COM 578
Empirical Methods in Machine Learning and Data Mining

Rich Caruana

http://www.cs.cornell.edu/Courses/cs578/2005fa

Today

- Dull organizational stuff
  - Course Summary
  - Grading
  - Office hours
  - Homework
  - Final Project
- Fun stuff
  - Historical Perspective on Statistics, Machine Learning, and Data Mining

Staff, Office Hours, …

Rich Caruana Tue 4:30-5:00pm caruana@cs.cornell.edu	Upson Hall 4157 Wed 10:00-11:00am
TA: Cristian Bucila Thu 11:30-12:00 cristi@cs.cornell.edu	Upson Hall 322 Fri 2:00-3:00
TA: Lars Backstrom TBA lbackstrom@cs.cornell.edu	Upson Hall ????
TA: Alex Niculescu-Mizil TBA alexn@cs.cornell.edu	Upson Hall 5154
Admin: Amy Fish M-F 9:00am-4:00pm	Upson Hall 4146

Topics

- Decision Trees
- K-Nearest Neighbor
- Artificial Neural Nets
- Support Vector Machines
- Association Rules
- Clustering
- Boosting/Bagging
- Cross Validation
- Performance Metrics
- Data Transformation
- Feature Selection
- Missing Values
- Case Studies:
  - Medical prediction
  - Protein folding
  - Autonomous vehicle navigation

~30% overlap with CS478

Grading

- 4 credit course
- 25% take-home mid-term (mid-October)
- 25% open-book final (Fri Dec 9, 9:00-11:30)
- 30% homework assignments (3 assignments)
- 20% course project (teams of 1-4 people)
- late penalty: one letter grade per day
- 90-100 = A-, A, A+
- 80-90 = B-, B, B+
- 70-80 = C-, C, C+

Homeworks

- short programming and experiment assignments
  - e.g., implement backprop and test on a dataset
  - goal: get familiar with a variety of learning methods
- two or more weeks to complete each assignment
- C, C++, Java, Perl, shell scripts, or Matlab
- must be done individually
- hand in code with summary and analysis of results
- emphasis on understanding and analysis of results, not generating a pretty report
- short course in Unix and writing shell scripts

Project

- Data Mining Mini Competition
- Train best model on problem(s) we give you
  - decision trees
  - k-nearest neighbor
  - artificial neural nets
  - SVMs
  - bagging, boosting, model averaging, ...
- Given train and test sets
  - Have target values on train set
  - No target values on test set
  - Send us predictions and we calculate performance
  - Performance on test sets is part of project grade
- Due before exams & study period: Fri, December 2

Text Books

- **Required Text:**
  - ***Machine Learning*** by Tom Mitchell
- **Optional Texts:**
  - ***Elements of Statistical Learning: Data Mining, Inference, and Prediction*** by Hastie, Tibshirani, and Friedman
  - ***Pattern Classification***, 2nd ed., by Richard Duda, Peter Hart, & David Stork
  - ***Data Mining: Concepts and Techniques*** by Jiawei Han and Micheline Kamber
- Selected papers

Fun Stuff

Statistics, Machine Learning, and Data Mining

Past, Present, and Future

Once upon a time...

before statistics

Pre-Statistics: Ptolmey-1850

- First "Data Sets" created
  - Positions of mars in orbit: Tycho Brahe (1546-1601)
  - Star catalogs
    - Tycho catalog had 777 stars with 1-2 arcmin precision
  - Messier catalog (100+ "dim fuzzies" that look like comets)
  - Triangulation of meridian in France
- Not just raw data - processing is part of data
  - Tychonic System: anti-Copernican, many epicycles
- No theory of errors - human judgment
  - Kepler knew Tycho's data was never in error by 8 arcmin
- Few models of data - just learning about modeling
  - Kepler's Breakthrough: Copernican model and 3 laws of orbits

Pre-Statistics: 1790-1850

- The Metric System:
  - uniform system of weights and measures
- Meridian from Dunkirk to Barcelona through Paris
  - triangulation
- Meter = Distance (pole to equator)/10,000,000
- Most accurate survey made at that time
- 1000's of measurements spanning 10-20 years!
- Data is available in a 3-volume book that analyses it
- No theory of error:
  - surveyors use judgment to "correct data" for better consistency and accuracy!

Statistics: 1850-1950

- Data collection starts to separate from analysis
- Hand-collected data sets
  - Physics, Astronomy, Agriculture, ...
  - Quality control in manufacturing
  - Many hours to collect/process each data point
- Usually Small: 1 to 1000 data points
- Low dimension: 1 to 10 variables
- Exist only on paper (sometimes in text books)
- Experts get to know data inside out
- Data is clean: human has looked at each point

Statistics: 1850-1950

- Calculations done manually
  - manual decision making during analysis
  - Mendel's genetics
  - human calculator pools for "larger" problems
- Simplified models of data to ease computation
  - Gaussian, Poisson, …
  - Keep computations tractable
- Get the most out of precious data
  - careful examination of assumptions
  - outliers examined individually

Statistics: 1850-1950

- Analysis of errors in measurements
- What is most efficient estimator of some value?
- How much error in that estimate?
- Hypothesis testing:
  - is this mean larger than that mean?
  - are these two populations different?
- Regression:
  - what is the value of y when $x=x_i$ or $x=x_j$?
- How often does some event occur?
  - $p(fail(part_1)) = p_1$; $p(fail(part_2)) = p_2$; $p(crash(plane)) = ?$

Statistics would look very different if it had been born after the computer instead of 100 years before the computer

Statistics meets Computers

Machine Learning: 1950-2000...

- Medium size data sets become available
  - 100 to 100,000 records
  - Higher dimension: 5 to 250 dimensions (more if vision)
  - Fit in memory
- Exist in computer, usually not on paper
- Too large for humans to read and fully understand
- Data not clean
  - Missing values, errors, outliers,
  - Many attribute types: boolean, continuous, nominal, discrete, ordinal
  - Humans can't afford to understand/fix each point

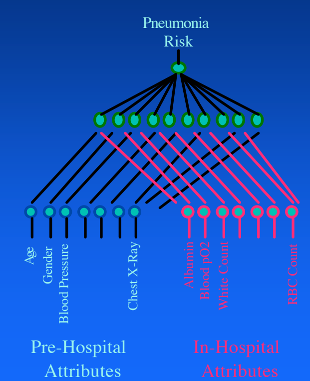
Machine Learning: 1950-2000...

- Computers can do ***very*** complex calculations on medium size data sets
- Models can be much more complex than before
- Empirical evaluation methods instead of theory
  - don't calculate expected error, measure it from sample
  - cross validation
  - e.g., 95% confidence interval from data, not Gaussian model
- Fewer statistical assumptions about data
- Make machine learning as automatic as possible
- Don't know right model => OK to have multiple models (vote them)

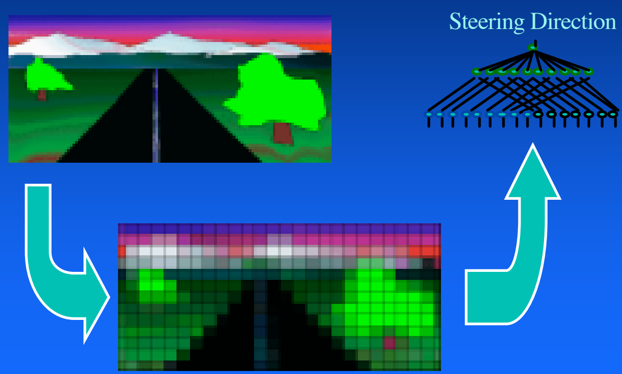
Machine Learning: 1950-2000...

- Regression
- Multivariate Adaptive Regression Splines (MARS)
- Linear perceptron
- Artificial neural nets
- Decision trees
- K-nearest neighbor
- Support Vector Machines (SVMs)
- Ensemble Methods: Bagging and Boosting
- Clustering

ML: Pneumonia Risk Prediction



ML: Autonomous Vehicle Navigation



Can't yet buy cars that drive themselves, and few hospitals use artificial neural nets yet to make critical decisions about patients.

Machine Learning: 1950-2000...

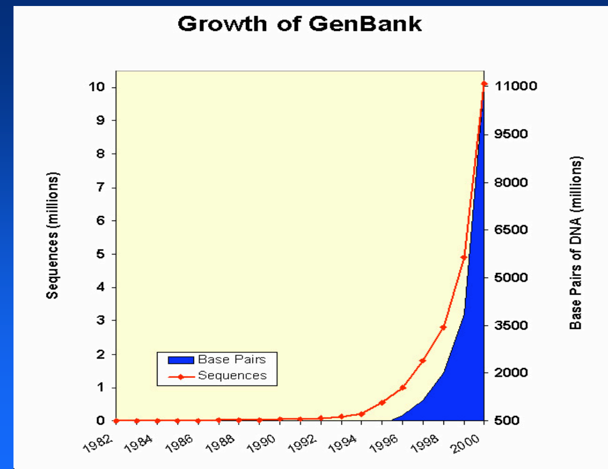
- New Problems:
  - Can't understand many of the models
  - Less opportunity for human expertise in process
  - Good performance in lab doesn't necessarily mean good performance in practice
  - Brittle systems, work well on typical cases but often break on rare cases
  - Can't handle heterogeneous data sources

Machine Learning Leaves the Lab

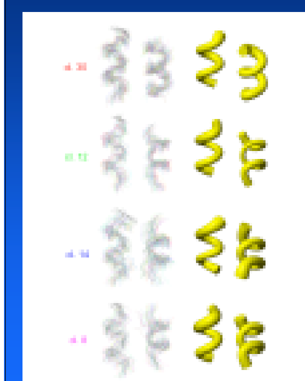
Computers get Bigger/Faster
but
Data gets Bigger/Faster, too

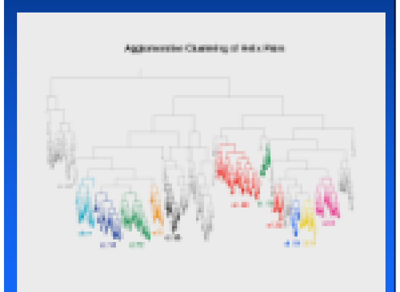
Data Mining: 1995-20??

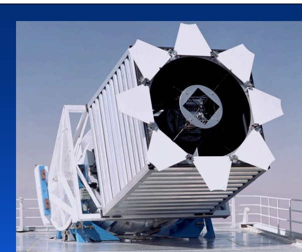
- Huge data sets collected fully automatically
  - large scale science: genomics, space probes, satellites
  - Cornell's Arecibo Radio Telescope Project:
    - terabytes per day
    - petabytes over life of project
    - too much data to move over internet -- they use FedEx!

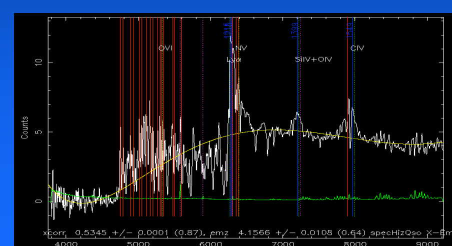


Protein Folding









What is the Sloan Digital Sky Survey?

Simply put, the Sloan Digital Sky Survey is the most ambitious astronomical survey project ever undertaken. The survey will map in detail one-quarter of the entire sky, determining the positions and absolute brightnesses of more than 100 million celestial objects. It will also measure the distances to more than a million galaxies and quasars. Apache Point Observatory, site of the SDSS telescopes, is operated by the Astrophysical Research Consortium (ARC).

Data Mining: 1995-20??

- Huge data sets collected fully automatically
  - large scale science: genomics, space probes, satellites
  - consumer purchase data
  - web: > 500,000,000 pages of text
  - clickstream data (Yahoo!: gigabytes per hour)
  - many heterogeneous data sources
- High dimensional data
  - "low" of 45 attributes in astronomy
  - 100's to 1000's of attributes common
  - linkage makes many 1000's of attributes possible

Data Mining: 1995-20??

- Data exists only on disk (can't fit in memory)
- Experts can't see even modest samples of data
- Calculations done completely automatically
  - large computers
  - efficient (often simplified) algorithms
  - human intervention difficult
- Models of data
  - complex models possible
  - but complex models may not be affordable (Google)
- Get something useful out of massive, opaque data
  - data "tombs"

Data Mining: 1990-20??

- What customers will respond best to this coupon?
- Who is it safe to give a loan to?
- What products do consumers purchase in sets?
- What is the best pricing strategy for products?
- Are there unusual stars/galaxies in this data?
- Do patients with gene X respond to treatment Y?
- What job posting best matches this employee?
- How do proteins fold?

Data Mining: 1995-20??

- New Problems:
  - Data too big
  - Algorithms must be simplified and very efficient (linear in size of data if possible, one scan is best!)
  - Reams of output too large for humans to comprehend
  - Very messy uncleaned data
  - Garbage in, garbage out
  - Heterogeneous data sources
  - Ill-posed questions
  - Privacy

Statistics, Machine Learning, and Data Mining

- Historic revolution and refocusing of statistics
- Statistics, Machine Learning, and Data Mining merging into a new multi-faceted field
- Old lessons and methods still apply, but are used in new ways to do new things
- Those who don't learn the past will be forced to reinvent it
- => Computational Statistics, ML, DM, …

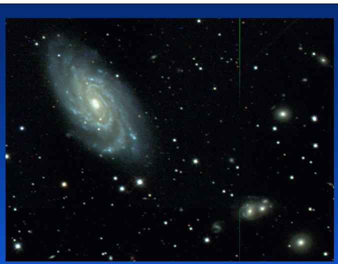
Change in Scientific Methodology

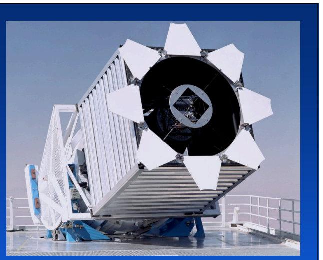
Traditional:

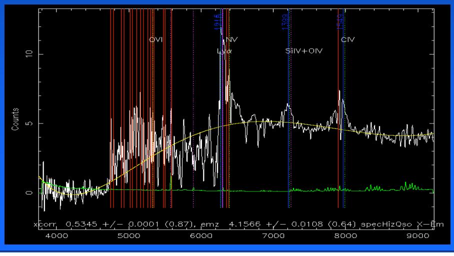
- Formulate hypothesis
- Design experiment
- Collect data
- Analyze results
- Review hypothesis
- Repeat/Publish

New:

- Design large experiment
- Collect large data
- Put data in large database
- Formulate hypothesis
- Evaluate hyp on database
- Run limited experiments to drive nail in coffin
- Review hypothesis
- Repeat/Publish







ML/DM Here to Stay

- Will infiltrate all areas of science, engineering, public policy, marketing, economics, …
- Adaptive methods as part of engineering process
  - Engineering from simulation
  - Wright brothers on steroids!
- But we can't manually verify models are right!
- Can we trust results of automatic learning/mining?