



Ph.D. Program in Machine Learning

Student Handbook

Revised 8/24/07

Introduction

The Machine Learning Department is a department within the School of Computer Science at Carnegie Mellon University. The Department Head of Machine Learning is Tom Mitchell, Fredkin Professor of Artificial Intelligence and Learning.

The Ph.D. program in Machine Learning is designed to train students to become tomorrow's leaders in the rapidly growing area of Machine Learning. The program builds on ML's world-class faculty across several academic departments. The program is designed to:

- (1) provide students with a unique multi-disciplinary curriculum taught by experts from a variety of disciplines,
- (2) expose students to the latest data mining research results developed at ML and elsewhere, and
- (3) provide practical hands-on experience with data mining problems and techniques, contributed in part by ML's corporate partners.

By exposing students to this combination of interdisciplinary coursework, hands-on applications, and cutting-edge research, we expect our graduates will be uniquely positioned to pioneer new data mining efforts, and to pursue research on the next generation of data mining tools, algorithms, and systems.

The Co-Directors of the program are:

Stephen Fienberg, Professor of Statistics and Social Sciences
Email: (fienberg@stat.cmu.edu) Phone: 412-268-2723

Geoffrey Gordon, Associate Research Professor
Email: (ggordon@cs.cmu.edu) Phone: 412-268-7399

The Machine Learning Department administrative offices are located in Wean Hall 4612, 4619 and 5315. Student offices are located in Doherty & Wean Halls. For administrative assistance, contact Diane Stidle (diane@cs.cmu.edu) x1299, Monica Hopes (meh@cs.cmu.edu) x5527 or Sharon Cavlovich (sharonw@cs.cmu.edu) x5196.

Program Information

To complete the Ph.D. degree, we require that each student

- Participate in directed research
- Pass 96 university units worth of graduate courses, with certain distribution requirements
- Serve as a teaching assistant at least twice
- Demonstrate communication and programming skills
- Write and orally defend a thesis, a significant piece of original research in a specialized area of Machine Learning.

We are committed to the principle that students may achieve competence through a variety of methods, including courses, seminars, projects, and independent study. We consider each student's individual strengths, weaknesses, and interests in designing the best method for the student to fulfill these requirements. Our program is unique in that we encourage and expect students to engage in research from their first day in the Department.

Immigration Course (IC):

The Immigration course is intended primarily to introduce you to the Machine Learning Department and the School of Computer Science. Faculty give research presentations, with time for individual and group discussions. The research presentations will give you an idea of the kind of research that goes on here in Machine Learning and throughout SCS. The Immigration course is held either immediately before or at the beginning of the Fall term. After the IC, students are strongly encouraged to arrange individual meetings with potential faculty advisors.

The Research-Matching Process in CSL

Carnegie Mellon is a research institution. We are strongly committed to scientific excellence, both in research and education. In particular, we believe that a close personal interaction among students, faculty, and staff is of the utmost importance for educating the next generation of leaders in academia and industry. CSL students are therefore matched to a faculty advisor in the very beginning of the program who will guide their research and advise them in academic matters.

THE MATCHING PROCESS:

Approximately two weeks after the end of the IC, both CSL students and Machine Learning (and affiliated) faculty submit a form, indicating preferences for advisorship relations. Based on these forms, the Co-Directors of the CSL program will then match students with faculty advisors. Each student either will

be assigned to one advisor, or will be co-advised by two faculty advisors. A student's advisor may change if the research direction changes and there is no longer a match.

ROLE OF THE ADVISOR:

The faculty advisor is a student's primary contact, both in research and in academic matters. Typically, a student has strong interests in the research area of the faculty advisor, and she/he will closely collaborate with the faculty member. The advisor is typically the primary person directing the student research, and is also expected to provide financial support (stipend and tuition) for the student.

Program Requirements:

During the IC we will be discussing the courses you should register for during the first semester. As a full-time student you are required to register for 48 units each semester. You will be conducting research with your advisor throughout the year and you must register for Research & Reading (10-920) each semester, which is a 12, 24 or 36 unit course. This course is doing research with your advisor, 50% of your time should be spent on research.

Prerequisites, Computer Science:

15-211: Fundamental Structures of Computer Science I, 12 units, Fall, Spring
Fundamental programming concepts are presented together with supporting theoretical bases and practical applications. This course emphasizes the practical application of techniques for writing and analyzing programs: data abstraction, program verification, and performance analysis. These techniques are applied in the design and analysis of fundamental algorithms and data structures. Prerequisites: 15-111 Intermediate/Advanced Programming.

15-212: Fundamental Structures of Computer Science II, 12 units, Fall, Spring
This course continues the presentation of fundamental programming concepts begun in 15-211, focusing on more sophisticated methods for describing and reasoning about computer programs. High-level languages are introduced including language mechanisms for user-defined data types, and formal methods are presented for reasoning about program specifications and correctness.
Prerequisite: 15-211.

Prerequisites, Statistics:

36-325: Probability and Mathematical Statistics, I 9 units, Fall
This course is a rigorous introduction to the mathematical theory of probability, and it provides the necessary background for the study of mathematical statistics and probability modeling. A good working knowledge of calculus is required. Topics include combinatorial analysis, conditional probability, generating functions, sampling distributions, law of large numbers, and the central limit theorem. Students studying Computer Science, or considering graduate work in Statistics or Operations Research, should carefully consider taking this course instead of 36-

225 after consultation with their advisor. Not open to students who have received credit for 36-217 or 36-225. Prerequisite: 21-118 or 21-122 or 21-256.

36-326: Probability and Mathematical Statistics II, 9 units, Spring An introduction to the mathematical theory of statistical inference. Topics include likelihood functions, estimation, confidence intervals hypothesis testing, Bayesian inference, regression, and the analysis of variance. Not open to students who have received credit for 36-226. Students studying Computer Science should carefully consider taking this course instead of 36-220 or 36-226 after consultation with their advisor. Prerequisite: 36-217 or 36-325.

Core Courses:

The core courses you must take are:

36-705(A) Intermediate Statistics, 12 units, Fall

Some elementary concepts of statistics are reviewed, and the concepts of sufficiency, likelihood, and information are introduced. Several methods of estimation, such as maximum likelihood estimation and Bayes estimation, are studied, and some approaches to comparing different estimation procedures are discussed.

10-701(A)/15-781(A): Machine Learning, 12 units, Fall

Introduction to Machine Learning, which covers a survey of basic techniques with numerous applications.

10-702(A)/ 36-712: Statistical Machine Learning 12 units, Spring

This course builds on the material presented in 10-701, introducing new learning methods and going more deeply into their statistical foundations and computational aspects. Applications and case studies from statistics and computing are used to illustrate each topic. Aspects of implementation and practice are also treated.

15-826*(A): Multimedia Databases and Data Mining, 12 units, Spring

The course covers advanced algorithms for learning, analysis, data management and visualization of large datasets. Topics include indexing for text and DNA databases, searching medical and multimedia databases by content, fundamental signal processing methods, compression, fractals in databases, data mining, privacy and security issues, rule discovery and data visualization. **15-826 is only offered every other year. 15-721 (A) Database Management Systems may be taken in place of 15-826 but there is a strict prerequisite, 15-415 Database Applications (or equivalent). 15-415 is also only offered in the spring.*

15-750 Algorithms, 12 units, Spring or 15-853 Algorithms in the Real World, 12 units, Fall

Data Analysis Project

Students are required to demonstrate their grasp of fundamental data analysis concepts and techniques in the context of a focused project. This project should involve the analysis of one or more substantive data sets and the application of state-of-the-art machine learning/data mining methods, as exemplified by, but not limited to, the concepts and methodology studied in course 10-701, 10-702, and 10-705. The project may involve the development of new methodology or extensions to existing methodology, but this is not a requirement. However, it is required that the project involves a significant data set and an analysis of the data; the project is not intended for purely theoretical or methodological investigations. Students should seek out a project (co)advisor who can provide access to such a data set, or can use a data set to which they already have access through one of the core courses or through their PhD advisor. Resources for this purpose include the Immigration Course, faculty home pages, and the ML Projects Page at:

http://www.ml.cmu.edu/research/student_research_projects.htm

The Data Analysis Project is to be carried out under the supervision of a Machine Learning Department faculty member, and possibly joint supervision of a substantive expert. It is to be concluded by a written report. The ideal report would demonstrate an ability to approach machine learning problems in a way that cuts across existing disciplinary boundaries. It should demonstrate a capacity to write about technical topics in machine learning in a cogent and clear manner for a professional/scientific audience.

Administratively, the student can sign up for Data Analysis Project course 10-910, or can simply work on the Data Analysis Project part of the Reading & Research course 10-920. Note that a new course number is under consideration, which would include work on both the data analysis project as well as on the speaking requirement, so this administrative procedure may change in the near future.

Data Analysis Project Requirements:

The intent is that the Data Analysis Project will be less formal in structure and more flexible in focus than a typical Masters thesis + defense requirement might allow. The Project is a requirement for those in other departments receiving a MS degree in Machine Learning as well as for PhD students in Machine Learning. The requirement will typically be completed during a student's 2nd year in the program. The requirements are:

- 1) A presentation of the work as a regular Machine Learning seminar.

The presentation stands in lieu of a defense of the Data Analysis project, and helps to disseminate the work to the rest of the Machine Learning community. All

Machine Learning faculty and students will be invited. There will be a limited set of dates available for such presentations---generally, at most one per week---so students should be sure to sign up early.

The presentation should be suitable for a general machine learning audience. I.e., it should provide sufficient background for a non-domain-expert to understand the results, and should adequately summarize the relationship of the project to previous work.

2) A stand-alone, single or lead author written paper that is approved by the faculty who is advising the Project.

The paper should be of high quality, both in terms of exposition of technical details and overall English and organization. It should be suitable for submission to a journal or to a refereed conference, but will typically be substantially longer than the usual conference proceedings papers. Although it does not have to be published, publishing the paper may be desirable and helpful to the student. It should be noted that the project papers will become part of the MLD archives, and will serve as examples to future students.

The student must provide a draft of the paper at least one month before the scheduled oral presentation. Both student and advisor(s) must certify that this draft is substantially complete. The final version of the paper, incorporating any feedback received at the oral presentation, should be submitted for review no later than one month after the oral presentation.

Electives:

Electives may be chosen from Carnegie Mellon's large number of graduate courses, in consultation with the student's advisor & Co-Directors, to meet the interdisciplinary distribution requirements. One of these three electives is taken from the offerings in Statistics. The other two advanced electives, chosen in consultation with the student's advisor, form a concentration in one of the allied disciplines with SCS, Biology, Statistics, or Tepper School of Business. For those candidates seeking an academic position after completing the CSL Ph.D. degree, the thoughtful selection of these three elective courses is particularly important. As in the each of the first two years, coursework is supplemented by 24 units/term of research.

Proficiencies in Programming, Teaching, Conference Presentation and Research Skills:

- The programming skill requirement is normally demonstrated during the student's first two years of research, carried out under the supervision of the student's research advisor.

- Each Ph.D. candidate must participate in two terms of instruction, either through TA duties or serving as the instructor for a class.
- The demonstration of conference presentation and related research skills is normally achieved through the KDD project requirement.

Speaking Requirement:

To satisfy the oral communication skill requirement, each student should give a public talk at Carnegie Mellon. The talk should be accessible to a general SCS audience. The talk is scheduled by the PhD Administrator, and members of the speaking committee attend and evaluate the presentation, as well as provide oral and written feedback to the student.

All members of the speaking committee and the student's advisor should be in attendance. Immediately after the talk, the speaking committee members and the student's advisor confer among themselves (with the student absent) about the presentation. The committee members also fill out a Speaking Review Form.

As with writing, speaking well takes practice. Satisfying this requirement might take a few tries on the student's part, and no stigma is attached to those who have to try more than once. The recommended time to first be evaluated for the speaking requirement is during the KDD Project Presentation. If the student does not pass the speaking requirement at the time of the KDD project presentation, the student can either satisfy the speaking skills requirement at an internal CMU seminar, or at the dissertation proposal. They are, however, encouraged to satisfy this requirement as soon as possible, and should consult with their advisor to choose an appropriate schedule.

Directed Research

During a student's first two years, he or she should be doing directed research at least half time; once all coursework is completed and before doing thesis research, full time (except when teaching). Different students, and different advisors, have different ideas of what directed research means and how progress can be demonstrated. It is the responsibility of both the student and his or her advisor to formulate for each semester a set of reasonable goals, plans, and criteria for success in conducting directed research.

At each semi-annual faculty meeting, known as Black Friday, the faculty review the student's previous semester's research progress and the student's next semester's research plans to ensure that the student is making satisfactory progress. The evaluation of a student's progress in directed research often depends on the student having produced some tangible result; examples include the implementation of pieces of a software system, a written report on research explorations, an annotated bibliography in a major area, or, as part of preparation for doing research, a passing grade in a graduate course (beyond the required 96 required units).

Advisors are individually responsible for adequately supervising this portion of the Ph.D. program.

Ph.D. Thesis

Required coursework is to be completed by the end of the third year. By the start of the fourth year a Ph.D. candidate will present a thesis Prospectus to the Machine Learning community. The prospectus should include:

- a clear statement of the proposed research problem
- including arguing for the significance of the proposed research
- a review of relevant literature relating to the problem
- a review of the candidate's work leading up to the thesis
- a tentative schedule for completing the work.

Advising on scheduling the prospectus, and guiding in the formation of the dissertation committee, is the thesis advisor's responsibility. Normally, the thesis advisor is one of the Machine Learning faculty, but this is not mandatory. The thesis committee should be composed of at least four members, one of whom is an external member and at least one of whom is a Machine Learning faculty member. The external member may be from another department at Carnegie Mellon, or (typically) from outside the University. All thesis committees are subject to departmental approval.

Normally, the dissertation is completed during the student's fifth year. The final defense is a public presentation, in accord with the College and University requirements for the Ph.D. It is the candidate's responsibility to ensure that the College and University's guidelines are followed for publicity of the defense, and the availability of the thesis at least one week prior to the defense.

Advising and Student Evaluation

The CSL program is supervised by two faculty co-directors. Evaluation and feedback on a student's progress are important both to the student and to the faculty. Students need information on their overall progress to make long range plans.

The faculty meet at the end of each semester to make a formal evaluation of each student in the program. For historical reasons this meeting is called "Black Friday." The purpose of having all the faculty meet together to discuss all the students is to ensure uniformity and consistency in the evaluation by all of the different advisors. The faculty measure each student's progress against the goal of completing the program in a reasonable period of time. In their evaluation the

faculty consider courses taken, directed research, teaching if applicable, skill, development, papers written and lectures.

The faculty's primary source of information about the student is the student's advisor. The advisor is responsible for assembling the above information and presenting it at the faculty meeting. The student should make sure the advisor is informed about participation in activities and research progress made during the semester. Each student is asked to submit a summary of this information to the advisor at the end of each semester.

Based on the above information, the faculty decide whether a student is making satisfactory progress in the program. If so, the faculty usually suggest goals for the student to achieve over the next semester. If not, the faculty make more rigid demands of the student.

Ultimately, permission to continue in the program is contingent on whether or not the student continues to make satisfactory progress in their home department and toward the ML Masters degree. If a student is not making satisfactory progress, the faculty may choose to drop the student from the program. Graduate students meet with these co-directors for approval of their curriculum, particularly for elective selections. In addition each graduate student is matched with a faculty research advisor in the fall of the first year, who oversees the student's required Research and Reading course. Twice each year, at "Black Friday" meetings, the faculty review the progress of each student in the program. The co-directors and faculty research advisors communicate in written and oral form the assessment from these Black Friday meetings to the graduate students.

Financial Support

The Machine Learning Department is committed to providing full tuition and stipend support for the academic year, for each full-time CSL Ph.D. student, for a period of 5 years. Research opportunities are constrained by funding availability. The funding commitments assume that the student is making satisfactory progress in the program, as reported to the student at the end of each academic term. Students are strongly encouraged to compete for outside fellowships and other sources of financial support. The department will supplement these outside awards in order to fulfill its obligations for tuition and stipend support.

Travel Support

The department encourages students to travel to conferences and workshops to enhance their professional and career development.

Policy: If a student wants to attend a conference or workshop, the student's advisor or research sponsor should support the trip through either a research contract or a discretionary account. Student travel is unlimited as long as there is

money available from research contracts and/or discretionary funds of a sponsoring faculty member.

If no such funding is available to the student, then limited departmental funds may be available upon request from the Department Head. Since departmental funds are limited, the department will reimburse you up to \$350. Department funding is only available to the student for one trip per year and will not be transferred to the following year.

Process: To obtain travel support, the student and his or her faculty advisor/research sponsor must first agree that the student should take the trip. Then in advance of the trip the student should print the Student Travel Authorization Form, and then get the advisor's signature before forwarding the form to the PhD Administrator. The faculty member must either (i) indicate the amount of support the student may receive and its source (be sure the charge number is filled in!), or (ii) state on the Comments line that no funds are available from any research or discretionary account.

Leave of Absence Policy

Students who wish to leave the program temporarily may request a leave of absence by submitting a request to the PhD Administrator. Leaves are initially granted for a period of no more than one year, but an extension of up to one additional year may be granted under exceptional circumstances. When an extension is granted, the conditions for return must be negotiated with the advisor and the Ph.D. Program Co-Directors, prior to returning to the program.

Students on leave of absence should contact the PhD Administrator two months prior to the end of the leave to indicate their plans for the next year.

Grievances

Students and advisors enjoy a close working relationship in our program. If students have problems, whether related to their research or not, they should feel free to speak to their advisor. If doing so is awkward or if students simply want a second opinion, they should feel free to discuss their problems with either or both of the Co-Directors or the PhD Administrator.

General Information

Computers

Every incoming student will have a PC computer on his or her desk. All students will be given a CS computer account. The School of Computer Science has a Help Center located at 3613 Wean Hall. If you have any computer or account problems contact help@cs.cmu.edu or call 8x4231 from a campus phone.

Photocopier/Printers

In order to use the photocopiers in Wean Hall you must have an access code. The code for the black & white copier in Wean 4215 is #80761. In order to use the color copier/printer spectrum, in 4215, you must have a code for copies but not for printing. The code for Machine Learning students is #3597. Printers in Wean 4215: Onyx, Spectrum (color) Stone. Printers in Wean 8108: Graphite, Iron, Sunbow (color). Color transparency printers are located in the Help Center, Wean 3613: Crayon & Slide1. To find out other names/locations of printers in Wean, please see the SCS Facilities website <http://www.cs.cmu.edu/~help/printing/>

The copier codes are to be used by Machine Learning students only and are not to be given out to anyone not currently in the program.

Copy Center

There is a copy center in Wean 4602 where you can drop off items to be copied. You will need a charge number if this is for the program or you can pay cash if this is for personal use.

CS Main Office

Functions of the Main Office (4212)

- Send mail
- Pick up packages or faxes
- Get office supplies (make sure you sign them out)
- Send overnight packages
- Sign out keys for conference rooms

Your mail will be in the ML Lab (4616 Wean Hall).

US Post Office is located in the basement of University Center.

Seminars

The Machine Learning Department sponsors seminars by researchers from within and outside Carnegie Mellon, which are attended by faculty, staff and graduate students. Students are encouraged to meet and interact with visiting scholars. This is extremely important, both to get a sense of the academic projects that are pursued outside of Carnegie Mellon and to get to know the leaders of such projects. That applies not only to seminars directly relevant to a student's research interests: the seminars provide an opportunity to widen one's perspective on the field.

The Emigration Course

The Emigration Course grooms finishing students for their career afterward. It is structured as a series of talks offered throughout the year and focuses on five topics: Jobs, The Real World, Money, Ethics, and Communication. These talks cover nuts-and-bolts issues like how to job interview, how to apply for grant money, and how to write a technical paper. They also expose students to traditional and non-traditional career paths in academia, industry, and

government. Participation is open to the entire SCS community and is completely voluntary. More senior students, especially those planning to finish in any given year, are encouraged to attend sessions offered that year; however, even junior students can benefit from attending, to prepare for a smooth transition from life as a student to life in the real world.