

Lecture I. Introduction & Exploratory Analysis

What is Data Mining?

Algorithms

- Automatic search for models
- Alternative types of models
- Criteria for picking models
- Finding variables for a regression
- Networks, recursive tree
- AIC, BIC, cross-validation

Huge databases

- Vast numbers of observations, variables
- “Wide”
- Unstructured

“Machine learning”

- Part of computer science
- Rapidly progressing, exploiting pace of processing speeds
- Moving on, building upon methods developed hundreds of years ago

Different objectives

- Predictive models rather than “cause and effect”

Discussion

Criticisms

- Heuristic, empirical hocus-pocus
- No clear model statement
- Unsuitable for inference, raw empiricism
- Practical failures: works well for your data, but not for new data
- Business hype, poor match for social science

Response

- Progress comes slowly in statistics
 - Better match to how data analysis is done in practice
 - Lessons for the way we look at data
 - Diagnostic for missing structure
-

Syllabus

Outline of lectures

- Exploratory analysis for multivariate data (today)
- Regression analysis
- Souped up regression analysis
- Logistic regression
- Neural networks
- Trees

Software

- JMP from SAS
- R

Bibliography

Expectations

- Backgrounds of participants
 - Regression
 - Perhaps some multivariate methods, perhaps
- Only way to learn a subject is to do it: Computer labs, TA
- Documentation project

Data

- ICPSR data 2004, 2008 ANES election studies
- Other data

Data: Election Survey

2004 ANES

- Need a question to focus analysis: How well can we predict
 - Choices of a registered voter? (vote?, Bush?) categorical response
 - Whether someone “likes” a candidate? “numerical” response
- Different types of data
 - Two phases of interviews
 - Many other columns
- Implications: get out the vote, predict missing data

Exploratory Data Analysis

What is EDA?

- Visual, lots of graphics
- Looking for patterns
- Gain familiarity with data

Browsing ANES 2004

- Square data table: 1,200 rows and columns
- Reliability of data: relevant in predictive modeling?
- Continuous and categorical data
- Feeling thermometers
- Need a codebook!

Marginal distributions, contingency tables

- Prevalence of missing data
- Response variables are close to national percentages

Continuous variables

- Role for principal components
- Cluster analysis

JMP tools

- Analyze menu
- Distribution
- Fit Y by X (two variables)
- Multivariate (principal components , create more vars)
- Nominal, ordinal, continuous variables (affect JMP's analysis of data)
- Formulas, calculator