 minutes
- Discuss a couple of ways in which MBR could be utilized and hence useful to an organization (enterprise, govt agency, etc.)
- Teams present ideas

Descriptive:

- Clustering / similarity matching
- Association rules and variants / link analysis
- Genetic Algorithms

1. What's Clustering

- Clustering is a kind of unsupervised learning.
- Clustering is a method of grouping data that share similar trend and patterns.
- Clustering of data is a method by which large sets of data is grouped into clusters of smaller sets of similar data.
  - Example:

After clustering:

Thus, we see clustering means grouping of data or dividing a large data set into smaller data sets of some similarity.

The usage of clustering

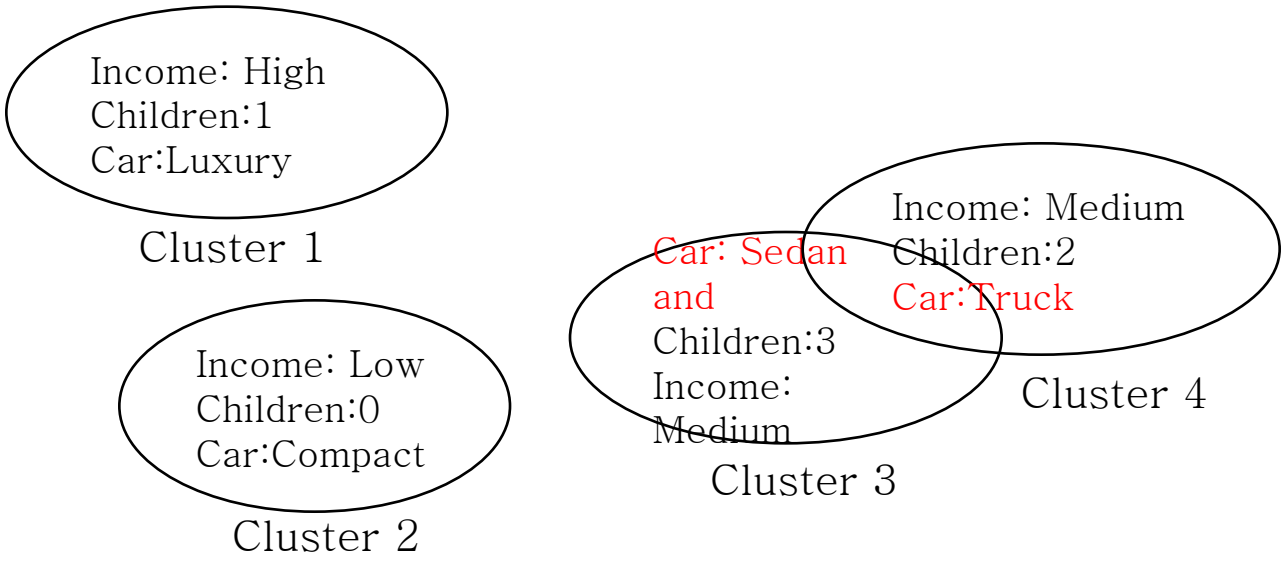
- Some engineering sciences such as pattern recognition, artificial intelligence have been using the concepts of cluster analysis. Typical examples to which clustering has been applied include handwritten characters, samples of speech, fingerprints, and pictures.
- In the life sciences (biology, botany, zoology, entomology, cytology, microbiology), the objects of analysis are life forms such as plants, animals, and insects. The clustering analysis may range from developing complete taxonomies to classification of the species into subspecies. The subspecies can be further classified into subspecies.
- Clustering analysis is also widely used in information, policy and decision sciences. The various applications of clustering analysis to documents include votes on political issues, survey of markets, survey of products, survey of sales programs, and R & D.

A Clustering Example

Retailers want to know where similarities exist in their customer base so they can create and understand different groups to which they can sell and market. They will use a database with rows of customer information and attempt to create customer segments.

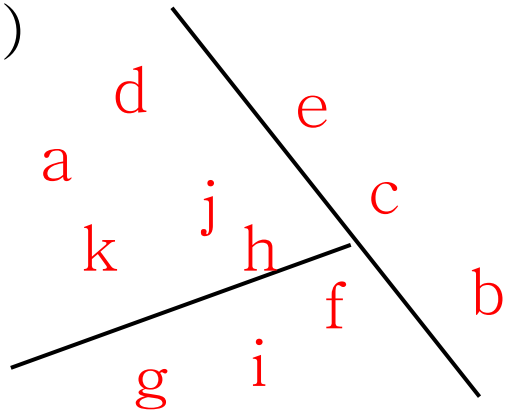
Clustering techniques try to look for similarities and differences within a data set and group similar rows together into clusters or segments. For example, a data set may contain many affluent customers with no children and also may have customers with lower incomes and one parent in the family. During the discovery process, this difference can be used to separate the data into two natural groupings. If more such similarities and differences exist, the data set could be further subdivided.

A Clustering Example

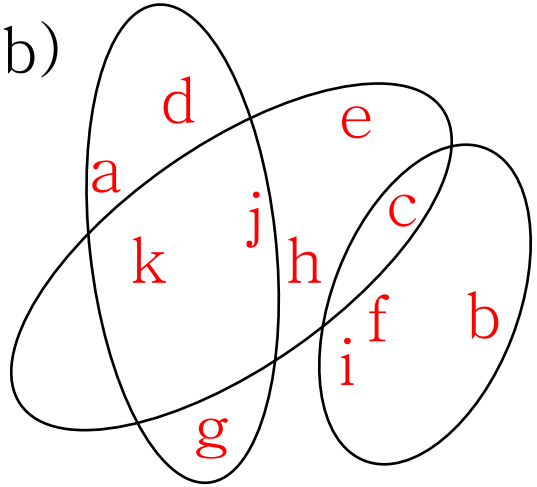


Different ways of representing clusters

(a)



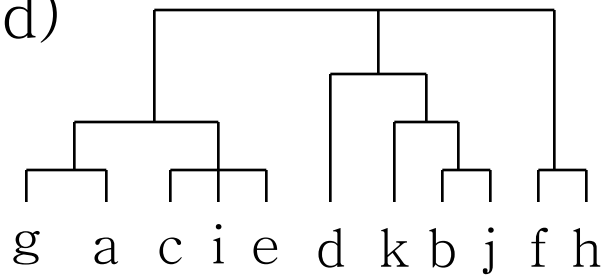
(b)



(c)

	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
…			

(d)



K Means Clustering (Iterative distance-based clustering)

- K means clustering is an effective algorithm to extract a given number of clusters of patterns from a training set. Once done, the cluster locations can be used to classify patterns into distinct classes.

K means clustering (Cont.)

Select the k cluster centers randomly.

Classify the entire training set. For each patern X_i in the training set, find the nearest cluster center C^* and classify X_i as a member of C^*.

For each cluster, recompute its center by finding the mean of the cluster:

$$M_k = \frac{1}{N_k} \cdot \sum_{j=1}^{N_k} X_{jk}$$

where M_k is the new mean, N_k is the number of training patterns in cluster k, and X_{jk} is the j-th pattern belonging to cluster k.

Store the k cluster centers.

Loop until the change in cluster means is less the amount specified by the user.

The drawbacks of K-means clustering

- The final clusters do not represent a global optimization result but only the local one, and complete different final clusters can arise from difference in the initial randomly chosen cluster centers. (fig. 1)
- We have to know how many clusters we will have at the first.

Drawback of K-means clustering (Cont.)

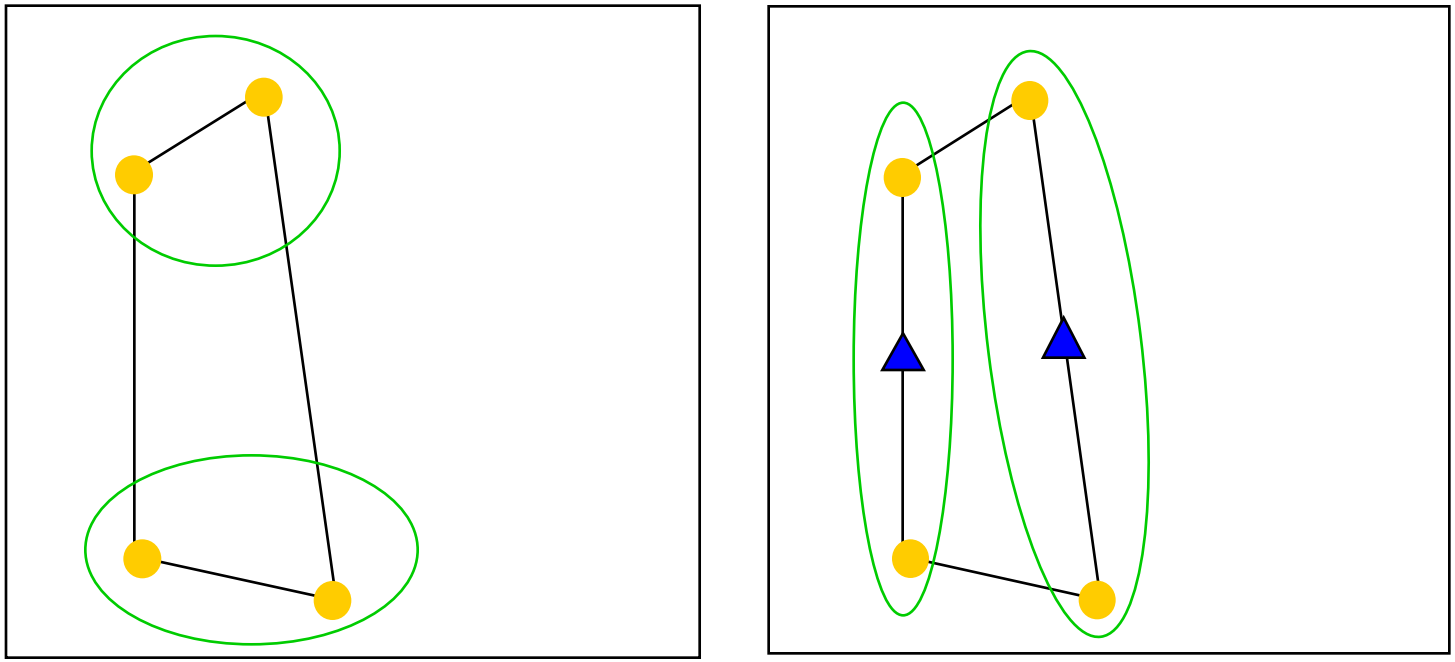


Figure 1

Cluster-based approaches

- External attributes of people and movies to cluster
  - age, gender of people
  - actors and directors of movies.
  - [May not be available]
- Cluster people based on movie preferences
  - misses information about similarity of movies
- Repeated clustering:
  - cluster movies based on people, then people based on movies, and repeat
  - ad hoc, might smear out groups

2.Introduction of GAs

- Inspired by biological evolution.
- Act like Bacteria growing in a petri dish
- Many operators mimic the process of the biological evolution including
  - Natural selection
  - Crossover
  - Mutation

Elements consisting GAs

- Individual (chromosome):
  - feasible solution in an optimization problem
- Population
  - Set of individuals
  - Should be maintained in each generation

Elements consisting GAs

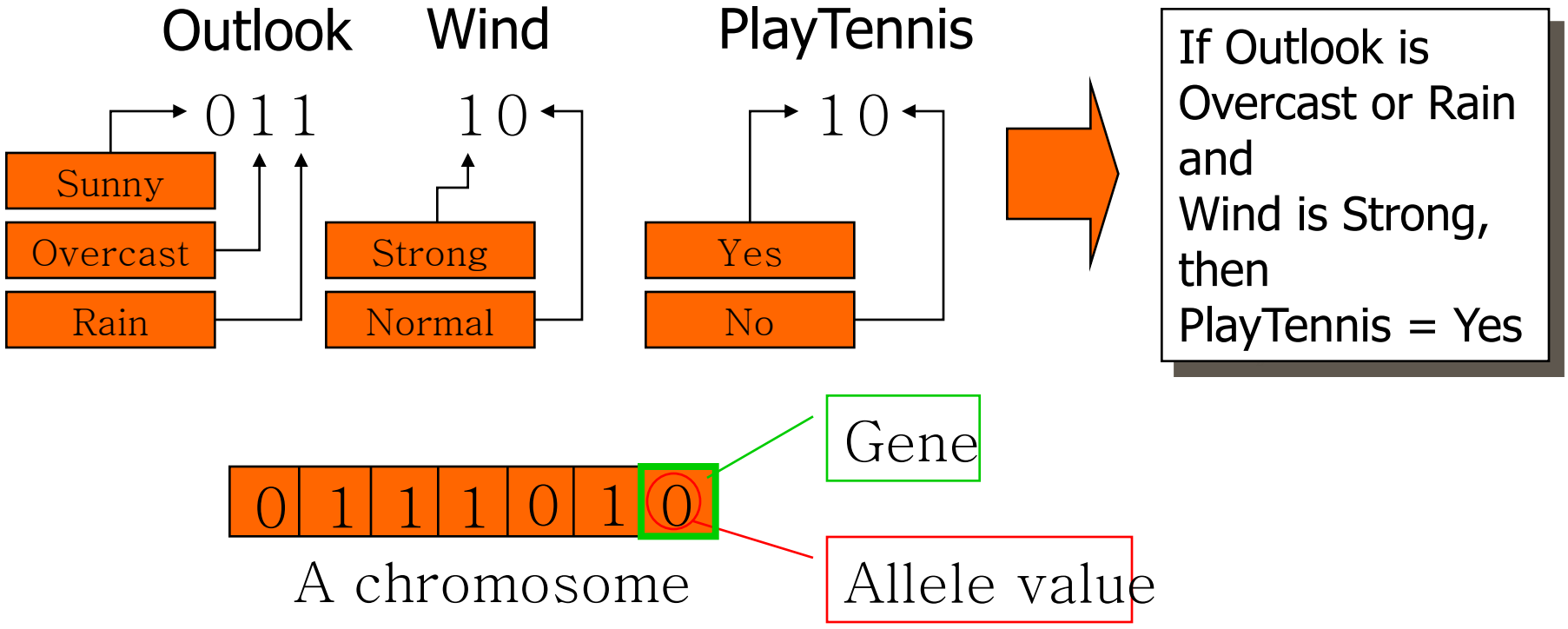
- Genetic operators. (crossover, mutation…)
- Define the fitness function.
  - The fitness function takes a single chromosome as input and returns a measure of the goodness of the solution represented by the chromosome.

Genetic Representation

- The most important starting point to develop a genetic algorithm
- Each gene has its special meaning
- Based on this representation, we can define
  - fitness evaluation function,
  - crossover operator,
  - mutation operator.

Genetic Representation (Cont.)

- **Examples 1**



Genetic Representation (Cont.)

- **Examples 2** (In clustering problem)
  - Each chromosome represents a set of clusters; each gene represents an object; each allele value represents a cluster. Genes with the same allele value are in the same cluster.

1	2	1	4	3	5	5
A	B	C	D	E	F	G

3.Association rules/Link analysis

- Given set T of groups of items
- Example: set of item sets purchased
- Goal: find all rules on itemsets of the form a-->b such that
  - support of a and b > user threshold s
  - conditional probability (confidence) of b given a > user threshold c
- Example: Milk --> bread
- Purchase of product A --> service B

T

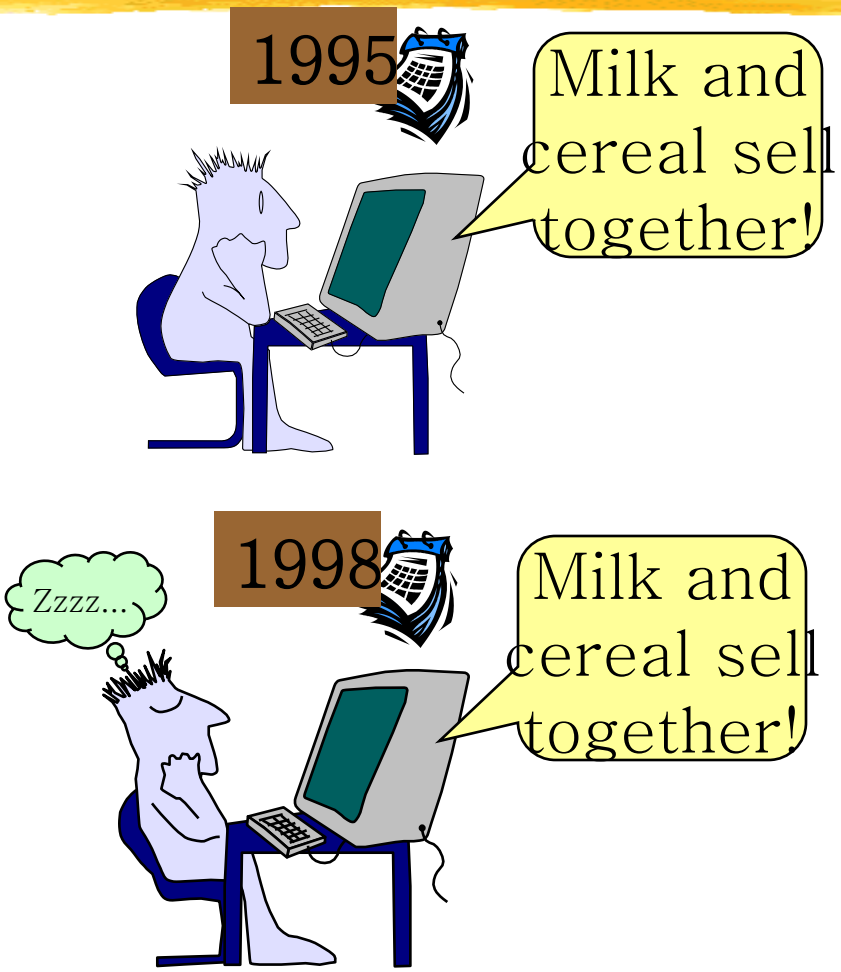
Milk, cereal
Tea, milk
Tea, rice, bread
cereal

Variants

- High confidence may not imply high correlation
- Use correlations. Find expected support and large departures from that interesting..
  - see statistical literature on contingency tables.
- Still too many rules, need to prune...

Prevalent ≠ Interesting

- Analysts already know about prevalent rules
- Interesting rules are those that *deviate* from prior expectation
- Mining's payoff is in finding *surprising* phenomena



What makes a rule surprising?

- Does not match prior expectation
  - Correlation between milk and cereal remains roughly constant over time
- Cannot be trivially derived from simpler rules
  - Milk 10%, cereal 10%
  - Milk and cereal 10% … surprising
  - Eggs 10%
  - Milk, cereal and eggs 0.1% … surprising!
  - Expected 1%

Data Mining in Practice

Application Areas

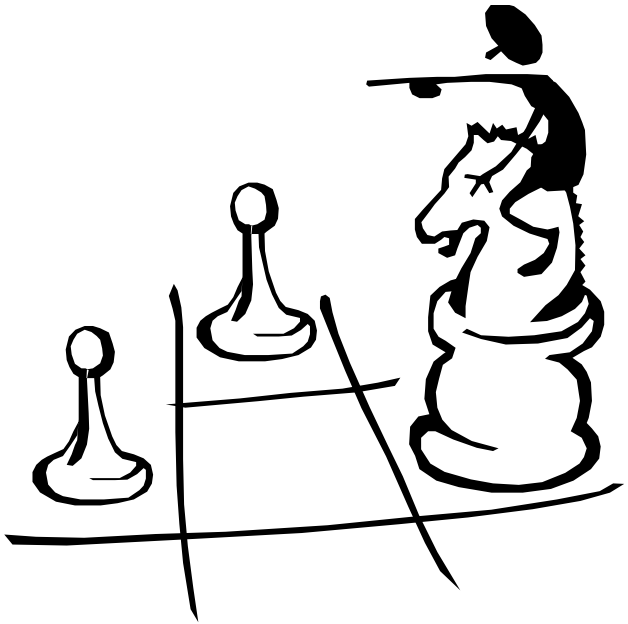
Industry	Application
Finance	Credit Card Analysis
Insurance	Claims, Fraud Analysis
Telecommunication	Call record analysis
Transport	Logistics management
Consumer goods	promotion analysis
Data Service providers	Value added data
Utilities	Power usage analysis

Why Now?

- Data is being produced
- Data is being warehoused
- The computing power is available
- The computing power is affordable
- The competitive pressures are strong
- Commercial products are available

Data Mining works with Warehouse Data

- Data Warehousing provides the Enterprise with a memory



- Data Mining provides the Enterprise with intelligence

Usage scenarios

- Data warehouse mining:
  - assimilate data from operational sources
  - mine static data
- Mining log data
- Continuous mining: example in process control
- Stages in mining:
  - data selection ➔ pre-processing: cleaning ➔ transformation ➔ mining ➔ result evaluation ➔ visualization

Mining market

- Around 20 to 30 mining tool vendors
- Major tool players:
  - Clementine,
  - IBM's Intelligent Miner,
  - SGI's MineSet,
  - SAS's Enterprise Miner.
- All pretty much the same set of tools
- Many embedded products:
  - fraud detection:
  - electronic commerce applications,
  - health care,
  - customer relationship management: Epiphany

Vertical integration: Mining on the web

- Web log analysis for site design:
  - what are popular pages,
  - what links are hard to find.
- Electronic stores sales enhancements:
  - recommendations, advertisement:
  - Collaborative filtering: Net perception, Wisewire
  - Inventory control: what was a shopper looking for and could not find..

OLAP Mining integration

- OLAP (On Line Analytical Processing)
  - Fast interactive exploration of multidim. aggregates.
  - Heavy reliance on manual operations for analysis:
  - Tedious and error-prone on large multidimensional data
- Ideal platform for vertical integration of mining but needs to be interactive instead of batch.

State of art in mining OLAP integration

- Decision trees [Information discovery, Cognos]
  - find factors influencing high profits
- Clustering [Pilot software]
  - segment customers to define hierarchy on that dimension
- Time series analysis: [Seagate's Holos]
  - Query for various shapes along time: eg. spikes, outliers
- Multi-level Associations [Han et al.]
  - find association between members of dimensions
- Sarawagi [VLDB2000]

Data Mining in Use

- The US Government uses Data Mining to track fraud
- A Supermarket becomes an information broker
- Basketball teams use it to track game strategy
- Cross Selling
- Target Marketing
- Holding on to Good Customers
- Weeding out Bad Customers

Some success stories

- Network intrusion detection using a combination of sequential rule discovery and classification tree on 4 GB DARPA data
  - Won over (manual) knowledge engineering approach
  - http://www.cs.columbia.edu/~sal/JAM/PROJECT/ provides good detailed description of the entire process
- Major US bank: customer attrition prediction
  - First segment customers based on financial behavior: found 3 segments
  - Build attrition models for each of the 3 segments
  - 40-50% of attritions were predicted == factor of 18 increase
- Targeted credit marketing: major US banks
  - find customer segments based on 13 months credit balances
  - build another response model based on surveys
  - increased response 4 times -- 2%