

ComS 573 – Machine Learning

Vasant Honavar
Artificial Intelligence Research Laboratory
Department of Computer Science
226 Atanasoff Hall
Iowa State University
Ames, Iowa 50011

January 10, 2005.

1 Course Objectives

This course aims to provide an introduction to the basic principles, techniques, and applications of Machine Learning. Programming assignments are used to help clarify basic concepts. The emphasis of the course is on the fundamentals, and not on providing a mastery of specific commercially available software tools or programming environments.

In short, this course is about the principles, design and implementation of learning agents – programs that improve their performance on some set of tasks with experience. Upon successful completion of the course, you will have a broad understanding of machine learning algorithms and their use in data-driven knowledge discovery and program synthesis. You will have designed and implemented several machine learning algorithms in Java. You will also be able to identify, formulate and solve machine learning problems that arise in practical applications. You will have a knowledge of the strengths and weaknesses of different machine learning algorithms (relative to the characteristics of the application domain) and be able to adapt or combine some of the key elements of existing machine learning algorithms to design new algorithms as needed. You will have an understanding of the current state of the art in machine learning and be able to begin to conduct original research in machine learning.

2 Prerequisites

The prerequisites necessary for fully benefiting from the material covered in this course include knowledge of data structures (e.g., lists, trees), design and analysis of algorithms, programming language concepts (e.g., functional programming, object-oriented programming, recursion, abstract data types), and selected topics in mathematics (e.g., boolean algebra, set theory, probability theory, calculus). If you are not sure whether you have the necessary background, please talk to the instructor.

Students will be expected to familiarize themselves with Java on their own, with the help of online resources and lab assignments.

Laboratory assignments will require you to program in Java a modern a platform-independent language for object-oriented design and implementation of software systems in general, and dis-

tributed artificial intelligence systems in particular. If you do not know Java already, you are expected to quickly acquire a working knowledge of Java on your own. Java should be an easy language to learn for those students who have a good grasp of object-oriented programming in a modern high-level language.

3 Course Staff

The instructor for the course is Dr. Vasant Honavar (honavar@cs.iastate.edu). The teaching assistant for the course is Ms. Oksana Yakhnenko (oksayakh@cs.iastate.edu). Additional information (office hours, etc.), will be posted on the course Web page at: <http://www.cs.iastate.edu/~cs573x/>.

The instructor and the TA will be available to answer your questions during the scheduled office hours, or at a time arranged by prior appointment, and at other times if necessary (if our schedules permit). The course Web page will be used to convey or update information concerning homework assignments, post lecture outlines, etc.

We might also use electronic mail to reach you when necessary. You are therefore strongly encouraged to get into the habit of reading your electronic mail and checking the course Web page once a day.

4 Computer Accounts

We will be using the computing facilities of the Department of Computer Science for all course-related assignments. If you do not already have a login on the departmental computer systems, please make sure that you have one by the end of first week of classes.

The Computation Center and the Computer Science Department hold tutorials that are designed to help a new users to get familiar with their facilities. Please contact them for a schedule and information on signing up for one of these tutorials.

The course web page located at will be used to post course materials including assignments, study guides, etc.

As a student in a course offered through the Computer Science department and a user of the ISU computer facilities, you are to abide by the department's Code of Computer Ethics a copy of which will be provided to you. *Please note that any suspected violations of the code of ethics are viewed extremely seriously by the Computer Science department and treated in accordance with the university's policies on academic misconduct.*

5 Assignments, Examinations, and Grading

There will be regularly scheduled written assignments, assignments, examinations to help you learn the material and to help us evaluate your progress. You will also need to complete a term project.

Each student will be assigned to transcribe lecture notes for 2 lectures during the course. LaTeX templates to be used in preparing lecture notes (and when available, drafts of notes that might be useful to start with) will be supplied. The notes will be edited by the TA and the instructor before being made available to the class.

There will be two examinations — given approximately around the 8th and the 15th week of the semester respectively. The examinations may have a take-home component. Written assignments will be handed out roughly every two to three weeks. Laboratory assignments will be assigned every two or three weeks. You should expect roughly 4-6 written assignments and about 4 laboratory assignments. Some of the written assignments will require you to read, understand, and critique current research papers that are related to the topics being covered in class.

The term project has to be a small research or design project cuminating in a written report and possibly a brief oral presentation. You may choose to work in small groups (consisting of 2-3 members each) on the project.

You should be actively thinking about potential topics for projects or term papers right from the beginning of the semester. A list of suggested topics as well as guidelines for the preparation of project reports and term papers will be made available in due course. The instructor and the TA will be available for consultation and guidance on the projects or papers as needed.

The course grades will be based on transcribing lecture notes (7.5%) written assignments, (17.5%), laboratory assignments (17.5%), two examinations (17.5% each), term project (17.5%), and participation in discussions in class (5%).

6 Policy on Collaboration, Late Assignments, Etc.

The primary purpose of the assignments is to clarify and enhance the understanding of the concepts covered in the lectures. Past experience with this course has shown that this is helped by increased interaction among students. Discussion of *general concepts and questions* concerning the problem sets and laboratory assignments among students is encouraged. However, each student is expected to work on the solutions individually. Sharing of solutions (including segments of code) to assignments is forbidden unless explicitly instructed otherwise. If you are unclear about this policy, please talk to the instructor *before* you act. **Suspected cases of academic misconduct will be pursued fully in accordance with ISU policies.**

On late assignments, there is a late penalty of 5% of the grade per day up to a maximum of 7 days from the specified due date. Assignments that are turned in later than 7 days after the due date will be assigned zero credit. Rare exceptions to this policy might be made (at the discretion of the course staff) under demonstrably extenuating circumstances.

7 Syllabus

A tentative list of topics to be covered in the course (not necessarily in the order in which they will be covered) is: Algorithmic models of learning. Learning classifiers, functions, relations, grammars, probabilistic models, value functions, behaviors and programs from experience. Bayesian,

maximum a posteriori, and minimum description length frameworks. Parameter estimation, sufficient statistics, decision trees, neural networks, support vector machines, Bayesian networks, bag of words classifiers, N-gram models; Markov and Hidden Markov models, probabilistic relational models, association rules, nearest neighbor classifiers, locally weighted regression, ensemble classifiers. Computational learning theory, mistake bound analysis, sample complexity analysis, VC dimension, Occam learning, accuracy and confidence boosting. Dimensionality reduction, feature selection and visualization. Clustering, mixture models, k-means clustering, hierarchical clustering, distributional clustering. Reinforcement learning; Learning from heterogeneous, distributed, data and knowledge. Selected applications in data mining, automated knowledge acquisition, pattern recognition, program synthesis, text and language processing, and bioinformatic and computational biology.

More detailed list of topics will be posted on the course web page. Specific reading assignments and brief lecture outlines will be placed on the course homepage periodically.

8 Textbooks, Lecture Notes, and References

There is no assigned textbook for this course. Lectures will draw on a number of sources including several textbooks:

1. Mitchell, T. (1997). *Machine Learning*. New York: Mc Graw-Hill.
2. Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of Statistical Learning - Data Mining, Inference, and Prediction*. Berlin: Springer-Verlag.
3. Cowell, R.G., Dawid, A.P., Lauritzen, S.L., and Spiegelhalter, D.J. (1999). *Graphical Models and Expert Systems*. Berlin: Springer.
4. Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. New York: Oxford University Press.
5. Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. London: Cambridge University Press.
6. Kearns, M. and Vazirani, U. (1994). *Computational Learning Theory*. Cambridge, MA: MIT Press.
7. Baldi, P., Frasconi, P., Smyth, P. (2003). *Modeling the Internet and the Web - Probabilistic Methods and Algorithms*. New York: Wiley.
8. Chakrabarti, S. (2003). *Mining the Web*. Palo Alto, CA: Morgan Kaufmann.
9. Baldi, P. and Brunak, S. (2002). *Bioinformatics: A Machine Learning Approach*. Cambridge, MA: MIT Press.
10. Cohen, P.R. (1995) *Empirical Methods in Artificial Intelligence*. Cambridge, MA: MIT Press.

Specific readings that will be assigned on a weekly basis are an integral part of the course. Reading assignments and pointers to relevant materials will be posted as part of the study guide on the course web page. A number of books and references will be available on reserve in the Parks library. Additional sources of useful material include major journals and conference proceedings in these areas. You are strongly encouraged to explore various machine learning resources on the Web (Check the course Web page for pointers).