CMSC828G Principles of Data Mining Lecture #26

- Today's Lecture:
  - What we covered
  - What we didn't cover
  - What next?
- Due Today:
  - Final Project Reports
  - Course Evaluations
- Final Exam: Saturday May 18, 10:30AM –12:30PM

## 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.

## 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

## Data Mining Tasks cont.

- Descriptive Modeling
  - goal: describe all of the data or the process that generated the data
  - techniques: density estimation, cluster analysis (partition-based, hierarchical, probabilistic model-based)
  - algorithms: Bayesian networks, k-means, hierarchical agglomerative clustering, mixture models

## 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: decision trees, nearest neighbor, naïve bayes, linear regression, neural networks, SVMs ….

## 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
  - algorithms: apriori, FP trees, …

## Component View

- Data mining algorithm components:
  - task
  - structure or model
  - score function
  - search or optimization method
  - data management technique

## 3 Algorithms

| | CART | Backpropagation | A Priori |
|---|---|---|---|
| Task | classification and regression | regression | rule pattern discovery |
| Structure | decision tree | neural network | association rules |
| Score Function | cross-validated loss function | squared error | support/accuracy |
| Search Method | greedy search over structure | gradient descent on parameters | breadth-first with pruning |
| Data Management Technique | unspecified | unspecified | linear scans |

## What we didn't cover, ☹

- Causal Models
- Privacy, Security issues
- Spatial and Temporal data mining
- Visual data mining, Image-based mining
- Graph mining

## What next?

- In next few weeks, concentrated graph mining reading group. Please email me if you are interested.
- In the fall, data mining reading group. Please email me if you are interested.
- Many related courses: Databases, Information Retrieval, Semantic Web. With high probability, I will teach machine learning in Spring 2003.
- Thanks, and good luck on the final exam!