Lecture 14

Midterm Review

Tuesday, October 12, 1999

William H. Hsu
Department of Computing and Information Sciences, KSU
http://www.cis.ksu.edu/~bhsu

Readings:
Chapters 1-7, Mitchell
Chapters 14-15, 18, Russell and Norvig

Lecture 0: A Brief Overview of Machine Learning

- Overview: Topics, Applications, Motivation
- Learning = Improving with Experience at Some Task
  - Improve over task *T,*
  - with respect to performance measure *P*,
  - based on experience *E.*
- Brief Tour of Machine Learning
  - A case study
  - A taxonomy of learning
  - Intelligent systems engineering: *specification of learning problems*
- Issues in Machine Learning
  - Design choices
  - The performance element: intelligent systems
- Some Applications of Learning
  - Database mining, reasoning (inference/decision support), acting
  - Industrial usage of intelligent systems

Lecture 1: Concept Learning and Version Spaces

- Concept Learning as Search through *H*
  - Hypothesis space *H* as a state space
  - Learning: finding the correct hypothesis
- General-to-Specific Ordering over *H*
  - Partially-ordered set: Less-Specific-Than (More-General-Than) relation
  - Upper and lower bounds in *H*
- Version Space Candidate Elimination Algorithm
  - *S* and *G* boundaries characterize learner's uncertainty
  - Version space can be used to make predictions over unseen cases
- Learner Can Generate Useful Queries
- Next Lecture: When and Why Are Inductive Leaps Possible?

Lecture 2: Inductive Bias and PAC Learning

- Inductive Leaps Possible Only if Learner Is Biased
  - Futility of learning without bias
  - Strength of inductive bias: proportional to restrictions on hypotheses
- Modeling Inductive Learners with Equivalent Deductive Systems
  - Representing inductive learning as theorem proving
  - Equivalent learning and inference problems
- Syntactic Restrictions
  - Example: *m-of-n* concept
- Views of Learning and Strategies
  - Removing uncertainty ("data compression")
  - Role of knowledge
- Introduction to Computational Learning Theory (COLT)
  - Things COLT attempts to measure
  - Probably-Approximately-Correct (PAC) learning framework
- Next: Occam's Razor, VC Dimension, and Error Bounds

Lecture 3: PAC, VC-Dimension, and Mistake Bounds

- COLT: Framework Analyzing Learning Environments
  - Sample complexity of *C* (what is *m*?)
  - Computational complexity of *L*
  - Required expressive power of *H*
  - Error and confidence bounds (PAC: $0 < \varepsilon < 1/2$, $0 < \delta < 1/2$)
- What PAC *Prescribes*
  - Whether to try to learn *C* with a known *H*
  - Whether to try to reformulate *H* (apply change of representation)
- Vapnik-Chervonenkis (VC) Dimension
  - A formal measure of the complexity of *H* (besides $|H|$)
  - Based on *X* and a worst-case labeling game
- Mistake Bounds
  - How many could *L* incur?
  - Another way to measure the cost of learning
- Next: Decision Trees

Lecture 4: Decision Trees

- Decision Trees (DTs)
  - Can be boolean ($c(x) \in \{+, -\}$) or range over multiple classes
  - When to use DT-based models
- Generic Algorithm *Build-DT*: Top Down Induction
  - Calculating best attribute upon which to split
  - Recursive partitioning
- Entropy and Information Gain
  - Goal: to measure *uncertainty removed* by splitting on a candidate attribute *A*
    - Calculating information gain (change in entropy)
    - Using information gain in construction of tree
  - *ID3* ≡ *Build-DT* using *Gain*(•)
- ID3 as Hypothesis Space Search (in State Space of Decision Trees)
- Heuristic Search and Inductive Bias
- Data Mining using *MLC++* (Machine Learning Library in C++)
- Next: More Biases (Occam's Razor); Managing DT Induction

Lecture 5: DTs, Occam's Razor, and Overfitting

- Occam's Razor and Decision Trees
  - Preference biases versus language biases
  - Two issues regarding Occam algorithms
    - Why prefer smaller trees? (less chance of "coincidence")
    - Is Occam's Razor well defined? (*yes*, under certain assumptions)
  - MDL principle and Occam's Razor: more to come
- Overfitting
  - Problem: fitting training data too closely
    - General definition of overfitting
    - Why it happens
  - Overfitting prevention, avoidance, and recovery techniques
- Other Ways to Make Decision Tree Induction More Robust
- Next: Perceptrons, Neural Nets (Multi-Layer Perceptrons), Winnow

Lecture 6: Perceptrons and Winnow

- Neural Networks: Parallel, Distributed Processing Systems
  - Biological and artificial (ANN) types
  - Perceptron (LTU, LTG): model neuron
- Single-Layer Networks
  - Variety of update rules
    - Multiplicative (Hebbian, Winnow), additive (gradient: Perceptron, Delta Rule)
    - Batch versus incremental mode
  - Various convergence and efficiency conditions
  - Other ways to learn linear functions
    - Linear programming (general-purpose)
    - Probabilistic classifiers (some assumptions)
- Advantages and Disadvantages
  - "Disadvantage" (tradeoff): simple and restrictive
  - "Advantage": perform well on many realistic problems (e.g., some text learning)
- Next: Multi-Layer Perceptrons, Backpropagation, ANN Applications

Lecture 7: MLPs and Backpropagation

- Multi-Layer ANNs
  - Focused on feedforward MLPs
  - Backpropagation of error: distributes penalty (loss) function throughout network
  - Gradient learning: takes derivative of error surface with respect to weights
    - Error is based on difference between desired output (t) and actual output (o)
    - Actual output (o) is based on activation function
    - Must take partial derivative of $\sigma \Rightarrow$ choose one that is easy to differentiate
    - Two σ definitions: sigmoid (*aka* logistic) and hyperbolic tangent (*tanh*)
- Overfitting in ANNs
  - Prevention: attribute subset selection
  - Avoidance: cross-validation, weight decay
- ANN Applications: Face Recognition, Text-to-Speech
- Open Problems
- Recurrent ANNs: Can Express Temporal Depth (Non-Markovity)
- Next: Statistical Foundations and Evaluation, Bayesian Learning Intro

Lecture 8: Statistical Evaluation of Hypotheses

- Statistical Evaluation Methods for Learning: Three Questions
  - Generalization quality
    - *How well does observed accuracy estimate generalization accuracy?*
    - Estimation bias and variance
    - Confidence intervals
  - Comparing generalization quality
    - *How certain are we that h_1 is better than h_2?*
    - Confidence intervals for paired tests
  - Learning and statistical evaluation
    - *What is the best way to make the most of limited data?*
    - *k*-fold CV
- Tradeoffs: Bias versus Variance
- Next: Sections 6.1-6.5, Mitchell (Bayes's Theorem; ML; MAP)

Lecture 9: Bayes's Theorem, MAP, MLE

- Introduction to Bayesian Learning
  - Framework: using probabilistic criteria to search *H*
  - Probability foundations
    - Definitions: subjectivist, objectivist; Bayesian, frequentist, logicist
    - Kolmogorov axioms
- Bayes's Theorem
  - Definition of conditional (posterior) probability
  - Product rule
- Maximum *A Posteriori* (MAP) and Maximum Likelihood (ML) Hypotheses
  - Bayes's Rule and MAP
  - Uniform priors: allow use of MLE to generate MAP hypotheses
  - Relation to version spaces, candidate elimination
- Next: 6.6-6.10, Mitchell; Chapter 14-15, Russell and Norvig; Roth
  - More Bayesian learning: MDL, BOC, Gibbs, Simple (Naïve) Bayes
  - Learning over text

Lecture 10: Bayesian Classfiers: MDL, BOC, and Gibbs

- Minimum Description Length (MDL) Revisited
  - Bayesian Information Criterion (BIC): justification for Occam's Razor
- Bayes Optimal Classifier (BOC)
  - Using BOC as a "gold standard"
- Gibbs Classifier
  - Ratio bound
- Simple (Naïve) Bayes
  - Rationale for assumption; pitfalls
- Practical Inference using MDL, BOC, Gibbs, Naïve Bayes
  - MCMC methods (Gibbs sampling)
  - Glossary: http://www.media.mit.edu/~tpminka/statlearn/glossary/glossary.html
  - To learn more: http://bulky.aecom.yu.edu/users/kknuth/bse.html
- Next: Sections 6.9-6.10, Mitchell
  - More on simple (naïve) Bayes
  - Application to learning over text

Lecture 11:
Simple (Naïve) Bayes and Learning over Text

- More on Simple Bayes, *aka* Naïve Bayes
  - More examples
  - Classification: choosing between two classes; general case
  - Robust estimation of probabilities: SQ
- Learning in Natural Language Processing (NLP)
  - Learning over text: problem definitions
  - Statistical Queries (SQ) / Linear Statistical Queries (LSQ) framework
    - Oracle
    - Algorithms: search for *h* using only (L)SQs
  - Bayesian approaches to NLP
    - Issues: word sense disambiguation, part-of-speech tagging
    - Applications: spelling; reading/posting news; web search, IR, digital libraries
- Next: Section 6.11, Mitchell; Pearl and Verma
  - Read: Charniak tutorial, "Bayesian Networks without Tears"
  - Skim: Chapter 15, Russell and Norvig; Heckerman slides

Lecture 12:
Introduction to Bayesian Networks

- Graphical Models of Probability
  - Bayesian networks: introduction
    - Definition and basic principles
    - Conditional independence (causal Markovity) assumptions, tradeoffs
  - Inference and learning using Bayesian networks
    - Acquiring and applying CPTs
    - Searching the space of trees: max likelihood
    - Examples: *Sprinkler*, *Cancer*, *Forest-Fire*, generic tree learning
- CPT Learning: Gradient Algorithm *Train-BN*
- Structure Learning in Trees: MWST Algorithm *Learn-Tree-Structure*
- Reasoning under Uncertainty: Applications and Augmented Models
- Some Material From: http://robotics.Stanford.EDU/~koller
- Next: Read Heckerman Tutorial

Lecture 13:
Learning Bayesian Networks from Data

- Bayesian Networks: Quick Review on Learning, Inference
  - Learning, eliciting, applying CPTs
  - In-class exercise: *Hugin* demo; CPT elicitation, application
  - Learning BBN structure: constraint-based versus score-based approaches
  - *K2*, other scores and search algorithms
- Causal Modeling and Discovery: Learning Cause from Observations
- Incomplete Data: Learning and Inference (Expectation-Maximization)
- Tutorials on Bayesian Networks
  - Breese and Koller (AAAI '97, BBN intro): http://robotics.Stanford.EDU/~koller
  - Friedman and Goldszmidt (AAAI '98, Learning BBNs from Data): http://robotics.Stanford.EDU/people/nir/tutorial/
  - Heckerman (various UAI/IJCAI/ICML 1996-1999, Learning BBNs from Data): http://www.research.microsoft.com/~heckerman
- Next Week: BBNs Concluded; Review for Midterm (10/14/1999)
- After Midterm: More EM, Clustering, Exploratory Data Analysis

Meta-Summary

- Machine Learning Formalisms
  - Theory of computation: PAC, mistake bounds
  - Statistical, probabilistic: PAC, confidence intervals
- Machine Learning Techniques
  - Models: version space, decision tree, perceptron, winnow, ANN, BBN
  - Algorithms: candidate elimination, *ID3*, backprop, MLE, Naïve Bayes, *K2*, EM
- Midterm Study Guide
  - Know
    - Definitions (terminology)
    - How to solve problems from Homework 1 (problem set)
    - How algorithms in Homework 2 (machine problem) work
  - Practice
    - Sample exam problems (handout)
    - Example runs of algorithms in Mitchell, lecture notes
  - Don't panic! ☺