Data Mining

Bob Stine

Department of Statistics, The Wharton School

University of Pennsylvania

Philadelphia PA

www-stat.wharton.upenn.edu/∼bob

June 10, 2000

- Describe the process and terminology (examples)
- Relationship to statistics
- Teaching business statistics
  - In stat classes themselves
  - In relationship to other classes
- Bottom line... Data mining has interesting
  - Data analysis problems
  - Algorithms and estimators

  We can/should use them in our courses...

What is Data Mining?

Data mining

- *Process* of using data analysis to discover patterns
- Patterns, relationships lead to predictions.
- Statistics is (could be) a key step in model building

Driven by technology

- Confluence of

$$\begin{array}{c}\text{Data}\\ \text{warehouse}\end{array} + \begin{array}{c}\text{Fast, inexpensive}\\ \text{hardware}\end{array} \rightarrow \text{Data mining}$$

- ROI from database, transaction records
- "Easy to use" GUI interface
  - Never see what's under the hood

Data mining = data dredging?

- Stepwise regression is back!
- Replace thinking with searching
- At odds with careful analysis
- Just hype... (Friedman, 97)
  Profit lies in mining the miners.

Examples of Data Mining

Credit modeling, scoring

Can you predict whether someone will declare bankruptcy?

Risk factors for a disease

Which factors indicate risk for osteoporosis?

Direct mail advertising

Who should receive a solicitation for a donation?

Internet/e-commerce

If you bought this CD, which others might you buy?

Financial forecasting

Which factors predict movement in stock returns?

Aren't these great statistics problems?

- Contemporary (e-commerce after all)
- Big dollars at stake
- Role in business strategy
- Lots of computing with rich data structures

Predicting Bankruptcy

Goal

Predictive model for personal bankruptcy:

Based on the recent history of an *individual* credit-card holder, estimate the probability that the card holder will declare bankruptcy during the next credit cycle. (Too late?)

Data

- Large data set: 250,000 bankcard accounts
- Short monthly time series for each account
- Credit limits, spend, payments, bureau info
- Demographic background
- No transaction data, but you could

Needle in the haystack

About 2,500 bankruptcies out of

$$24 \times 250,000 = 6 \text{ million account-months}$$

Trade-off

Profitable customers look risky. Want to lose them?

"Borrow lots of money and pay it back slowly."

Predicting Osteoporosis

Goal

Estimate likelihood of osteoporosis in women without requiring an expensive diagnostic x-ray procedure.

Preference for self-reported measurements.

Osteoporosis

Degenerative loss of calcium as we age leads to weakened bones which easily fracture in a fall, leading to hospitalization and further complications.

Data

- Two samples
  - Training data: 1000 women from clinics
  - Validation data: *subsequent* sample of 250
- Self-reported forms
- Clinical data obtained by MD
- Lab measurements (blood sample analysis)

Trade-off

Sensitivity vs specificity:

Want to find those at high risk without sending too many for expensive diagnosis.

Selling CDs

Manage customer life-cycle

- CDNow web based CD vendor
- Repeat purchase, maintain interest in web site
- E-commerce!

Collaborative filtering

- Purchase data is a large, sparse matrix

	BB King	BBoys	Spears	JJack	YY Ma
Lyle	y	y			
Ellen	y	y			y
Dean		y	y	y	
Jason	y				y

- Karen bought B Boys – What to recommend?
- Net Perceptions
- Combinatorial chemistry and new drug discovery

Cross-selling

- Amazon book groups
- Cross-selling related products
- Oops, that was a gift.

Steps of Data Mining Process

Problem identification

As part of general business strategy.

1. Data preparation

- Extract from database
- Transformation
  - Subsets and group indicators
  - Interactions, other forms of nonlinearity
  - Discretize continous predictors
  - Substantive transformation
- "Rectangularize"

2. Data analysis

- Classification and regression
  - "Computer science" (neural nets, boosting)
- Associative methods (clustering)

3. Validation

- Cross-validation, experiments
- Decision theory and feature selection

Implementation of results

Step 1: Data Preparation

Black hole of data mining

60% to 90% of effort to prepare data

data bases $\Rightarrow$ rectangular array

Merging data

Bankruptcy data arrived in 35 files:

24 monthly, 8 quarterly, 2 annual, plus descriptive

- Convert incoherent formats, identify errors
  "its just a space" "that's another version of Y"
- Align events in time
  months/quarters prior to bankruptcy
- Devise sampling scheme
  subsample with all bankruptcy events included

Missing data

- Structural: Subsets with more extensive data.
- Temporal: "We didn't collect that back then."
- Some are informative?
- Handle by adding indicator, viewing as interaction

Feature set grows rapidly

Step 2: Data Analysis – Modeling

Not the standard textbook case

- Often violate most of the standard assumptions:
  independence, constant variance, normal sample

Don't know the real model

- Unsure of form, much less which predictors
- Consistency in usual sense is not an issue.
- Not the traditional test of $H_0 : \beta_3 = 0$

Many potential predictors

- Access to large database, data warehouse
- Automated data collection streams
- Nonlinearity, interactions, subsets:
  "**the** MBA problem"

Differing expectations for predictors

Consider the osteoporosis example...

- Some *likely* to be useful ($|t| \approx 10$)
- Some *might* be useful ($|t| \approx 2$) $\Leftarrow$
- Others that somebody collects and are available.

What model?

Do you want a model?

- Software often hides form from miner.
- Interpretation of model?
- Diagnostics?

Tables may be enough

- Queries for a database
  "Find me all of the customers from Florida who bought Backstreet Boys and Boyz II Men last month"
- (Relational) on-line analytic processing (ROLAP)
- May be enough for the purpose at hand

Data visualization

- Fly through the data

Under the Hood

Models actually used vary

Familiar choices

- Linear regression model
- Additive models (smoothing), splines (MARS)
- Decision trees (CART)
- Bayesian methods that average over many models

Less familiar

- Neural nets (projection pursuit regression)
- Boosting
- Decision trees (C4.5, CHAID)
- Rule induction
- Association rules

Transformations remain an art form

- Informed combinations e.g. *limit – curr balance*
- Time lags, other transformations e.g. logs
- Interactions, "less-informed" combinations

Alternative Model: Neural Nets

Are NNs really new/useful?

Another variation on regression, closely related to projection pursuit regression and the origins of smoothing.

Structure

- Sigmoid basis functions, like the logistic CDF

$$\varphi(x) = \frac{1}{1 + e^{-x}}$$

- Linear combinations of features (direction)

$$Z_j = \alpha_{1j} X_1 + \alpha_{2j} X_2 + \cdots + \alpha_{pj} X_p, \quad \sum_i \alpha_{ij}^2 = 1$$

- Weight and combine

$$\hat{Y} = \beta_0 + \beta_1 \varphi(Z_1) + \cdots + \beta_k \varphi(Z_k)$$

Projection pursuit

Harder since φ replaced by nonparametric function.

Similar problems

- Interpretation — but do you care?
- Estimation in presence of collinearity

Alternative "Model": Boosting

Classification problem

- Response Y_i is discrete, here 0/1
- Predict using CART or similar tree-based model

Boosting

- Build simple classifier $\hat{Y}(1)$, a "stump"
- Define sample weights, putting more weight at points previously missclassified

$$W_{i1} = w(Y_i - \hat{Y}_{i1})$$

- For $m = 2, \ldots, M$, iterate
  1. Compute weighted classfier $\hat{Y}(m)$ using W_{m-1}
  2. Compute new weights W_m
- Combine classifications by weighted majority vote
  - More weight given to better classifiers

Success?

- Avoids over-fitting in many applications
- Shown to be similar to additive logistic regression (Friedman, Hastie, Tibshirani 1998)
- Model averaging, as in Bayes methods

Do you need new models?

1998 KDD Cup

- 21 groups compete to build best model.
- Includes SAS, IBM, specialized vendors
- http://www.kdnuggets.com/

Direct mail problem

- Paralyzed Veterans of America
- List of 200,000 former donors
- Results of prior mail campaign

Could you have done better?

Modeling data

- 100,000 cases available to the modeler
- 500 *raw* predictors (prior donation, demographics)

Costs in hold-back sample

- 5% respond, with net of $74,000
- Mail to all lowers net to $10,600
- Half of data miners beat full mailing cost
- Other half are less profitable

KDD Winning Approach

Winning analysis

View as two questions, a two-stage model:

- Who will respond: logistic regression
- How much will they give: regression

Net gain

Winner nets $14,700, versus $10,600 for mailing to all.

Feature selection

Search over expanded set of 2,000 features, with many substantive additions.

Familiar trade-off

Maximize *profit*, not prediction accuracy.

$$E(\text{profit}) = \text{pr}(\text{donate}) \times (\text{donation \$})$$

$$-(1 - \text{pr}(\text{donate})) \times (\text{contact \$})$$

Gains chart

Identifying Predictors

Feature selection is key

Various models share the common problem of choosing predictors from a very large set of candidates.

Recent progress in statistics

Competitive analysis: How well can you compete with others who have more information about an underlying 'true' model.

Oracle analysis

Oracle knows which predictors to include, and must only estimate their coefficients:

- What data-based rule minimizes *ratio* of MSEs?

$$\min_{\hat{q}} \max_{\beta} \frac{E \left\| Y - \hat{Y}(\hat{q}) \right\|^2}{q\sigma^2}$$

- Answer: "hard thresholding" (Donoho&Johnstone, Foster&George 94)

$$\text{Pick } X_j \text{ whose } |t_j| > \sqrt{2 \log p}$$

- Almost Bonferroni (harsher than Bonferroni)

Adaptive Variable Selection

Sources of prediction error

- Include an extraneous predictor
- Omit a useful predictor $\Leftarrow$
- Random estimation error

Hard thresholding omits useful predictors.

Step up/step down tests

- Bonferroni threshold unpopular in *multiple comparisons* because of low power.
- Simes method – step-up/step-down tests:

$$|t_{(1)}| \geq |t_{(2)}| \geq \cdots \geq |t_{(p)}|$$

1. Compare $t_{(1)}$ to $\sqrt{2 \log p}$
2. Compare $t_{(2)}$ to $\sqrt{2 \log p/2}$
3. ... compare $t_{(q)}$ to $\sqrt{2 \log p/q}$

- Related methods:
  emp Bayes, information theory, half-normal plots

Prediction error

Within a constant factor of prediction error of "expert" who knows the true β_j, but not the coordinates!

Step 3. Validation

Over-fitting

With feature selection from among so many choices...

- Tukey: "Optimization capitalizes on chance"
- Selection bias
- Coincidences

Dilemma for cross-validation

If you use

- Most of the data to identify predictors, then you have little left for validation.
- Little of the data to identify predictors, then you have poor model to validate.

Will this problem go away with enough data?
(recall bankruptcy case – only 2500 events)

Baseball model

- Large training data set (40,000)
- "Small" validation data set (1,000)
- Model pretty clear
- Intentionally bias results

Validation Samples

"Zero Law of Data Mining"

The more similar the training data is to the prediction data, the more accurate the predictions will be.

Comparable?

- Osteoporosis:
  - Validation gathered in separate study.
  - Different demographics
- Time series:
  - Are the time periods really comparable?
  - Creative ways to cross-validate time series.

Convincing MBAs

Convey weaknesses How to convey the weaknesses of data mining in the context of a class that introduces them to regression modeling?

MBA regression project

- 20-30 predictors, concern for interaction
- *Very* competitive, like big R^2
- Discover stepwise regression, basic data mining
- Knowledge spreads

Question

How to convey problems of stepwise and, more generally, models discovered by data mining?

Possible solutions

- Prohibition
- Math
- Example...

Stepwise Regression Example

Predict stock returns

- Monthly returns on S&P 500 index
- Predict next year using previous five
- We "give" them a collection of 10-20 predictors

Stepwise regression

- Standard implementation
  - Easy forward selection
  - Harsh rule for backward elimination
- Response surface option
  - All pairwise interactions
  - Includes main effects

Results

- Fit is good on paper
- $R^2 \approx 0.85$
- About 10 predictors pass Bonferroni

Pictures...

What happened?

Over-fitting

- Fits well in-sample, poor out of sample
- Predictions worse than fitting constant

What predictors were those?

- Random noise
- Choices from big database

Message sometimes garbled

"What's your random number generator?"

How to use stepwise productively

- Use only as a "polishing step"
- Don't include with interpretation
- Standard protection fails (Bonferroni)
  - Biased estimate of σ^2
  - Find an unbiased estimate

DM and Statistics

Where can you use data mining examples?

Statistics can play a role in every phase of data mining

- Data selection, preparation
- Modeling
- Validation

Data preparation

- Over-sampling rare events, stratified
- Reliability of internet-gathered samples
- Experiments vs observational study

Modeling

- Regression remains canonical method
- Over-fitting
- Awareness of Bonferroni ideas
- Multiple comparisons in Anova
- Random clustering

Validation

- Cross-validation
- Prediction *intervals*, ranges

Relating to Other Courses

Data mining also suggests connections to other courses in the standard business program...

Information systems

- Data base systems
- Information systems, DSS

Legal, ethics

- Privacy concerns
  - Double-click cookies
  - Amazon grouping
- International laws, differences

Management

- Managing the data mining process
- Ownership, responsibility, goals
- Strategy
  - Most valuable resource of many is data.
  - Capital One: test-and-learn strategy

Marketing

- E-commerce marketing
- Real-time testing, feedback

Some Lessons for Us

If ther're so similar, why has business become fascinated by Data Mining when Statistics has been around for a while?

Economics of models

- DMs explicitly address the financial trade-off.
- Dollars, not goodness of fit (R^2), is last word.
  - Bankruptcy, credit scoring
  - Risk of osteoporosis
  - CD Now cross selling
  - Direct mailing

Size of problems

- Statistics ignored their problems, data
- Data mining embraced computers, computer science

Better names

supervised learning	classification
gains or lift chart	receiver operating characteristic
neural network	projection pursuit regression

Looking Ahead

"Data mining" will not go away

Data mining also represents the trend for other disciplines to develop and offer tools for data analysis, leveraging technology.

Participate

Statisticians need to be engaged, rather than saying "That's not statistics."

Show off the value of statistical thinking. Efron: "Those who ignore Statistics are comdemned to reinvent it."

If you teach it, will they come?

- Lure to interest MBAs in statistics
  - Very popular modeling course
- Wharton Executive Education
  - They were not looking for a "statistics" course