Data Mining: An Overview

David Madigan

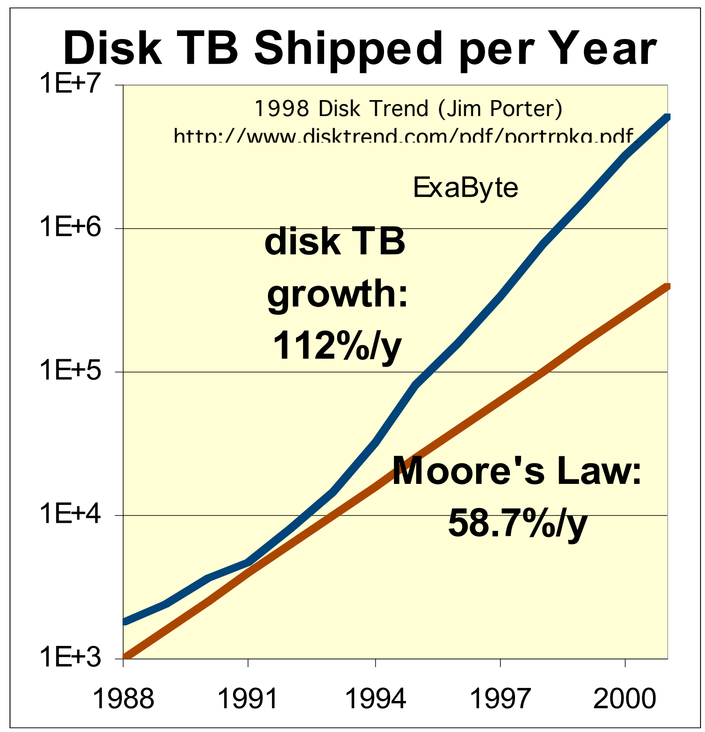
http://www.stat.columbia.edu/~madigan

Overview

- Brief Introduction to Data Mining
- Data Mining Algorithms
- Specific Examples
  - Algorithms: Disease Clusters
  - Algorithms: Model-Based Clustering
  - Algorithms: Frequent Items and Association Rules
- Future Directions, etc.

Of “Laws”, Monsters, and Giants…

- Moore’s law: processing “capacity” doubles every 18 months : CPU, cache, memory
- It’s more aggressive cousin:
  - Disk storage “capacity” doubles every 9 months



What do the two “laws” combined produce?

A rapidly growing gap between our ability to generate data, and our ability to make use of it.

What is Data Mining?

Finding interesting structure in data

- *Structure:* refers to statistical patterns, predictive models, hidden relationships

- Examples of tasks addressed by Data Mining
  - Predictive Modeling (classification, regression)
  - Segmentation (Data Clustering)
  - Summarization
  - Visualization


Third IEEE International Conference on
Data Mining
Data Mining and Knowledge Discovery
www.kdd2004.com
KDD08
LAS VEGAS | 24-27 AUGUST 2008
The 14th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining
Rapidly build and deploy data mining solutions with Clementine 8.0
Clementine
SPSS
Mastering Data Mining
The Art and Science of Customer Relationship Management
A. Berry Linoff
sas
SAS® Enterprise Miner™ finds patterns in the most complex data structures. See how.
Insightful Miner
Easy to Use & Extensible Data Mining
Build predictive models easily
Modern visual interface
Advanced analytic methods
Scalable capabilities
Free Webcast & Whitepaper!
CRISP-DM
CRoss Industry Standard Process for Data Mining
ORACLE DATABASE 10g
Oracle Data Mining
IBM
DB2 Intelligent Miner
PRINCIPLES OF Data Mining
David Hand
Heikki Mannila
Padhraic Smyth

Data understanding

- Generate subsets of data automatically from graphs and tables
- Show summary statistics, histograms, and distribution graphics for each data field, and display them in an easy-to-read matrix with the data audit node. This provides you with a comprehensive first look at your data.
- Visually interact with your data
  - Select node or field and view information in a table
  - Create histograms, distributions, line plots, and point plots
  - Display 3-D, panel, and animated graphs
  - Use Web association detection

Modeling

- Prediction and classification
  - Neural networks (multi-layer perceptrons trained using error-back propagation with momentum, radial basis function, and Kohonen network)
  - Decision trees and rule induction [C5.0 and Classification and Regression Trees (C&RT)]
  - Linear regression, logistic regression, and multinomial logistic regression
- Clustering and segmentation
  - Kohonen network, K-means, and TwoStep
  - View summary statistics and distributions for fields between clusters using the Cluster Viewer
- Association detection
  - GRI, apriori, and sequence
- Data reduction
  - Factor analysis and principle components analysis

Stories – Non–actionable Segment

- A bank discovered a cluster of customers that have left the bank:
  - Older than the average customer.
  - Less likely to have a mortgage.
  - Less likely to have a credit card.

They were also...

Ronny Kohavi, ICML 1998

Data Mining Algorithms

> "A data mining algorithm is a well-defined procedure that takes data as input and produces output in the form of models or patterns"

Hand, Mannila, and Smyth

"well-defined": can be encoded in software

"algorithm": must terminate after some finite number of steps

Algorithm Components

1. The *task* the algorithm is used to address (e.g. classification, clustering, etc.)

2. The *structure* of the model or pattern we are fitting to the data (e.g. a linear regression model)

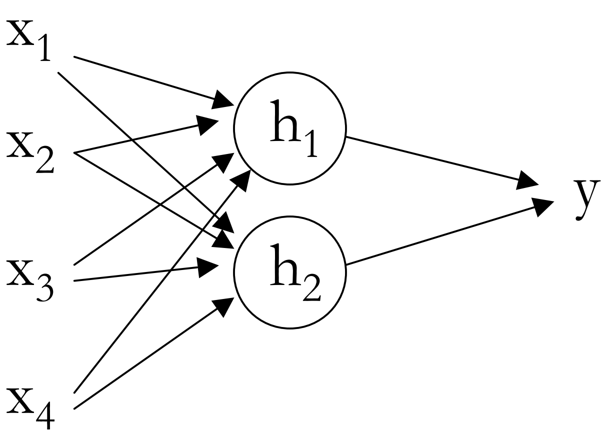
3. The *score function* used to judge the quality of the fitted models or patterns (e.g. accuracy, BIC, etc.)

4. The *search or optimization method* used to search over parameters and/or structures (e.g. steepest descent, MCMC, etc.)

5. The *data management technique* used for storing, indexing, and retrieving data (critical when data too large to reside in memory)

	CART	**Backpropagation**	**A Priori**
Task	Classification and Regression	Regression	Rule Pattern Discovery
Structure	Decision Tree	Neural Network (Nonlinear functions)	Association Rules
Score Function	Cross-validated Loss Function	Squared Error	Support/Accuracy
Search Method	Greedy	Gradient Descent	Breadth-First with Pruning
Data Management Technique	Unspecified	Unspecified	Linear Scans

Backpropagation data mining algorithm



$$s_1 = \sum_{i=1}^{4} \alpha_i x_i ; s_2 = \sum_{i=1}^{4} \beta_i x_i$$

$$h(s_i) = 1 \big/ (1 + e^{-s_i})$$

$$y = \sum_{i=1}^{2} w_i h_i$$

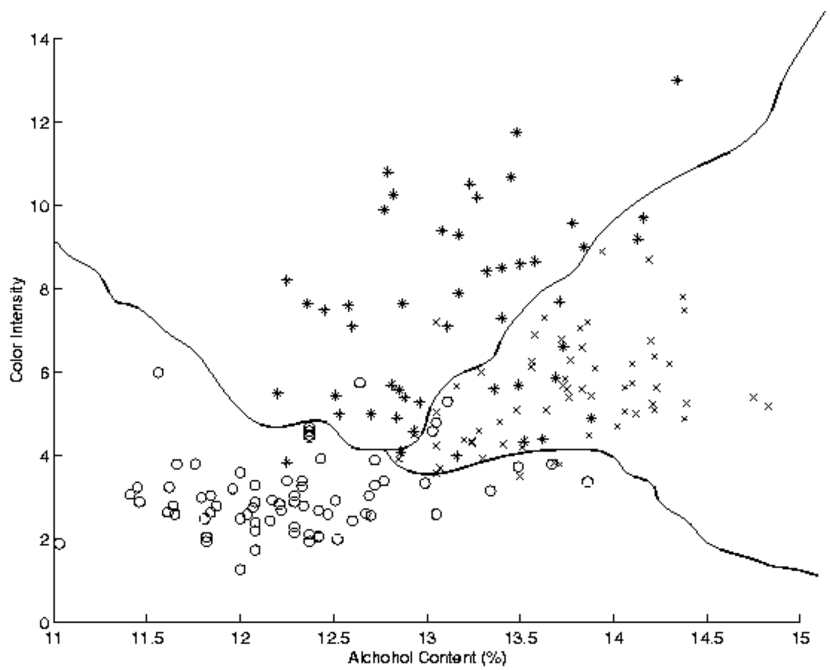
- vector of p input values multiplied by $\overset{4}{p} \times \overset{2}{d_1}$ weight matrix
- resulting d_1 values individually transformed by non-linear function
- resulting d_1 values multiplied by $d_1 \times \overset{1}{d_2}$ weight matrix

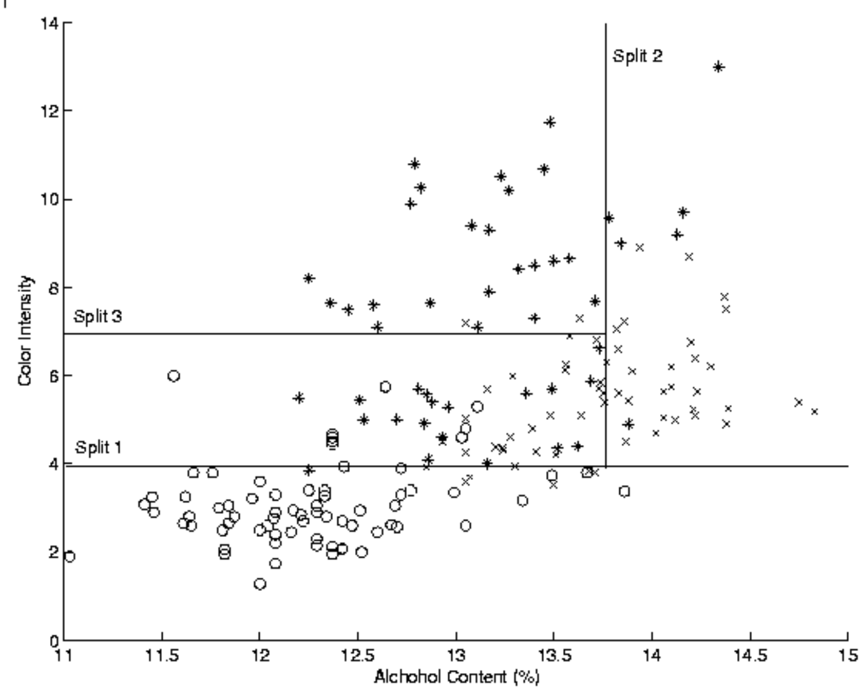
Backpropagation (cont.)

Parameters: $\alpha_1, \ldots, \alpha_4, \beta_1, \ldots, \beta_4, w_1, w_2$

Score: $S_{SSE} = \sum_{i=1}^{n} (y(i) - \hat{y}(i))^2$

Search: steepest descent; search for structure?





Models and Patterns

Models

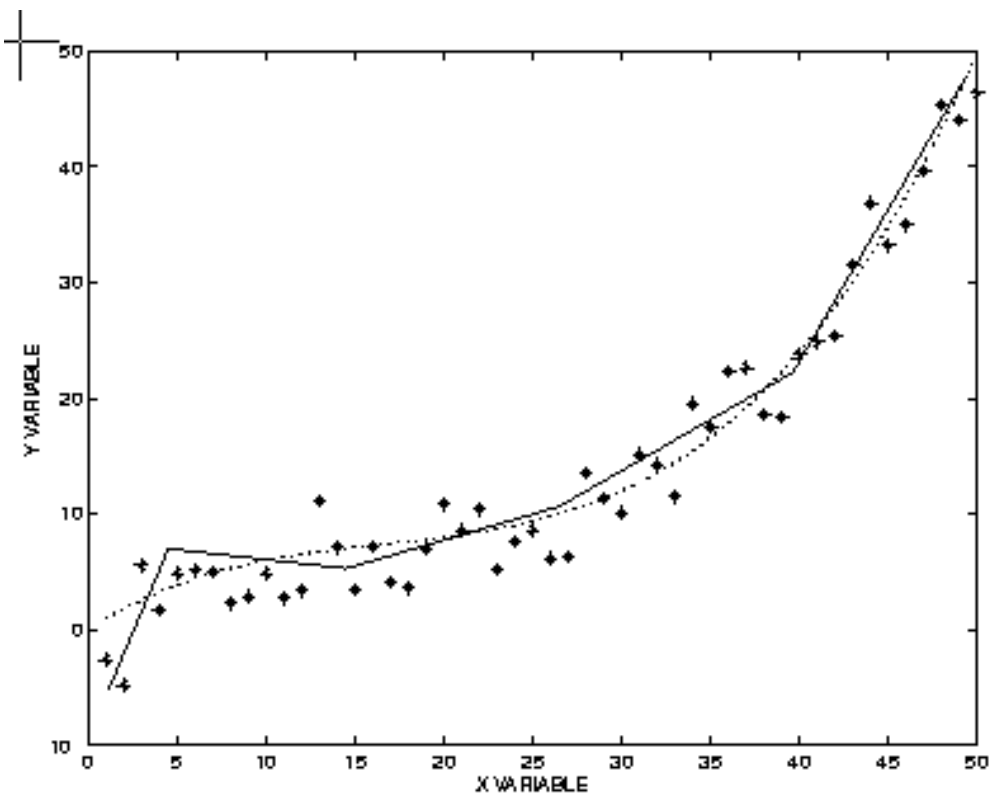
Prediction

Probability Distributions

Structured Data

•Linear regression

•Piecewise linear



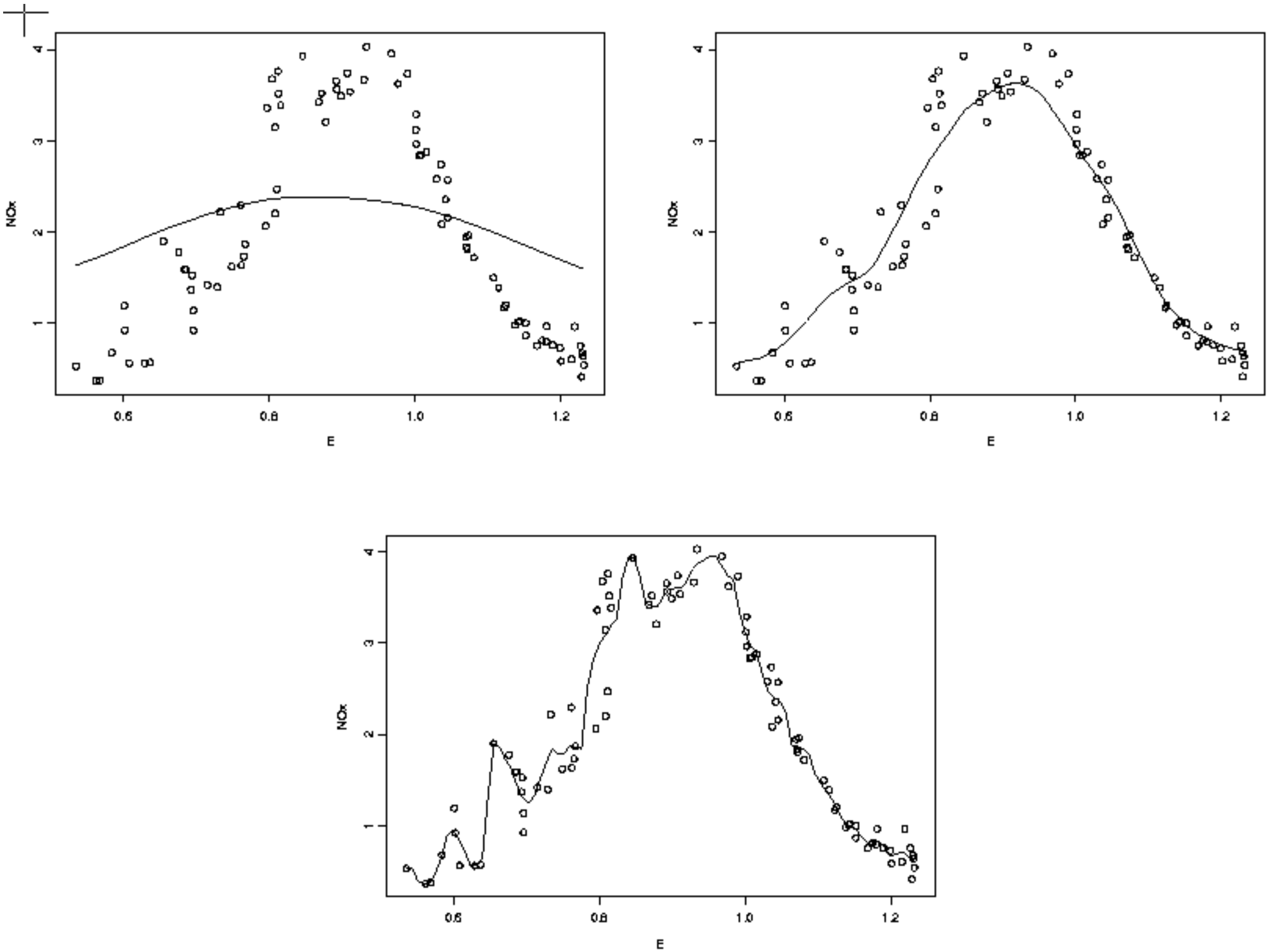
Models

Prediction

- Linear regression
- Piecewise linear
- Nonparamatric regression

Probability Distributions

Structured Data



NOx
E
0.6
0.8
1.0
1.2
1
2
3
4

Models

Prediction

Probability Distributions

Structured Data

- Linear regression
- Piecewise linear
- Nonparametric regression
- Classification

logistic regression

naïve bayes/TAN/bayesian networks

NN

support vector machines

Trees

etc.

Models

Prediction

Probability Distributions

Structured Data

•Linear regression

•Piecewise linear

•Nonparametric regression

•Classification

•Parametric models

•Mixtures of parametric models

•Graphical Markov models (categorical, continuous, mixed)

Models

Prediction

- Linear regression
- Piecewise linear
- Nonparametric regression
- Classification

Probability Distributions

- Parametric models
- Mixtures of parametric models
- Graphical Markov models (categorical, continuous, mixed)

Structured Data

- Time series
- Markov models
- Mixture Transition Distribution models
- Hidden Markov models
- Spatial models

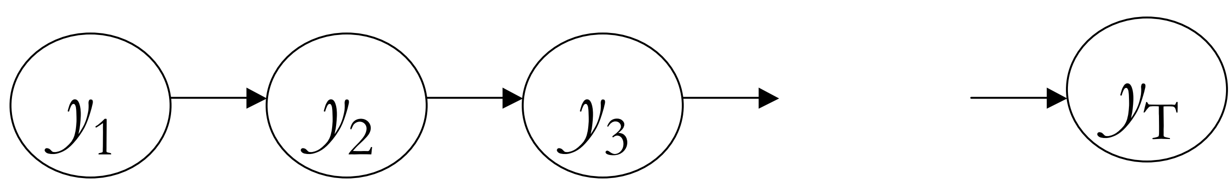
Markov Models

First-order: $$p(y_1,\ldots,y_T) = p_1(y_1)\prod_{t=2}^{T} p_t(y_t \mid y_{t-1})$$

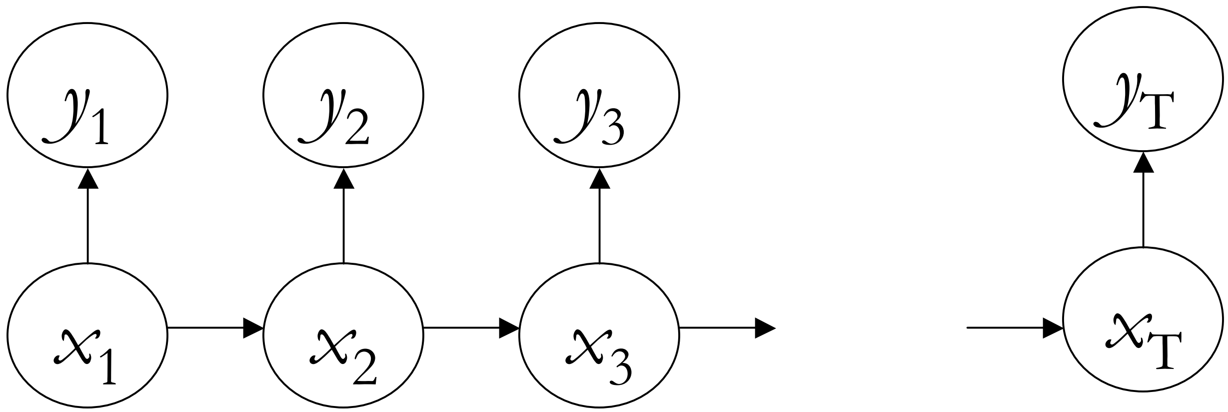
e.g.: $$p(y_t \mid y_{t-1}) = \frac{1}{\sqrt{2\pi\sigma}} \exp - \frac{1}{2}\left(\frac{y_t - g(y_{t-1})}{\sigma}\right)^2$$

g linear $\Rightarrow$ standard first-order auto-regressive model

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + e \qquad e \sim N(0,\sigma^2)$$



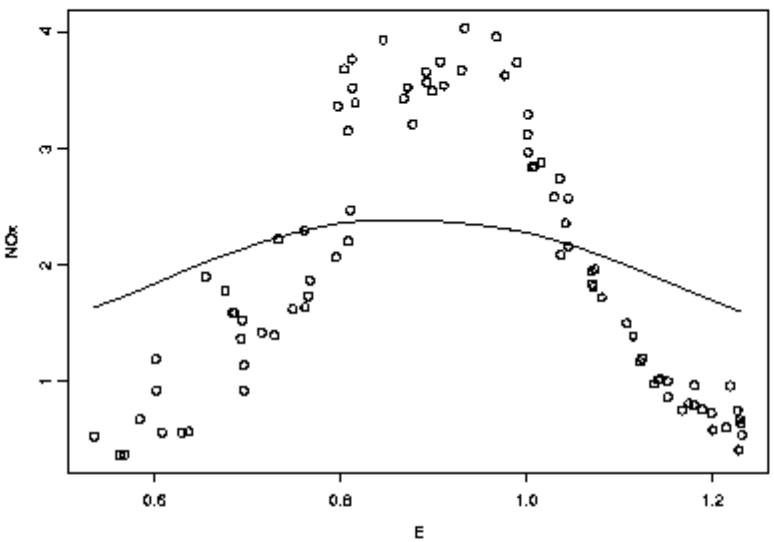
First-Order HMM/Kalman Filter

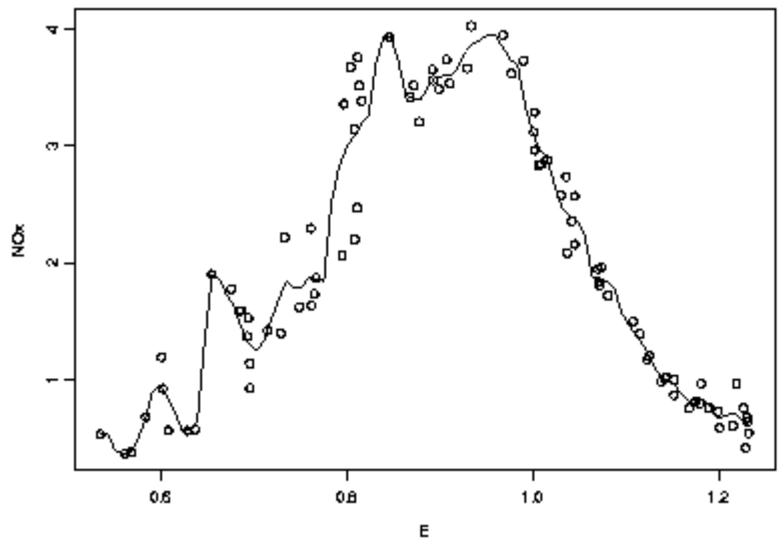


$$p(y_1,\dots,y_T,x_1,\dots,x_T) = p_1(x_1)\,p_1(y_1 \mid x_1)\prod_{t=2}^{T} p(y_t \mid x_t)\,p(x_t \mid x_{t-1})$$

Note: to compute $p(y_1,\dots,y_T)$ need to sum/integrate over all possible state sequences...

Bias-Variance Tradeoff





High Bias - Low Variance

Low Bias - High Variance

"overfitting" - modeling the random component

Score function should embody the compromise

The Curse of Dimensionality

$$X \sim \mathrm{MVN}_p\,(\mathbf{0}\,,\,\mathbf{I})$$

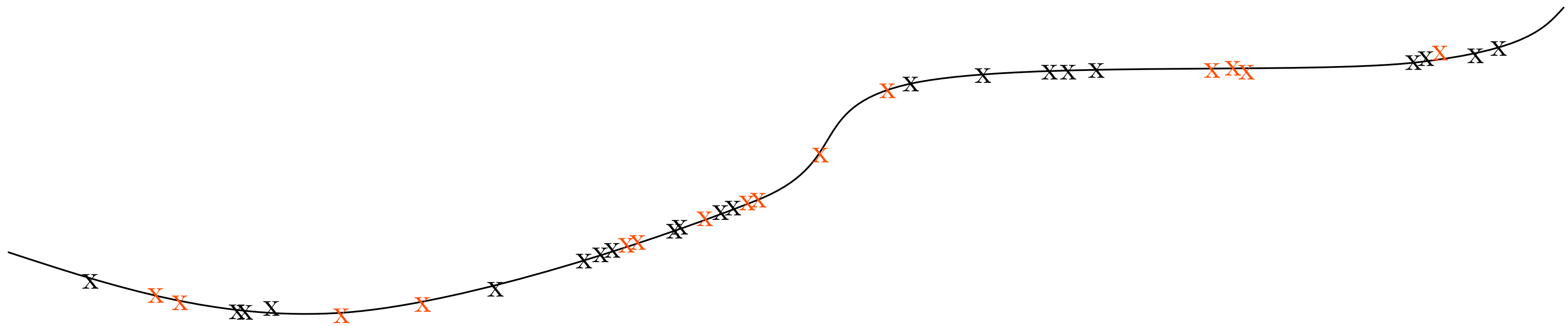
- Gaussian kernel density estimation
- Bandwidth chosen to minimize MSE at the mean
- Suppose want: $\left.\dfrac{E[(\hat{p}(x)-p(x))^2]}{p(x)^2} < 0.1\right|_{x=0}$

Dimension	**# data points**
1	**4**
2	**19**
3	**67**
6	**2,790**
10	**842,000**

Patterns

- Global
  - Clustering via partitioning
  - Hierarchical Clustering
  - Mixture Models
- Local
  - Outlier detection
  - Changepoint detection
  - Bump hunting
  - Scan statistics
  - Association rules

Scan Statistics via Permutation Tests



The curve represents a road

Each “x” marks an accident

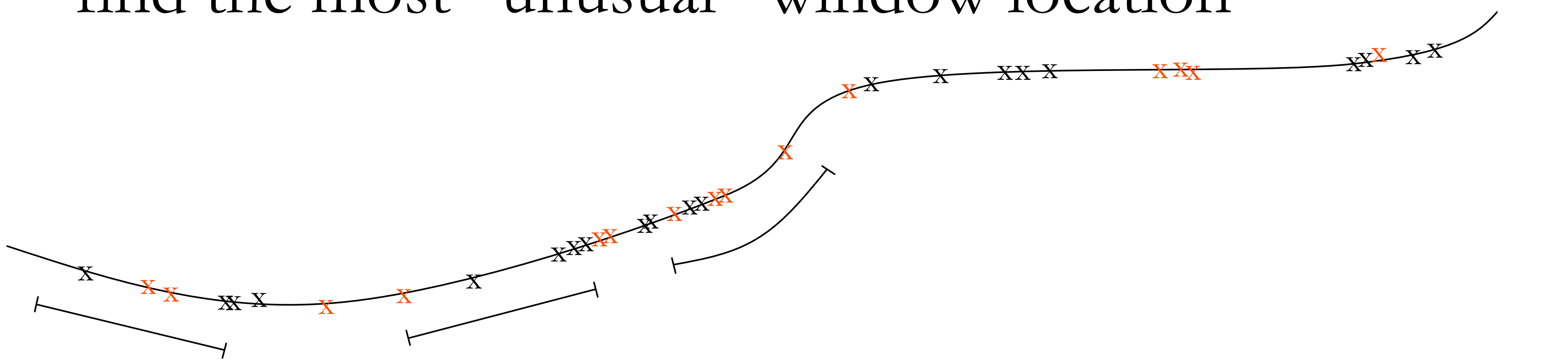
Red “x” denotes an injury accident

Black “x” means no injury

Is there a stretch of road where there is an unusually large fraction of injury accidents?

Scan with Fixed Window

- If we know the length of the "stretch of road" that we seek, e.g., we could slide this window long the road and find the most "unusual" window location



How Unusual is a Window?

- Let p_W and $p_{\neg W}$ denote the true probability of being red inside and outside the window respectively. Let (x_W, n_W) and $(x_{\neg W}, n_{\neg W})$ denote the corresponding counts
- Use the GLRT for comparing H_0: $p_W = p_{\neg W}$ versus H_1: $p_W \neq p_{\neg W}$

$$\lambda = \frac{[(x_W + x_{\neg W})/(n_W + n_{\neg W})]^{x_W + x_{\neg W}} [1 - ((x_W + x_{\neg W})/(n_W + n_{\neg W}))]^{n_W + n_{\neg W} - x_W - x_{\neg W}}}{(x_W / n_W)^{x_W} [1 - (x_W / n_W)]^{n_W - x_W} (x_{\neg W} / n_{\neg W})^{x_{\neg W}} [1 - (x_{\neg W} / n_{\neg W})]^{n_{\neg W} - x_{\neg W}}}$$

- lambda measures how unusual a window is

$-2 \log \lambda$ here has an asymptotic chi-square distribution with 1df

Permutation Test

- Since we look at the smallest λ over *all* window locations, need to find the distribution of smallest-λ under the null hypothesis that there are no clusters
- Look at the distribution of smallest-λ over say 999 random relabellings of the colors of the x's

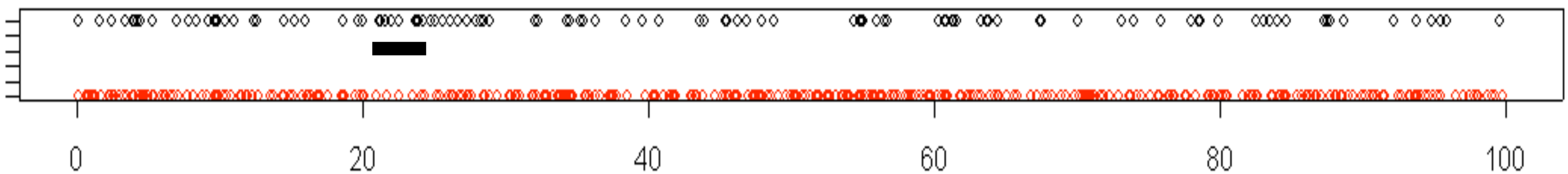
	smallest-λ
xx x xxx x xx x xx x	0.376
xx x xxx x xx x xx x	0.233
xx x xxx x xx x xx x	0.412
xx x xxx x xx x xx x	0.222

…

- Look at the position of observed smallest-λ in this distribution to get the scan statistic p-value (e.g., if observed smallest-λ is 5th smallest, p-value is 0.005)

Variable Length Window

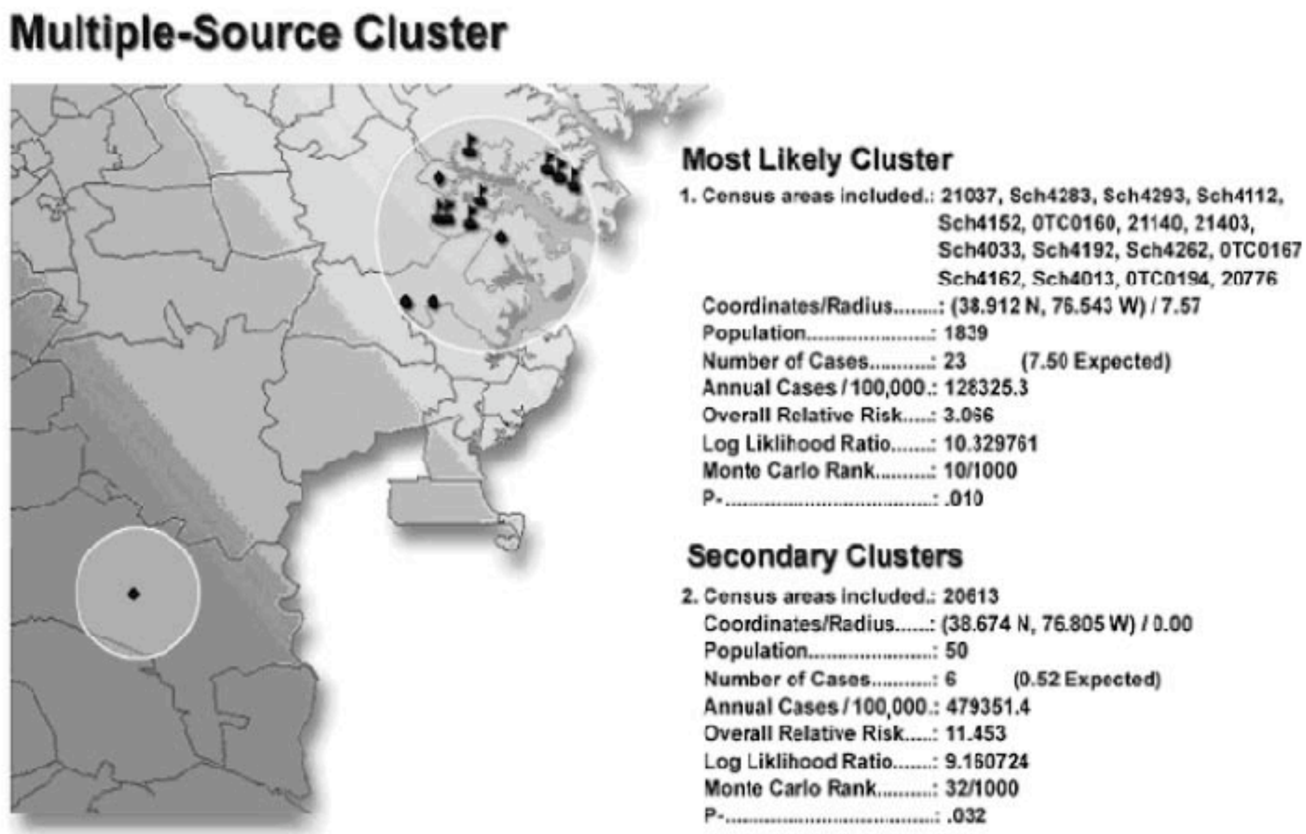
- No need to use fixed-length window. Examine all possible windows up to say half the length of the entire road



o = fatal accident
o = non-fatal accident

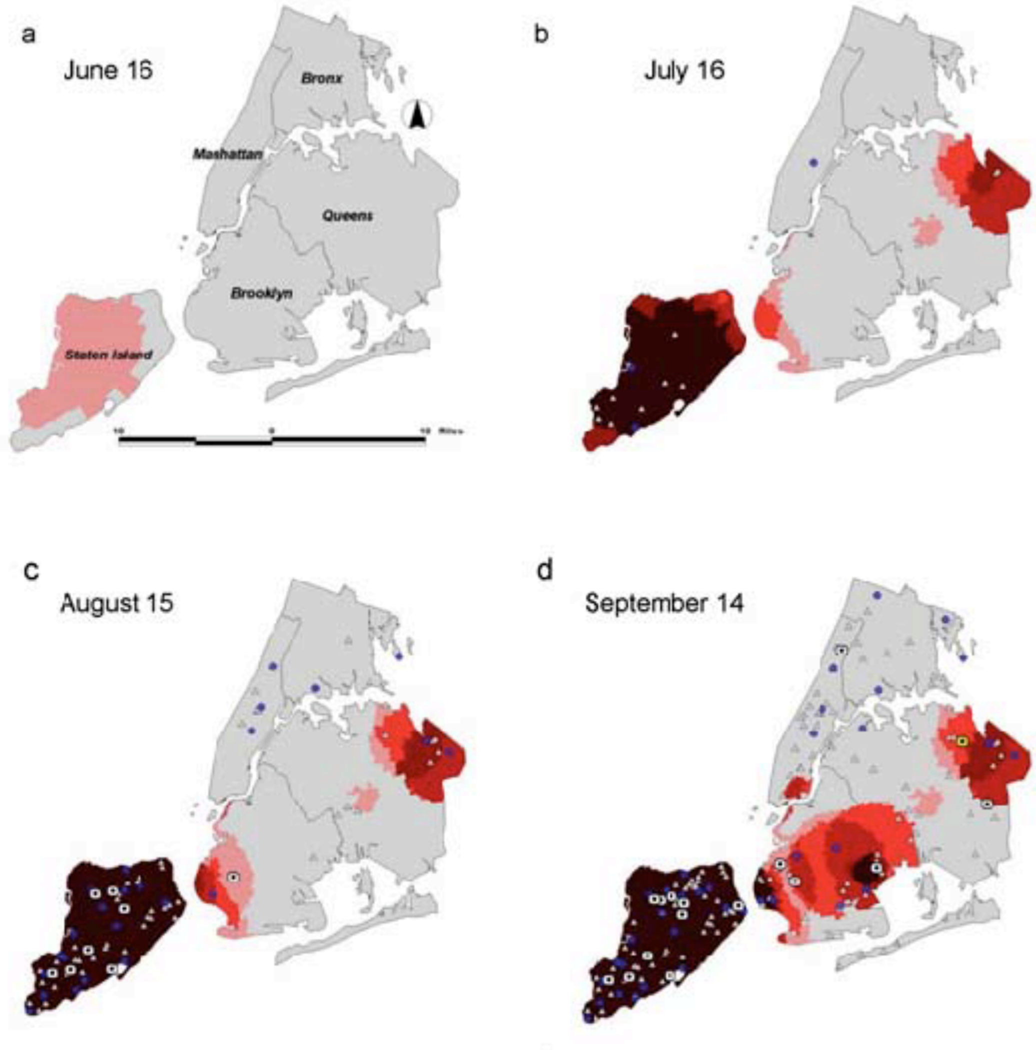
Spatial Scan Statistics

- Spatial scan statistic uses, e.g., circles instead of line segments



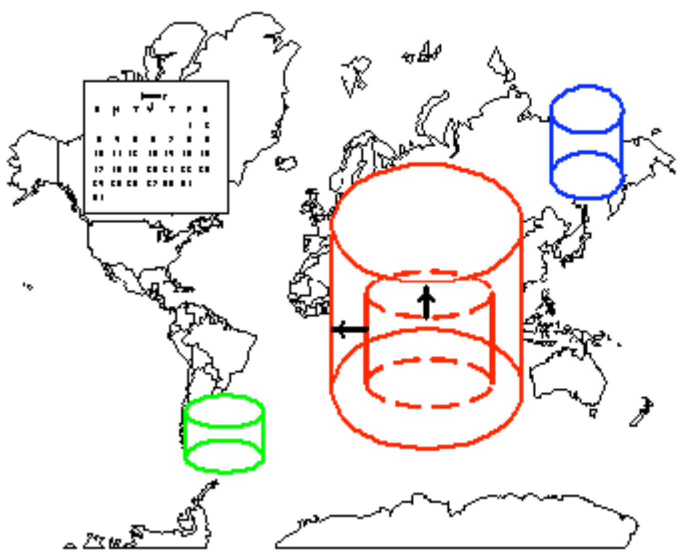
Dead Bird Clusters as an Early Warning System for West Nile Virus Activity

Farzad Mostashari,* Martin Kulldorff,† Jessica J. Hartman,* James R. Miller,* and Varuni Kulasekera*



Spatial-Temporal Scan Statistics

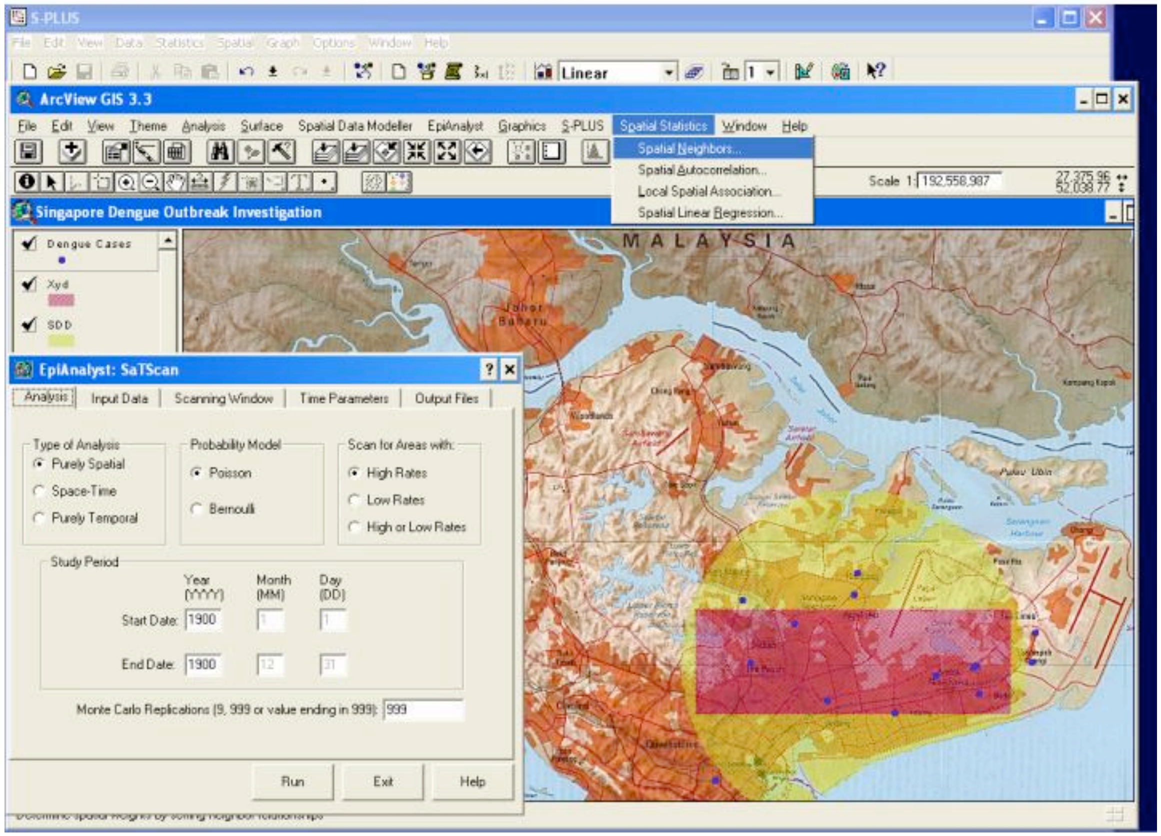
- Spatial-temporal scan statistic use cylinders where the height of the cylinder represents a time window



Other Issues

- Poisson model also common (instead of the bernoulli model)
- Covariate adjustment
- Andrew Moore's group at CMU: efficient algorithms for scan statistics

Software: SaTScan + others

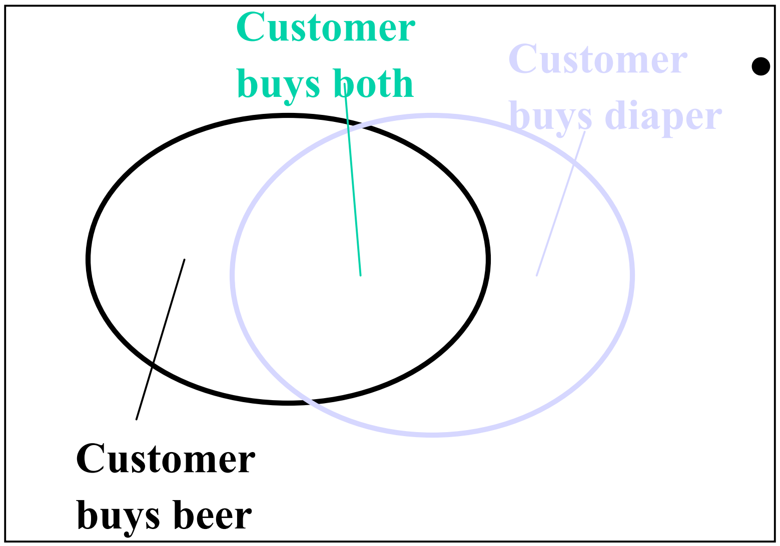


http://www.satscan.org

http://www.phrl.org

http://www.terraseer.com

Association Rules: Support and Confidence



- Find all the rules $Y \Rightarrow Z$ with minimum confidence and support
  - support, s, probability that a transaction contains $\{Y \& Z\}$
  - confidence, c, conditional probability that a transaction having Y also contains Z

Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Let minimum support 50%, and minimum confidence 50%, we have

- $A \Rightarrow C$ (50%, 66.6%)
- $C \Rightarrow A$ (50%, 100%)

Mining Association Rules—An Example

Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Min. support 50%
Min. confidence 50%

Frequent Itemset	Support
{A}	75%
{B}	50%
{C}	50%
{A,C}	50%

For rule $A \Rightarrow C$:

support = support($\{A \& C\}$) = 50%

confidence = support($\{A \& C\}$)/support($\{A\}$) = 66.6%

The Apriori principle:

Any subset of a frequent itemset must be frequent

Mining Frequent Itemsets: the Key Step

- Find the *frequent itemsets*: the sets of items that have minimum support
  - A subset of a frequent itemset must also be a frequent itemset
    - i.e., if $\{AB\}$ is a frequent itemset, both $\{A\}$ and $\{B\}$ should be a frequent itemset
  - Iteratively find frequent itemsets with cardinality from 1 to k (k-itemset)
- Use the frequent itemsets to generate association rules.

The Apriori Algorithm

- Join Step: C_k is generated by joining L_{k-1} with itself
- Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset
- Pseudo-code:

C_k: Candidate itemset of size k
L_k : frequent itemset of size k

L_1 = {frequent items};
for (k = 1; L_k !=∅; k++) **do begin**
  C_{k+1} = candidates generated from L_k;
  for each transaction t in database do
    increment the count of all candidates in C_{k+1} that are contained in t
  L_{k+1} = candidates in C_{k+1} with min_support
  end
return $\cup_k L_k$;

The Apriori Algorithm — Example

Database D

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Scan D

C_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

L_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{5}	3

C_2

itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

Scan D

C_2

itemset	sup
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

L_2

itemset	sup
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

C_3

itemset
{2 3 5}

Scan D

L_3

itemset	sup
{2 3 5}	2

Association Rule Mining: A Road Map

- Boolean vs. quantitative associations (Based on the types of values handled)
  - buys(x, "SQLServer") ^ buys(x, "DMBook") $\rightarrow$ buys(x, "DBMiner") [0.2%, 60%]
  - age(x, "30..39") ^ income(x, "42..48K") $\rightarrow$ buys(x, "PC") [1%, 75%]
- Single dimension vs. multiple dimensional associations (see ex. Above)
- Single level vs. multiple-level analysis
  - What brands of beers are associated with what brands of diapers?
- Various extensions (thousands!)

Representation Design and Brute-force Induction in a Boeing Manufacturing Domain

Patricia Riddle* **Richard Segal & Oren Etzioni**

If the nest's material is A and it is from batch B, then it is 4 times as likely to have a TypeC reject.

A nest which goes through station B is 2 times as likely to be rejected as a nest which goes through station A.

If a nest is at station A before station B for over 32 minutes, then it is 4.5 times as likely to be a TypeC reject.

If a nest is at station X over 51 minutes, then it is 3 times as likely to be rejected.

If the nest's material is X and it is at station Y before it goes to station Z, then it is 2 times as likely to be a TypeW reject.

A nest is 2 times as likely to get a TypeZ reject on a Friday.

Part A is 1.5 times as likely to get a TypeX reject.

A nest which spends less than 9 minutes in station X is 6 times more likely to be OCC3 than a nest which spends more than 9 minutes in station X. OCC3 means that a nest spent 6 to 21 minutes in station Z.

If the nest's material is A, then the probability of alarm X reduces by 25%.