CS4780/5780 - Machine Learning

Fall 2011

Thorsten Joachims
Cornell University
Department of Computer Science

Cornell University

- Who we are?
  - Prof: Thorsten Joachims
  - TAs: Karthik Raman, Chenhao Tan, Adith Swaminathan
  - Consultants: Mevlana Gemici, Anthony Chang, Nic Williamson,
    Heran Yang, Boiar Qin
- What is learning?
  - Why should a computer be able to learn?
  - Examples of machine learning.
  - What it takes to build a learning system?
- Syllabus
- Administrivia

- Definition [Mitchell]:
  - A computer program is said to learn from
    - experience E with respect to some class of
    - tasks T and
    - performance measure P,
  - if its performance at tasks in T, as measured by P,
  - improves with experience E.

- Concept Learning : Hypothesis space, version space
- Instance-Based Learning : k-nearest neighbor, collaborative filtering
- Decision Trees : TDIDT, attribute selection, pruning and overfitting
- ML Experimentation: hypothesis tests, resampling estimates
- Linear Rules : Perceptron, duality, mistake bound
- Support Vector Machines : optimal hyperplane, kernels, stability
- Generative Models : Naïve Bayes, linear discriminant analysis
- Hidden Markov Models : probabilistic model, estimation, Viterbi
- Structured Output Prediction : predicting sequences, rankings, etc.
- Learning Theory : PAC learning, mistake bounds
- Clustering : HAC Clustering, k-means, mixture of Gaussians

- Main Textbooks
  - Main: Tom Mitchell, "Machine Learning", McGraw Hill, 1997.
  - Cristianini, Shawe-Taylor, "Introduction to Support Vector Machines", Cambridge University Press, 2000. (online)
  - Schoelkopf, Smola, "Learning with Kernels", MIT Press, 2001. (online)
  - Course pack (one chapter)
- Additional References (optional)
  - Ethem Alpaydin, "Introduction to Machine Learning", MIT Press, 2004.
  - See other references on course web page.
- Course Notes
  - Slides available on course homepage
  - Material on blackboard

- Pre-Requisites
  - Programming skills (e.g. CS 2110)
  - Basic linear algebra (e.g. MATH2940)
  - Basic probability theory (e.g. CS 2800)
  - → Short exam to test prereqs
- Related Courses
  - CS4700: Foundations of Artificial Intelligence
  - CS4758: Robot Learning
  - CS4300: Information Retrieval
  - CS6780: Advanced Machine Learning
  - CS6784: Advanced Topics in Machine Learning
  - CS6740: Advanced Language Technologies

Cornell University

- Assignments
  - 5 homework assignments
  - Some problem sets, some programming and experiments
- Policies
  - Assignments are due at the beginning of class on the due date in hardcopy. Code must be submitted via CMS by the same deadline.
  - Assignments turned in late will drop 5 points for each period of 24 hours for which the assignment is late.
  - Everybody had 3 "free" late days. Use them wisely.
  - No assignments will be accepted after the solutions have been made available (typically 3 days after deadline).
  - Typically collaboration of two students (see each assignment for detailed collaboration policy).
  - We run automatic cheating detection. Must state all sources of material used in assignments or project. Please review Cornell Academic Integrity Policy!

- In-class Quizzes
  - A few per semester
  - No longer than 5 minutes
  - You can miss one of them
- Exams
  - Two Prelim exams
  - In class
  - No final exam

- Organization
  - Self-defined topic related to your interests and research
  - Groups of 3-4 students
  - Each group has TA as advisor
- Deliverables
  - Project proposal (~ 2 weeks after fall break)
  - Meetings with TA to discuss progress
  - Short presentation in class (last week of classes)
  - Project report (~ exam period)

- Deliverables
    - 2 Prelim Exams   (40% of Grade)
    - Final Project   (15% of Grade)
    - Homeworks (~5 assignments)   (35% of Grade)
    - Quizzes (in class)   (5% of Grade)
    - PreReq Exam   (2% of Grade)
    - Participation   (3% of Grade)
- Outlier elimination
    - For homeworks and quizzes, the lowest grade is replaced by the second lowest grade.

Cornell University

- Online
  - http://www.cs.cornell.edu/Courses/cs4780/2011fa/
  - Piazza forum
  - Videonote
- Email Addresses
  - Thorsten Joachims: tj@cs.cornell.edu
  - Karthik Raman, karthik@cs.cornell.edu
  - Chenhao Tan, chenhao@cs.cornell.edu
  - Adith Swaminathan, adith@cs.cornell.edu
  - Mevlana Gemici, Anthony Chang, Nic Williamson, Heran Yang, Boiar Qin
- Office Hours
  - Thorsten Joachims:
    - Tuesdays 2:40pm – 3:30pm, 4153 Upson Hall
  - Other office hours:
    - TBD