Machine Learning Intro

CPS 170

Ron Parr

Why Study Learning?

- Considered a hallmark of intelligence
- Viewed as way to reduce programming burden
- Many algorithms assume parameters that are difficult to determine exactly a priori

Examples

- SPAM classification
- Computational Biology/medicine
  - Distinguish healthy/diseased tissue (e.g., skin/colon cancer)
  - Find structure in biological data (regulatory pathways)
- Financial events
  - Predict good/bad credit risks
  - Predict price changes
  - Response to marketing
- Drilling sites likely to have oil
- Document categorization
- Learn to play games
- Learn to control systems
  - Fly Helicopter
  - Optimize OS components
- Public database of learning problems:
  - http://www.ics.uci.edu/~mlearn/MLSummary.html

Who Does Machine Learning?

- In AI
  - Core AI topic (AAAI, IJCAI)
  - Specialized communities (ICML, NIPS)
- Databases (data mining - KDD)
- Used in (CS):
  - Vision
  - Systems
  - Comp. Bio
- Statistics

Who Does Machine Learning (@Duke)

- CS:
  - Faculty: Pankaj Agarwal, Vince Conitzer, Alex Hartemink, Kamesh Munagala, Ron Parr, Carlo Tomasi, Jun Yang
- ISDS (everybody, but especially):
  - Scott Schmidler, Sayan Mukherjee
- IGSP:
  - Terry Furey, Uwe Ohler
- Engineering:
  - Larry Carin, Silvia Ferrari, Rebecca Willett

Who Hires in Machine Learning?

- Universities
- Microsoft Research
- Search: Google/Yahoo/Amazon
- Defense contractors
- Some financial institutions (quietly)
- Many startups

- ML viewed as good background for many other tasks (robotics, vision, systems, engineering)

What is Machine Learning

- Learning Element
  - The thing that learns
- Performance Element
  - Objective measure of progress
- Learning is simply an increase in the ability of the learning element over time to achieve the task specified by the performance element

ML vs. Statistics?

- Machine learning is:
  - Younger
  - More empirical
  - More algorithmic
  - (arguably) More practical
  - (arguably) More decision theoretic
- Statistics is:
  - More mature
  - (arguably) More formal and rigorous

ML vs. Data Mining

- Machine Learning is:
  - (Arguably) more formal
  - (Arguably) more task driven/decision theoretic
- Data Mining is:
  - More constrained by size of data set
  - More closely tied to database techniques

Types of Learning

- Inductive Learning
  - Acquiring new information that previously was not available
  - Learning concepts
- Speedup learning
  - Learning to do something you already "know" faster or better

Feedback in Learning

- Supervised Learning
  - Given examples of correct behavior
- Unsupervised Learning
  - No external notion of what is correct
  - Is this well-defined?
- Reinforcement Learning
  - Indirect indication of effectiveness

Learning Methodology

- Distinction between training and testing is crucial
- Correct performance on training set is just memorization!
- Researcher should ***never*** look at the test data
- Raises some troubling issues for "benchmark" learning problems

Computational Learning Theory

- Formal study of what can be learned from data
- Closely related to ML, but also to CS theory

- Assumptions:
  - Training examples must be representative
  - Algorithm needn't *always* work, but should scale well

- Goals:
  - Algorithms that have a low error rate with high probability
  - Good characterization of how performance scales

COLT

- Learning theory is elegant and mathematically rich. However,
  - It sometimes isn't constructive
  - It sometimes tells us how many data are needed, but not how to manipulate the data efficiently

- Through the late 90's, learning theory drifted away from practical learning algorithms

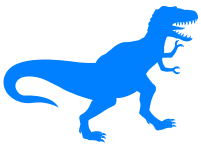
- New advances fresh thinking have led to a rapprochement, e.g.:
  - Support vector machines
  - Boosting

Example: Supervised Learning

- Classical framework
- Target concept, e.g., green
- Learner is presented with labeled instances
  - True: Green cones, green cubes, green spheres
  - False: Red cones, red cubes, red spheres, blue cones, blue cubes, blue spheres
- Learner must correctly identify the target concept from the training data

Performance Measure

- Training set won't have all possible objects
- Test set will contain novel objects
  - Blue cylinders, yellow tetrahedra
- To learn successfully, learner must have good performance when confronted w/novel objects
  - This is what we would expect from people
  - A blue Broccolisaurus is still blue



Why Learning Is Tricky

- Suppose we have seen:
  - Red tetrahedron(f), Blue sphere(t), Blue cone (t), green cube(f)
- Possible concepts:
  - Blue
  - (Blue Sphere) or (Blue Cone)
  - Objects a prime number from start
  - Objects with a circular cross-section
- What if some data are mislabeled?

Learning and Representation

- Learning is very sensitive to representation
- Every learning algorithm can be viewed as a *search through a space of concepts*
- Space of concepts determines
  - Difficulty of task
  - Appropriate algorithm
  - Restricting too aggressively can trivialize problem
  - Failure to restrict (or regularize) can trivialize the problem
- Example Space: Conjunctions of colors and shapes
  - Eliminates primes and (possibly) cross sections

Management of the Hypothesis Space

- Ockham's Razor:
  - All things being equal, favor the simplest consistent hypothesis
  - Guiding principle of science,e.g., Einstein:
    'In my opinion the theory here is the logically simplest relativistic field theory that is at all possible. But this does not mean that nature might not obey a more complex theory. More complex theories have frequently been proposed… In my view, such more complicated systems and their combinations should be considered only if there exist physical-empirical reasons to do so.'
- Ockham's razor is not provably correct, but
  - Computational learning theory shows us that the more choices we have, the more data we need to distinguish reliably among these choices
  - Well known trade off between bias and variance
    - How many points do you need to fit a degree 2 polynomial?
    - How many points do you need to fit a degree 100 polynomial?
- Ockham's razor is embodied in a wide range of methods

Learning Intro Final Thoughts

- Machine learning is one of the most successful areas of AI
  - Many practical applications
  - Many ways to succeed without solving the "whole problem"
  - Many fields view machine learning as a special sauce that will give them an advantage
- Machine learning conferences are almost as large as the general AI conferences

How to Succeed with Machine Learning

- Theoretical/algorithmic success
  - Maneuver through space of hypotheses efficiently
  - Efficiency
    - Make good use of data
    - Make good use of time
- Practical Success
  - Getting something to learn can be hard (my job!)
  - Know your problem!
    - Pick training data carefully
    - Craft hypothesis space