Cmput 466 / 551
Introduction to Machine Learning

R Greiner

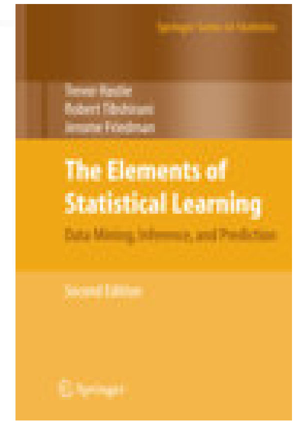
Department of Computing Science

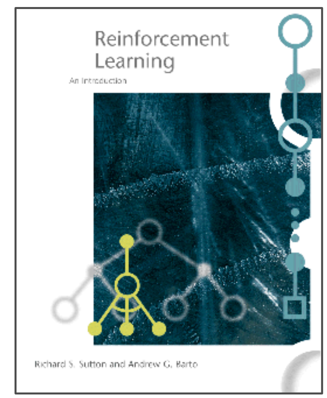
University of Alberta

http://www.cs.ualberta.ca/~greiner/C-466/

Summary

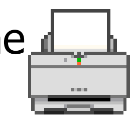
- http://www.cs.ualberta.ca/~greiner/C-466/
  - Assignments, Logistics, Slides
- REQUIRED Texts:
  - Hastie/Tibshirani/Friedman, *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, 2009. (2nd edition)
  - Barto/Sutton, *Reinforcement Learning, 1998.*
  - Other hand-outs ("Bayesian Networks")
- Evaluation:
  - Assignments: 70% (4 total; 3 with coding); Solo
    Late policy (4 "forgivable" days; ≤2 /HW)
  - Project: 30%, in ?2?3- or 4-person teams
    Must be on-time!





Contact Information

- Home Page: http://www.cs.ualberta.ca/~greiner/C-466/
  - Announcements
  - Assignments
  - Slides: http://www.cs.ualberta.ca/~greiner/C-466/SLIDES
    - Print when you see:



- Email to c466@cs.ualberta.ca reaches
  - prof
  - all 2?3 TAs
- Newsgroup: *via Moodle*
  - Public!

- Prof: *R Greiner*
  - Office hours: Tues, Thurs: 2:00-2:50, or by appointment
  - Phone: 780 492-5461
- Barnabas Poczos
  - Phone: 780 248-1435
- Contact TAs in Labs, by appointment
  - *Shahab Jabbari Arfaee*
  - *Gabor Balazs*

Mini- Research Project

- Investigate some interesting aspect of machine learning
  - a broad thorough literature review, overviewing general topic:
    Ex1: techniques for learning motifs in DNA
    Ex2: ways to cope with missing data
  - a deeper discussion of some specific subtopic
    Ex1: using HMMs to learn probabilistic motifs
    Ex2: statistically motivated ways to handle blocked attribute values
  - theoretical/empirical analysis of several systems for this task
    Ex1: empirical comparison of several gene-finding tools, on novel datasets
    Ex2: empirical, +? theoretical, analysis of several techniques, on data

- Either
  - “application pull”: seeking ways to solve some specific problem (Ex1)
  - “technology push”: exploring ways of coping with some specific technical challenge (Ex2)

Mini- Research Project (con't)

Tentative!

- ?3- or 4-person teams
- 30% of course grade
  - Everyone gets same
  - (All Grads; All UGrads)
- Schedule

  \+ bi-weekly meetings
- Presentations

Decide on topic (1page)	**6/Oct**
Presentation #1 "lay of the land"	**12,17/Nov**
Presentation#2: contributions	**1,3/Dec**
Final write-up	**17/Dec**

- http://www.cs.ualberta.ca/~greiner/C-466/project.html

Mini- Research Project (con’t)

- Report should include:
  - Problem: Why is problem interesting and challenging?
  - Background material, review/limitations of previous work
  - Technical solutions used to solve problem, successful? (why?)
  - Remaining problems; future research

  … 8 pages (NIPS style)
- Evaluation Criteria
  - Apparent effort
  - Clarity... Analysis, Examples, ...
  - Originality
  - Implementation
- Same grade (all grads; all ugrads)

Homework Issues

- Both Programming / nonProgramming questions
- Programming Questions
  - Typically C, C++, JAVA, Matlab
    - If you want another language: check with TAs
    - Your implementations must run on lab machines (CSC 219)
  - Neat, well-documented … include *convincing* examples and tests
  - The onus is on ***you*** to convince TAs that your code/idea works
  - Submit using 'ASTEP'
- NonProgramming Questions
  - Write legibly or type (better!)
  - Submit in class, or in "Box", or to ASTEP
- … don't annoy the TAs!

Assignment Guidelines

- Submit on due date/time
  - Program (ASTEP) + Hard copy (class, box)
  - Late policy: 4 “excused days”; ≤2 / HW#1
  - If >4 days: 15% penalty / day (until solution posted)
- Use MoodleGrades to see…
  - Current marks, #Late days, Class statistics
- If question about marking:
  - See TA first… then prof, only if necessary
- Don’t look for answers on the web…
- Don’t cheat… Code of Conduct

Code of Conduct

- Do not cheat on assignments:
  Discuss only general approaches to problem
- Do not take written notes on other's work
- Respect the lab environment.
- Do not:
  - Interfere with operation of computing system
  - Interfere with other's files
  - Change another's password
  - Copy another's program
  - …

- Cheating is reported to university whereupon it is out of our hands
- Possible consequences:
  - A mark of 0 for assignment
  - A mark of 0 for the course
  - A permanent note on student record
  - Suspension / Expulsion from university

Academic Integrity

The University of Alberta is committed to the highest standards of academic integrity and honesty. Students are expected to be familiar with these standards regarding academic honesty and to uphold the policies of the University in this respect. Students are particularly urged to familiarize themselves with the provisions of the Code of Student Behavior (online at www.ualberta.ca/secretariat/appeals.htm) and avoid any behavior which could potentially result in suspicions of cheating, plagiarism, misrepresentation of facts and/or participation in an offence. Academic dishonesty is a serious offence and can result in suspension or expulsion from the University.

Labs

- TAs in labs ONLY…
  - before HW is due
  - for tutorials – Matlab, Belief Nets?, …
  - by arrangement
- *ONLY* first hour
- Only Ugrads assigned Lab;
  Grads can attend if space
- 2008: used lab-sessions only
  2times/all year…

Goals of Course

- Obtain a (near)graduate-level understanding of
  Machine Learning
- Emphasis: systems that can learn about environment, to help them improve their performance on range of tasks.
- Covering…
  - ***general models:*** supervised learning, unsupervised learning, reinforcement learning + active learning
  - ***general techniques:*** gradient descent, consistency filtering, EM, …
  - ***practical aspects:*** algorithms for learning
    linear regressors, linear classifiers, SVMs, decision trees, neural networks, belief networks, HMMs,…+ mixtures, boosting, …;
  - ***theoretical concepts (foundations):*** relevant ideas from statistics, inductive bias, Bayesian learning and the PAC learning framework.

http://www.cs.ualberta.ca/~greiner/C-466/

Major Topics Covered

- Foundations
  - Probability
  - Statistics
  - Regularization & Learning Theory
- Supervised Learning
  - Linear classifiers, logistic regression, LDA, …
  - NeuralNetworks, NearestNeighbor, DecisionTrees, BayesianClassifiers, …
  - Support Vector Machines and Kernel Methods
- Unsupervised Learning
  - Dimensionality reduction: Clustering, PCA, ..
  - Learning graphical models (Belief net / MarkovNet / HMM)
- ? Reinforcement Learning ?
  - Decision Theory
  - MDPs
  - RL Algorithms

Major Topics *Not* Covered

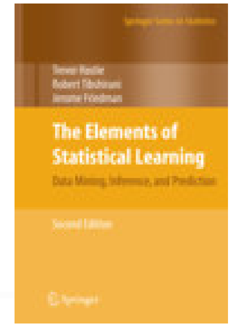
- Genetic algorithms
- Fuzzy sets
- Rough sets
- Biological basis of learning
  - Human
  - Animal models

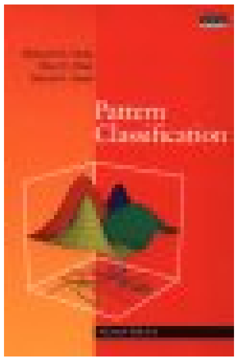
Prerequisites

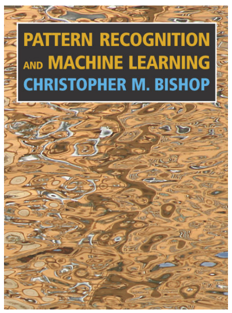
- Probabilities
  - Distributions, densities, marginalization…
- Basic statistics
  - Moments, typical distributions, regression…
- Algorithms
  - Dynamic programming, basic data structures, complexity…
- Programming
  - Mostly your choice of language, but Matlab will be very useful
- We provide some background, but the class will be fast paced

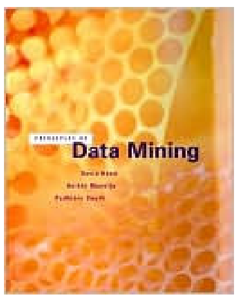
- Ability to deal with “abstract mathematical concepts”

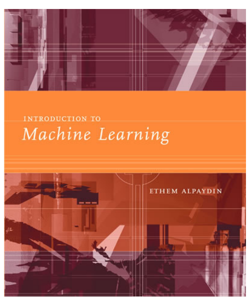
Textbooks

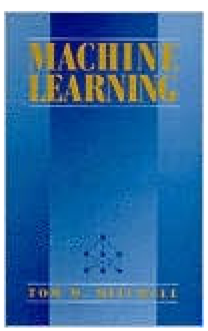


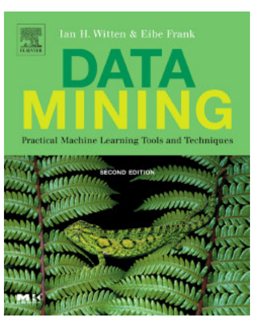




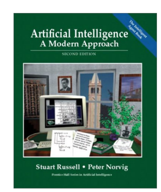






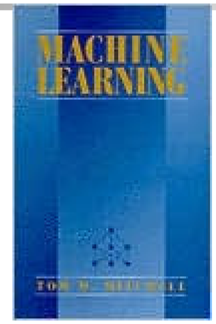


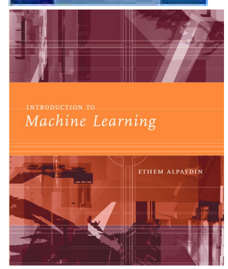
- Hastie/Tibshirani/Friedman, *The Elements of Statistical Learning*, Springer, 2009.
- Recommended
  - Duda/Hart/Stork, *Pattern Classification,* 2006*.*
  - C Bishop, *Pattern Recognition and Machine Learning,* Srpinger, 2006.
  - Hand/Mannila/Smyth, *Principles of Data Mining*, 2001.
  - Alpaydin, *Introduction to Machine Learning*, 2004.
  - Mitchell, *Machine Learning*, McGraw Hill, 1997.
  - Wittin/Frank, *Data Mining* (2nd) Morgan Kaufmann, 2005.
  - Russell/Norvig, *Artificial Intelligence: A Modern Approach*, 2003.

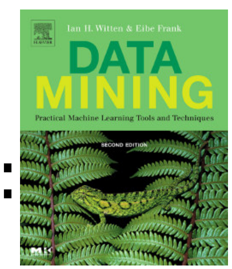


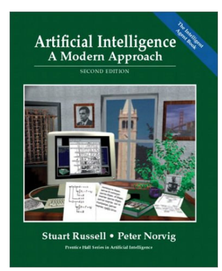
How best to *structure* ML?

- Learning Classifiers
  - Collection of Algs
    - Linear Separators
    - Decision Trees
    - SVM
    - NN, NN, …
  - Issues
    - Overfitting, …
  - Foundations
    - PAC learning
    - Bayesian theory
- Other Types of Learning:
  - Regression
  - Density Estimation
    - Graphical Models
  - Reinforcement Learning



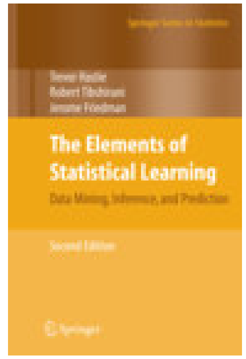


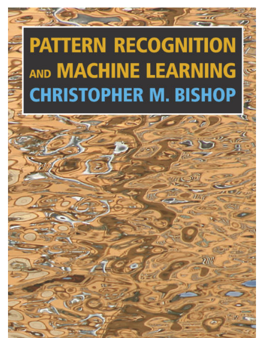


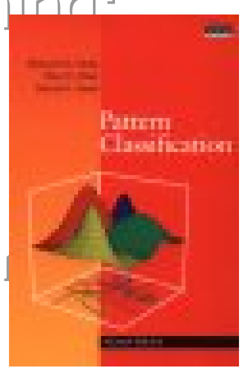


How best to *structure* ML?

- Learning Classifiers
  - Collection of Algs
    - Linear Separators
    - Decision Trees
    - SVM
    - NN, NN, ...
  - Issues
    - Overfitting, ...
  - Foundations
    - PAC learning
    - Bayesian theory
- Other Types of Learning
  - Regression
  - Density Estimation
    - Graphical Models
  - Reinforcement Learni







- Foundation of Learning
  - Classification/Regression
  - Bayesian theory
    - Gaussian, Gaussian, Gaussian, ...
    - Density Estimation (Graphical Models)
- Collection of Algorithms
  - Linear Separators
  - Decision Trees
  - SVM
  - NeuralNet, NearNghbr, ...
- Other stuff
  - PAC learning
  - Reinforcement Learning
  - ...

AI Seminar !!!

- http://www.cs.ualberta.ca/~ai/cal/
- Friday noons, CSC 3-33
- Neat topics, great speakers, FREE PIZZA!

http://www.cs.ualberta.ca/~ai/schedule.html

Other Issues

- Ask LOTS of questions
  - Really…
- Questionaire

- Should we cover Reinforcement Learning?
  - Yes: It is important part of machine learning!
    - … at least 2 lectures ??
  - No: Already covered in full semester-long course!

Class Size

- At most 10 teams
- At most 4 students/team

$\Rightarrow$ at most 40 students

How many really want to take course??

Specific Ideas

- **Big Ideas:**
  - Can be done – in practice, theory
  - $\neg\exists$ Universal knowledge-free learner!
    $\Rightarrow$ needs prior knowledge
  - Rel'n of Training Data (Size, Quality) to quality of results (Overfitting)
  - Computational complexity
- **Techniques:**
  - Specific algorithms for learning ...
    Linear Separators, Decision Trees, Neural Nets, Belief Nets, ...
  - General techniques:
    ConsistencyFiltering, Gradient Descent, EM, Reinforcement, Boosting, ...
- **Foundations, Formal theories:**
  Bayesian Theory, Hypothesis Eval'n, PAC-Learning
- **Applications:**
  - Classification/ Regression (Diagnosis), Reinforcement (Control)
  - Computational Biology, DataMining, Adaptive software (Web/Interfaces)

Syllabus

- Covers a wide range of Machine Learning techniques – from basic to state-of-the-art
- You will learn about the methods you heard about:
  - Naïve Bayes, logistic regression, nearest-neighbor, decision trees, boosting, neural nets, SVMs, HMMs, graphical models, PCA, …,
  - overfitting, regularization, dimensionality reduction, error bounds, VC dimension, kernels, margin bounds, K-means, EM, mixture models, …
  - semi-supervised learning, active learning, reinforcement learning…
- Covers algorithms, theory and applications
- **It's going to be fun and hard work ☺**