Machine Learning and Data Mining

Fundamentals, robotics, recognition

Machine Learning, Data Mining, Knowledge Discovery in Data Bases

Their mutual relations

Data Mining, Knowledge Discovery in Databases, Machine Learning, & All That Stuff

- What do these terms mean?
- **Data Mining:**
  - Application of algorithms to search for patterns and relationships that may exist in large *databases*.
  - Because data sets are so large, many relationships are possible. To search this space of possibilities, machine learning techniques are used.
  - Part of the overall knowledge discovery process
  - “correct” use of term data mining is that ***it is part of process*** concerned with finding patterns in data.
  - In industry, data mining is often used for the whole process

- **Knowledge Discovery in Databases** (KDD)
  - Identifying non-trivial, valid, useful patterns in data.
  - What is a “**pattern**”?
    - Says something about the probability distribution of variables
    - For example, credit card fraud:
      - Variables may be: amount spent in one day, items purchased, where items where purchased, etc.
      - Prob(fraud | lives in Newark and 5 color TV’s bought in NYC in 2 hours) = HIGH!
  - What is data (?what are data?):
    - For the purposes of data mining, ***database***
      - rows represent a record or instance
      - columns describe attributes of that instance.
      - like in constructive induction, decision trees, rough sets, etc.

- Knowledge discovery usually involves “domain expert” and data mining analyst.
  - Other steps include gathering, preparing, storing, and accessing data. Also, ultimately understanding the patterns discovered.
- Steps in **Knowledge discovery**: (from George Johns)
  - Understand and define problem
    - So that you don’t mindless apply inappropriate algorithms and get *worthless (or misleading!) results*
  - Extract data
    - Data usually already exists in database, but not in the form you need!
    - Also, you need to decide what data you need, this is where **expert** comes in handy
  - **Understand and clean data**
    - Make sure you have consistent data
    - Make sure data is not too consistent. (Is what you are trying to predict one of the variables you know?)

- “Data engineering” (perhaps part of understand & clean)
  - Deal with missing variables, rescale data, combine similar attributes
- Algorithm engineering (this is the machine learning part)
  - Figure out what algorithm to use or write one
- Run data mining algorithm
  - art and science in this.
  - Use part of data for training, part for testing
  - Adjust parameters
- Preliminary evaluation of results
- Refine data and problem
- Use the results to accomplish some goal.

- Machine Learning:
  - This is the algorithm part of the data mining process
  - But machine learning means more:
    - “Computer program that improves its performance at some task through experience.” - Tom Mitchell (1997 - *Machine Learning*)
    - “A learning system uses sample data to generate an updated basis for improved [performance] on subsequent data from the same source and expresses the new basis in intelligible symbolic form.”

      - Donald Michie (1991 - *Computer Journal*)
    - “Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the same task or tasks drawn from the same population more effectively the next time.”
      - Herbert Simon (1983 - *Machine Learning I*)

      Last two are from David W. Aha’s tutorial on Machine Learning (http://www.aic.nrl.navy.mil/~aha/

What is learning according to Webster?

- To gain knowledge or understanding of or skill in by study, instruction, or experience
- To come to be able, to come to realize
- Robot learning - getting a robot to do all this good stuff

Challenges Specific to Robot Learning

- Situatedness in the world
  - Noise, imperfection, occlusion, dynamics, hard to model, etc.
- Real time constraints
  - Slow learners die easily in real world
- Supervised/Unsupervised mix
  - Often supervised means a bad teacher
- Simultaneous and multimodal
  - Need to walk and talk (not just either or)

- How does this relate to data mining?
- What we will discuss is really machine learning algorithms
- Learning task is data driven.
  - For example in checkers learning problem, many board positions are generated. Data is position of pieces and final state of game
    (win, lose, draw)
- Learning reduces to searching a large space of hypotheses to find one that best fits observed data.
  - Find patterns or relationships that can be generalized
  - Really the same problem as data mining.
- Machine learning algorithms deal with how this search is done.

- How does this differ from what people have been doing for years?
  - Scientific method:
    - Hypothesis -> Design Experiment -> Gather Data -> Test hypothesis
  - Data mining is more “data centric”
    - Data may have been gathered as part of an overall process or may be just an “accidental” by-product.
  - Examples:
    - Data from discount cards at supermarket
    - Credit card information
    - Historical chemical process control data
    - Information on patients gathered by hospitals
    - Visual and chemical information collected by a robot in field

- Another important difference:
  - Not only can specific relationships be obtained, but also real "knowledge" may be deduced by a computer program
    - From observed positions of planets, could *Newton's law of gravitation* be deduced automatically?
    - **Causal inference:** Can causal relationships be inferred from statistical data?
  - In fact, data mining techniques (Autoclass) have been use to learn new, useful information from *astronomical data.*
  - Under certain assumptions, causation can be deduced from statistical data.
  - This is a different type of discovery than determining correctness of model by *data fitting*.

- Let’s get more specific -

  Some common examples of data mining:

  - **Credit worthiness:** Should a bank make a loan to someone?
    - Data consists of financial **attributes**
      - Income
      - Debt
      - Credit history
      - Length of time in job
      - Residence

      Each attribute a set of values (discrete or continuous)
    - You also have a set of **instances** (cases)
      - These may be used for training (supervised learning)

        (Sometimes you look for patterns, no training - unsupervised learning)
    - Ultimate goal is to predict the likelihood of default based on a new customer’s attributes.

- Marketing:
  - Use past buying habits to predict likelihood of customer purchasing some new product
- Health care
  - Look for causal relationships between environment and disease
- Investment
  - Predict the stock market?
- Textual datamining
  - internet
- Bioinformatics
  - image processing
- Astronomy
- Chemistry
- Speech recognition
- Machine learning methods applied to signal and image analysis
- **Robotics**
  - **speech, sensor arrays, images, sensor integration,prediction, recognition**

- Most data mining methods involve classification and clustering
  - Given a collection of observed attributes (possibly a very high dimensional space), can the value of some target attribute be predicted?
  - Classification: Supervised learning
    - A set of data (instances, attribute values) with known outcome (class) is used to train algorithm
    - New data is classified based upon trained algorithm
  - Clustering: Unsupervised learning
    - Patterns are discovered in data based upon is there significant clustering of cases such that assignment to classes can be made?

- Methods:
  - Regression (linear, non-linear, multiple variables)
  - Decision Trees
  - Neural networks
  - Bayesian classifiers
  - Baysian networks
  - Constructive Induction (last lecture)
  - Genetic algorithms
- Many resources available on Internet
  - A few to start with
    - http://www.kdnuggets.com/
    - http://www.ai.univie.ac.at/oefai/ml/ml-resources.html
    - http://www.cs.cmu.edu/~tom/
  - Assignment:
    - Download “Enhancements to the Data Mining Process” George Johns’ thesis (now a book as well)
    - Read Chapter 1 “What is Data Mining?

Questions and Problems

- What is your understanding of relations between Machine Learning, Data Mining and Knowledge Discovery in Data Bases
- Think about interesting data base that can be mined and propose the complete data mining method based on methods that you learned in this class.
- How data mining can be used in Intelligent Robotics?
- Propose a robot that will use Knowledge Discovery methods
- Propose an Internet Robot with Data Mining abilities