CMSC828G Principles of Data Mining Lecture #1

- Today's Reading:
  - HMS, chapter 1
  - How Much Information Is There In the World? Michael Lesk, 1997
  - Data Mining and Statistics: What's the Connection? Jerome Freidman, 1997
- Today's Lecture:
  - Logistics
  - Introduction to Data Mining
  - Project overview
- Due Today:
  - H0: Course Survey – Please fill out and give to the TA before you leave!

Logistics

- Web page: http://www.cs.umd.edu/class/spring2002/cmsc828g/
- **Time and Place: Tu, Th 11:00 - 12:15 AVW 3258**
- **Instructor:**
  Lise Getoor
  AVW 3205
  getoor@cs.umd.edu
  office hours: Tu 3:30-5:30PM and by appointment
- **Teaching Assistant:**
  Eiman Elnahrawy
  AVW 3228
  eiman@cs.umd.edu
  office hours: Tu, Th 1:30-2:30PM and by appointment
- Course Mailing list:
  - cmsc828Gg@cs.umd.edu

  Please subscribe to the mailing list today:
  Mail majordomo@cs.umd.edu
  In the body of your message put 'subscribe cmsc828g'

Logistics cont.

- Workload:
  - Readings:
    - Text, Principles of Data Mining, Hand, Mannila and Smyth (HMS)
    - Additional readings listed for each class
  - Homeworks:
    - 4 Homeworks; each home work will have a written component and a matlab exercise. You will have approximately 2 weeks for each homework. **(15%)**
    - Homeworks are due at the ***start*** of class.
    - You may work together. However, if you must turn in your own copy of the homework. Please list anyone you collaborated with.
    - ***Late policy: 25% reduction first 24 hours. 50% reduction 24-48 hours. No homework accepted after 2 days. Please turn in late homework to the TA.***
  - Exams:
    - Midterm: in class, tentative date Thu 3/21 **(20%)**
    - Final: 10:30-12:30, Saturday May 18 **(30%)**
  - Project:
    - More about this at the end of the class… **(35%)**

Logistics cont.

- **Prerequisites:**
  - AI (cmsc421), DB (cmsc424), probability and statistics
- **Other Useful courses:**
  - Machine Learning (cmsc726),
  - Adv. DB (cmsc724),
  - Information Retrieval (cmsc838L),
  - Information Visualization (cmsc828F,cmsc838B)

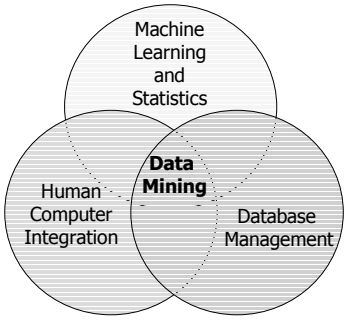
Approximate Syllabus

- http://www.cs.umd.edu/class/spring2002/cmsc828g/syllabus.html
- Topics covered:
  - Measurement and data, HMS ch.2
  - Information Visualization, HMS ch.3
  - Statistical foundations, HMS ch.4
  - Data mining components, HMS 5,6,7
  - Descriptive Modeling, HMS ch.9
    - Density estimation
    - Cluster analysis
  - Predictive Modeling for classification, HMS ch.10
  - Predictive Modeling for regression, HMS ch.11
  - Causal Models
  - Data Organization, HMS ch.12
  - Patterns and rules, HMS ch.13
  - Advanced topics, as time permits: text retrieval, recommender systems, link analysis, mining data streams, mining biomedical data, temporal and spatial data.

Introduction

- What is data mining?
- Why now?
- Knowledge Discovery in Databases
- Data mining process
- Data mining tasks

Data Mining



Definition

The analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.

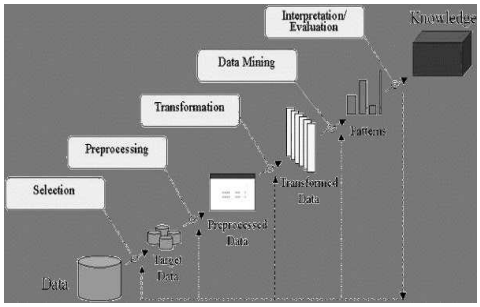
Principles of Data Mining, Hand, Mannila and Smyth

Why Now?

- Huge increase in digital data acquisition
- units of very large storage sizes:
  - gigabyte 1,000 megabytes
  - terabyte 1,000 gigabytes
  - petabyte 1,000 terabytes
  - exabyte 1,000 petabytes
- How much information is there in the world? Lesk, 1997

  There may be a few thousand petabytes of information all told; and the production of tape and disk will reach that level by the year 2000. So in only a few years,
  - we will be able save ***everything*** -- no information will have to be thrown out, and
  - the typical piece of information will ***never*** be looked at by a human being.

Knowledge Discovery in Databases (KDD)



Reference: Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview", in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, *Advances in Knowledge Discovery and Data Mining,* AAAI Press / The MIT Press, Menlo Park, CA, 1996, pp.1-34

KDD cont.

The overall process of finding and interpreting patterns from data involves the repeated application of the following steps:

1. Developing an understanding of
   - the application domain
   - the relevant prior knowledge
   - the goals of the end-user
2. Creating a target data set: selecting a data set, or focusing on a subset of variables, or data samples, on which discovery is to be performed.
3. Data cleaning and preprocessing.
   - Removal of noise or outliers.
   - Collecting necessary information to model or account for noise.
   - Strategies for handling missing data fields.
   - Accounting for time sequence information and known changes.
4. Data reduction and projection.
   - Finding useful features to represent the data depending on the goal of the task.
   - Using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.

KDD cont.

5. Choosing the data mining task.
   - Deciding whether the goal of the KDD process is classification, regression, clustering, etc.
6. Choosing the data mining algorithm(s).
   - Deciding which models and parameters may be appropriate.
   - Matching a particular data mining method with the overall criteria of the KDD process.
   - Selecting method(s) to be used for searching for patterns in the data.
7. Data mining.
   - Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.
8. Interpreting mined patterns.
9. Consolidating discovered knowledge.

Data Mining Process

- Determining the nature and structure of the representation to be used;
- Deciding how to quantify and compare how well different representations fit the data (score function)
- Choosing an algorithmic process to optimize the score function; and
- Deciding what principles of data management are required to implement the algorithms efficiently.

In this course, we will study each of these steps in depth

Example: Regression

- X is predictor variable
- Y is response variable
- Build a predictive model relating X to Y of the form:

$$Y = aX + b$$

- Predict person's annual credit-card spending given income
  - Score function?
  - Optimization algorithm?
  - Complexity?

Regression Applet

What's different?

- Emphasis on both:
  - Statistical issues
  - Computational issues
- Questions to ask about data mining algorithms:
  - Are they based on sound statistical principles?
  - How will the algorithms scale?

Data sets

- Set of measurements taken from some environment or process.
- n objects, each with same set of p measurements, n x p matrix
- Measurements may be:
  - quantitative, or numeric
    - Age, Income
  - categorical
    - Sex, Marital Status, Education
- structure: time-series data, spatial data
- but, most often we will represent our data sets as n x p matrix.

How can we do this for text documents? transaction data? web logs?

Data Mining Tasks

- Exploratory Data Analysis (EDA)
  - goal: explore data, without clear idea of what we are looking for
  - techniques: interactive and visual
  - as p increases, harder to visualize higher dimensions
    - project, in some intelligent way, onto lower dimensional space
  - as n increases, harder to visualize large number of individuals
    - abstract, in some intelligent way, into collections

Florence Nightingale's coxcomb plots (1)(2)(3)– a coxcomb plot is like a pie chart; rather than the angle of the wedges varying, the radii of the wedges differ

Data Mining Tasks cont.

- Descriptive Modeling
  - goal: describe all of the data or the process that generated the data
  - techniques: density estimation, cluster analysis and segmentation

Market segmentation – divide customers into homogeneous groups based on purchasing patterns and demographic data
VALS Types

Data Mining Tasks cont.

- Predictive Modeling, classification and regression
  - goal: build model that will predict the value of one variable from known values of other variables
  - classification the variable is categorical
  - regression the variable is quantitative
  - techniques: a large number of machine learning and statistical methods have been developed.

SKICAT Fayyad, Djorgovski and Weir (1996) – classify stars and galaxies from a 40-dimensional feature vector using classification tree. Performs as well as human experts.

Data Mining Tasks cont.

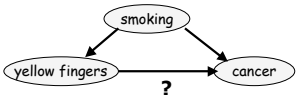
- Discovering Patterns and Rules
  - goal: detect patterns
  - spot fraudulent behavior, detect unusual stars, find items that occur frequently together in transaction databases
  - techniques: association rules
  - challenge: determining what constitutes truly unusual behavior

Professional basketball have detailed logs of every game, with time-stamped records of who took a particular type of shot, who scored, who passed to whom and so on. Advanced Scout system Bhandari et al. (1997) searches for rule-like patterns. Example: When Player X is on the floor, Player Y's shot accuracy decreases from 75% to 30%.

Data Mining Tasks cont.

- Learning Causal Models
  - goal: learn cause and effect
  - techniques: causal graphs
  - challenge: distinguishing between correlation and causation
  - example: Do yellow fingers cause lung cancer?

hidden cause: smoking



Data Mining: the downside

- Hype
- The cons of interdisciplinary fields
- Dredging, snooping and fishing
- historically, 'data mining' was a derogatory term in the statistics community
  - making inferences from small sample
    - spurious correlations such as 'over the past 30 years when the winner of the Super Bowl is from a particular league, a leading stock market index goes up in the following months'

Class Project

http://www.cs.umd.edu/class/spring2002/cmsc828g/project.htm

Next Time

- Reading:
  - HMS, chapter 2
- Topic:
  - Measurement and Data
- Due Today:
  - H0: Course Survey – Please fill out and give to the TA before you leave!
  - Please subscribe to the mailing list

References

- *Principles of Data Mining*, Hand, Mannila, Smyth. MIT Press, 2001.
- Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview", in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, AAAI Press / The MIT Press, Menlo Park, CA, 1996, pp.1-34
- How Much Information Is There In the World? Michael Lesk, 1997
- Data Mining and Statistics: What's the Connection? Jerome Freidman, 1997