CS 6220: Data Mining Techniques

Mirek Riedewald

Course Information

- Homepage: http://www.ccs.neu.edu/home/mirek/classes/2011-S-CS6220/
  - Announcements
  - Homework assignments
  - Lecture handouts
  - Office hours
- Prerequisites: CS 5800 or CS 7800, or consent of instructor
  - No exception for first-year Master's students—based on past experience

Grading

- Homework: 40%
- Midterm exam: 30%
- Final exam: 30%
- No copying or sharing of homework solutions allowed!
  - But you can discuss general challenges and ideas with others
- Material allowed for exams
  - Any handwritten notes (originals, no photocopies)
  - Printouts of lecture summaries distributed by instructor
  - Nothing else

Instructor Information

- Instructor: Mirek Riedewald (332 WVH)
  - Office hours: Tue 4:30-5:30pm, Thu 11am-noon
  - Can email me your questions (include TA)
  - Email for appointment if you cannot make it during office hours (or stop by for 1-minute questions)
- TA: Peter Golbus (472 WVH)
  - Office hours: TBD

Course Materials

- No single textbook covers everything at the right level of depth and breadth…
- Main textbook: Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques, 2nd edition, Morgan Kaufmann, 2006
  - Read it as we cover the material in class
- Other resources mentioned in syllabus
  - Consult them whenever the textbook is not sufficient

Course Content and Objectives

- Become familiar with landmark general-purpose data mining methods and understand the main ideas behind each of them
  - Classification and prediction: decision tree, regression tree, Naïve Bayes, Bayesian Belief Network, rule-based classification, artificial neural network, SVM, nearest neighbor
  - Ensemble methods: bagging, boosting
  - Frequent pattern mining: frequent itemsets, frequent sequences
  - Clustering: K-means, hierarchical, density-based, high-dimensional data, outliers

Course Content and Objectives

- Learn about major concepts and challenges in data mining and how to deal with them
  - Overfitting
  - Bias-variance tradeoff
  - Evaluating the quality of a classifier or predictor
  - Exponential search space and pruning for pattern mining
  - Pattern interestingness measures
  - Intuitive clustering versus clusters found by an algorithm; evaluating a clustering
- Gain practical experience in using some of these techniques on real data
  - Choose the right method for the task and tune it for the given data

What We Cannot Cover

- Specialized techniques
  - Text mining, genome analysis, recommender systems
- All possible variations of the presented landmark techniques
- Volumes of theoretical and technical results for most techniques
- Implementation details

How to Succeed

- Attend the lectures
  - Advanced material, not readily found in any single textbook
- Take notes during the lecture
  - Helps remembering (compared to just listening)
  - Capture lecture content more individually than our handouts
  - Free preparation for exams
- Go over notes, handouts, book chapter soon after lecture, e.g., Wed or Thu
  - Try to explain material to yourself or friend
    - Reveals if you really understood it
    - Helps identify questions early—ask us ASAP to resolve them
- Look at content from previous lecture right before the next lecture to "page-in the context"

How to Succeed

- Ask questions during the lecture
  - There is no "stupid" question
  - Even seemingly simple questions show that you are thinking about the material and are genuinely interested in understanding it
    - Helps you stay alert and makes instructor happy…
- Work on the HW assignment as soon as it comes out
  - Time to ask questions and deal with unforeseen problems
  - We might not be able to answer all last-minute questions if there are too many right before the deadline

What Else to Expect?

- Need good programming skills
  - E.g., be able to write algorithm that recursively partitions a set of multi-dimensional data points based on their values in different dimensions
- Tree structures and how to traverse them
  - Recursion!
- Cost of algorithms in big-O notation
- Fairly basic probability concepts
  - Random variable, joint distribution, expectation, variance, confidence interval, conditional probability, independence
- Basic logic and set operators
  - AND, OR, implication, union, intersection

Introduction

- Motivation: Why data mining?
- What is data mining?
- Real-world example

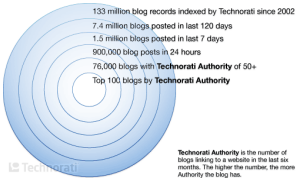
How Much Information?

- Source: http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/execsum.htm
- 5 exabytes (10^{18}) of new information from print, film, optical storage in 2002
  - 37,000 times Library of Congress book collections (17M books)
- New information on paper, film, magnetic and optical media doubled between 2000 and 2003
- Information that flows through electronic channels—telephone, radio, TV, Internet—contained 18 exabytes of new information in 2002

13

Web 2.0

- Billions of Web pages, social networks with millions of users, millions of blogs
  - How do friends affect my reviews, purchases, choice of friends
  - How does information spread?
  - What are "friendship patterns"
    - Small-world phenomenon: any two individuals likely to be connected through short sequence of acquaintances



14

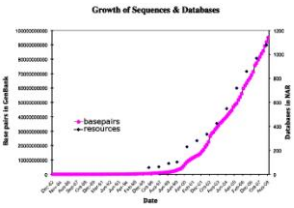
eScience

- …science and engineering data are constantly being collected, created, deposited, accessed, analyzed and expanded in the pursuit of new knowledge. In the future, U.S. international leadership in science and engineering will increasingly depend upon our ability to leverage this reservoir of scientific data captured in digital form, and to transform these data into information and knowledge aided by sophisticated data mining, integration, analysis and visualization tools. (National Science Foundation Cyberinfrastructure Council, 2007)
- Computing has started to change how science is done, enabling new scientific advances through enabling new kinds of experiments. These experiments are also generating new kinds of data – of increasingly exponential complexity and volume. Achieving the goal of being able to use, exploit and share these data most effectively is a huge challenge. (Towards 2020 Science, Report by Microsoft Research, 2006)

15

eScience Examples

- Genome data
- Large Hadron Collider
  - Petabytes of raw data
  - Find particles
  - Analyze scientific analysis process itself
    - How do experienced researchers attack a problem?
- SkyServer
  - 818 GB, 3.4 billion rows
- Cornell Lab of Ornithology
  - 80M observations, thousands of attributes



Source: Nature

16

Other Examples

- Fraudulent/criminal transactions in bank accounts, credit cards, phone calls
  - Billions of transactions, real-time detection
- Retail stores
  - What products are people buying together?
  - What promotions will be most effective?
- Marketing
  - Which ads should be placed for which keyword query?
  - What are the key groups of customers and what defines each group?
- How much to charge an individual for car insurance
- Spam filtering

17

Traditional Analysis: Regression

- Simple linear regression: $Y = a + b*X$
- Estimate parameters a and b from the data
  - Need to deal with noise, errors
  - Given a set of data points (X_i, Y_i), find a and b that minimize squared error, i.e., the sum of $(Y_i - (a + b*X_i))^2$
- Solution exists
- Can also compute confidence intervals

18

Problems?

- More complex functions: $Y = F(X_1, X_2, \ldots, X_{1000})$
  - What functional form to choose?
    - Does Y depend on X or log(X) or X^k? What k to choose?
    - Which variables interact and how?
      - Include term X_1X_2 or $X_1\log(X_2)$ or other?
- How to efficiently search the space of possible functions
- How to select one functional form over another
- Wrong assumptions => bad predictions
  - How do we know we found a good function?
- Expert-driven process, difficult to **scale** to large complex data

19

What Is Data Mining?

- Statistics + computers
- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from a huge amount of data
- Alternative names:
  - Knowledge discovery in databases, knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

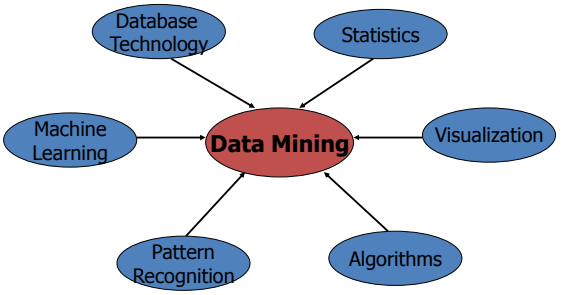
20

Data Mining Process

- Learning the application domain
  - Relevant prior knowledge and goals of application
- Creating a target data set: data selection
- Data cleaning and preprocessing (may take 60% of effort!)
- Data reduction and transformation
  - Find useful features, dimensionality reduction
- Choosing functions of data mining
  - Summarization, classification, regression, association, clustering
- Choosing the mining algorithm(s)
- Data mining: search for patterns of interest by tuning the algorithm
- Pattern evaluation and knowledge presentation
  - Visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

21

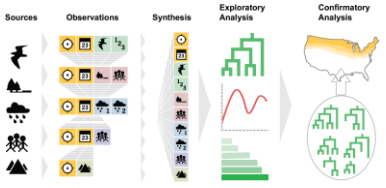
Data Mining: Confluence of Multiple Disciplines



22

Real-World Example

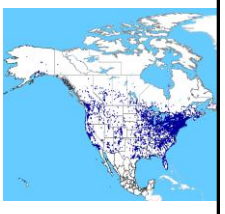
- Millions of bird observations
  - Which bird, when, where, observer effort
- Joined with data about climate, habitat, human population, geography,...
- Are bird species declining? Why? What are major migration patterns? Are they changing over the years?



23

Working with Observational Data

- Errors
- Missing values
- Skewed distribution
- Outliers



- Made-up example to illustrate problems caused by skew and interactions:
  - 100 birds live in an area in 2009, but only 60 in 2010
  - On sunny days, observer sees 80% of the birds
  - On rainy days, observer sees 10% of the birds
  - 2009: 30% observations on sunny days, 70% on rainy days
  - 2010: 70% observations on sunny days, 30% on rainy days
  - Average number of birds seen in 2009: 0.3*0.8*100 + 0.7*0.1*100 = 31
  - Average number of birds seen in 2010: 0.7*0.8*60 + 0.3*0.1*60 = 35.4
  - Conclusion: number of birds increased from 2009 to 2010... Wrong!

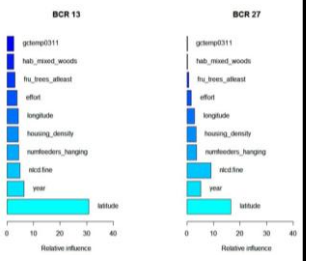
24

Our Approach

- Train a model
  - Function that maps a combination of input values (climate, habitat etc.) to a desired output (e.g., number of birds seen)
- If the models shows good generalization performance, use it as an approximation of reality
  - Use a model that can represent interactions
  - Analyze the model (= pseudo-experiment) to deal with skew
  - Estimate confidence

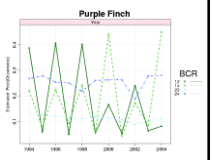
Most Important Attributes

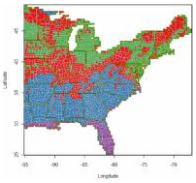
- Feature selection: train new model for each subset of attributes
  - Very expensive
- Sensitivity analysis: shuffle attribute values, run through model
  - Expensive
- Analyze model structure, e.g., records affected by split predicate in tree
  - Cheap



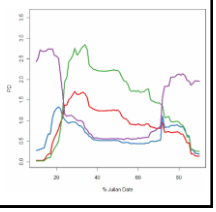
Model Summaries

- Model: #birds = F(time, location, observer effort, climate, weather, habitat, human population)
- Summaries
  - #birds = F_1(year)
  - #birds = F_2(day of year) for given regions





Tree Swallow migration



Our Scolopax Project

- Search for patterns in models based on user preferences

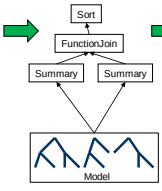
Make this as easy and fast as Web search

User-friendly query language (broad class of patterns)

Formal language (for query optimization)

Optimizer (execution in distributed system)

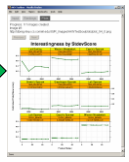
Pattern evaluation



Pattern ranking alg.

Function join alg.

Pattern creation alg. (low cost, parallel, approximate)



Data mining models (distributed training, evaluation, confidence computation)