Master of Science in Machine Learning

Student Handbook

Revised 3/21/13

Table of Contents

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 M.S. 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.

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 machine learning efforts, and to pursue research on the next generation of data mining tools, algorithms, and systems.

The Co-Directors of the program are:

Co-Directors of the program:
Geoffrey Gordon, Associate Research Professor, Machine Learning Dept.
Email: (ggordon@cs.cmu.edu) Phone: x7399

Rob Kass, Professor, Statistics Dept.
Email: (kass@stat.cmu.edu) Phone: x8723

Administrative Support:
Diane Stidle, Graduate Programs Administrator (diane@cs.cmu.edu) x1299,
8203 Gates-Hillman Center.

Program Requirements

Prerequisites, Computer Science:

15-150 Principals of Functional Programming

An introduction to programming based on a "functional" model of computation. The functional model is a natural generalization of algebra in which programs are formulas that describe the output of a computation in terms of its inputs---that is, as a function. But instead of being confined to real- or complex-valued functions, the functional model extends the algebraic view to a very rich class of data types, including not only aggregates built up from other types, but also functions themselves as values. This course is an introduction to programming that is focused on the central concepts of function and type. One major theme is the interplay between inductive types, which are built up incrementally; recursive functions, which compute over inductive types by decomposition; and proof by structural induction, which is used to prove the correctness and time complexity of a recursive function. Another major theme is the role of types in structuring large programs into separate modules, and the integration of imperative programming through the introduction of data types whose values may be altered during computation. NOTE: students must achieve a C or better in order to use this course to satisfy the pre-requisite.

15-210 Parallel and Sequential Data Structures and Algorithms

Teaches students about how to design, analyze, and program algorithms and data structures. The course emphasizes parallel algorithms and analysis, and how sequential algorithms can be considered a special case. The course goes into more theoretical content on algorithm analysis than 15-122 and 15-150 while still including a significant programming component and covering a variety of practical applications such as problems in data analysis, graphics, text processing, and the computational sciences. NOTE: students must achieve a C or better in order to use this course to satisfy the pre-requisite.

Previously offered Computer Science courses 15-211 and 15-212 would also fulfill the prerequisite requirement.

Prerequisites, Statistics:

36-225: Introduction to Probability Theory

This course is the first half of a year-long course which provides an introduction to probability and mathematical statistics for students in economics, mathematics and statistics. The use of probability theory is illustrated with examples drawn from engineering, the sciences, and management. Topics include elementary probability theory, conditional probability and independence, random variables, distribution functions, joint and conditional distributions, law of large numbers, and the central limit theorem. A grade of C or better is required in order to advance to 36-226. Not open to students who have received credit for 36-625. **36-217 Probability Theory and Random Processes, will also be accepted as a prerequisite.**

36-226: Introduction to Statistical Inference

This is mostly a theoretical course in statistics. First, we will give a formal introduction to point estimation and consider and evaluate different methods for finding statistical estimates. Then we will discuss interval estimation and hypothesis testing, which are necessary for most statistical analyses. In this first part of the course, the emphasis will be on definitions, theorems and mathematical calculations. Once we have covered the mathematical foundations of statistical inference, we will focus on the use of these concepts in concrete statistical situations. We will study statistical modeling and specific models such as ANOVA

and regression. Emphasis will be placed on understanding the qualities of a good statistical analysis, specifying correct models, assessing model assumptions and interpreting results.

Previously offered Statistics courses 36-625 and 36-626 would also fulfill the prerequisite requirement.

Core Courses:

The core courses you must take are:

10-705: Intermediate Statistics
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: Machine Learning
Machine learning studies the question "How can we build computer programs that automatically improve their performance through experience?" This includes learning to perform many types of tasks based on many types of experience. For example, it includes robots learning to better navigate based on experience gained by roaming their environments, medical decision aids that learn to predict which therapies work best for which diseases based on data mining of historical health records, and speech recognition systems that lean to better understand your speech based on experience listening to you. This course is designed to give PhD students a thorough grounding in the methods, theory, mathematics and algorithms needed to do research and applications in machine learning. The topics of the course draw from machine learning, from classical statistics, from data mining, from Bayesian statistics and from information theory. Students entering the class with a pre-existing working knowledge of probability, statistics and algorithms will be at an advantage, but the class has been designed so that anyone with a strong numerate background can catch up and fully participate.

10-702: Statistical Machine Learning
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: Multimedia Databases and Data Mining
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.

15-750 Graduate Algorithms
or
15-853 Algorithms in the Real World
This course covers how algorithms and theory are used in "real-world" applications. The course will cover both the theory behind the algorithms and case studies of how the theory is applied. It is organized by topics and the topics change from year to year.

Electives:

Electives may be chosen from Carnegie Mellon's large number of graduate courses, in consultation with the student's advisor, to fit with the student's educational program. Elective choices are subject to review by the co-directors. Elective courses may be counted toward a simultaneous PhD degree at CMU, but not toward any other Masters-level degree.

For those candidates seeking an academic position after completing the ML M.S. degree, the thoughtful selection of these three elective courses is particularly important.

Double Counting Courses:

Any course counted toward another master-level or bachelor-level degree may not be counted toward our Secondary Master in Machine Learning. If a course is counted toward your PhD degree it may also be counted in our Secondary Master in Machine Learning, so long as such double-counting is permitted by your PhD department.

Data Analysis Project (DAP)

Once admitted into the secondary Masters degree program in ML, students have until the end of the following semester to identify an advisor in ML who will serve as their DAP advisor.

Students are required to demonstrate their grasp of fundamental data analysis and machine learning concepts and techniques in the context of a focused project. The project should focus on a substantive problem involving the analysis of one or more data sets and the application of state-of-the art machine learning and data mining methods, or on suitable simulations where this is deemed appropriate. Or, the project may focus on machine learning methodology and demonstrate its applicability to substantial examples from the relevant literature. The project may involve the development of new methodology or extensions to existing methodology, but this is not a requirement.

Machine learning and data mining methods are exemplified by, but not limited to, those covered in the core courses 10-701, 10-702, and 15-826. In particular, the analysis methods should be adequately justified in terms of the theory taught in these courses.

The project is not intended for purely theoretical or methodological investigations, but these may form the heart of a project in appropriate cases. (In such cases, the project should also contain a component of applying the new theoretical or methodological tools to data. This component does not have to contain novel results; instead, its goal is to characterize how well or poorly the tools perform for the given data.) Students are encouraged to seek out a project (co)advisor who can provide access to data or substantive applications, or can use data sets to which they already have access through one of the core courses, through the literature and archives, or through their PhD advisor. Other resources for this purpose include the Immigration Course, faculty home pages, and the ML Research Projects webpage.

The Data Analysis Project is to be carried out under the supervision of a Machine Learning Department faculty member, and possibly under joint supervision of a subject matter 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 and scientific audience.

Research for the Data Analysis Project is typically done as part of the Reading & Research course, 10-920. The student must register for the ML Journal Club, 10-915, for the semester they intend to present their Data Analysis project.

DAP Committee

Student must form an official "DAP committee" of three faculty to evaluate the document. The committee will consist of the advisor, the Journal club instructor(s), and one other faculty member selected by the student. The third member is often someone with an interest in the analysis of the data set, and does not have to be an expert in ML or part of the student's thesis committee. The student should form the committee as early as possible during the DAP research process, and inform Diane of who the members are.
2 of 3 DAP Committee members, one of whom is the DAP advisor, must be in attendance for the DAP presentation.

DAP Prospectus

Student must write a 1-2 page prospectus, including the DAP's title, general topic, proposed data source, and a brief summary of proposed analysis methods, and circulate it to the committee. The student should do this as early as possible, preferably when the student forms the committee.

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.

DAP Requirements:

1) A presentation of the work during the Machine Learning Journal Club course. 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. 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 in the Machine Learning Journal Club. 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 of 3 DAP Committee members, one of whom is the DAP advisor, must be in attendance.

2) A stand-alone, single or “lead author’ written paper that is approved by the faculty member(s) 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 refereed conference. But, unlike some conference papers, it should be completely self-contained, including all descriptions necessary for a general machine learning audience to follow the theoretical development and reproduce the experimental results. This requirement may (but does not have to) result in the project paper being substantially longer than a conference proceedings paper on which it is based. Although it does not have to be published, publishing the paper may be desirable and helpful to the student. Project papers will become part of the MLD archives, and will serve as examples to future students.

3) The student must provide a near-final draft of the DAP document (approximately 15 pages) at least one month before the oral presentation to the DAP Committee. Both student and committee must certify that this draft is substantially complete. Within two weeks of submission, the instructor(s) will either approve the project for presentation (at which point the presentation can be advertised to the members of the department), or notify the student that changes will be required before presentation. This approval is for the general topic and content, and not for the final contents of the document. 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.

Student Evaluation

The faculty meet at the end of each academic semester to make a formal evaluation of each student in the program. For historical reasons this meeting is called "Black Friday." The co-directors and faculty research advisors communicate in written and oral form the assessment from these Black Friday meetings to the graduate students.

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.

At each semi-annual “Black Friday” meeting, 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).

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 degree. If a student is not making satisfactory progress, the faculty may choose to drop the student from the program.

Terms of progress in Black Friday letters from faculty:

SP = In the semiannual evaluation of all our students the faculty reviewed your progress toward the Ph.D. We are happy to report that you are in good standing in the Machine Learning PhD program.

USP = We have determined that your current level of progress is unsatisfactory:

N-2 = We have determined that there are significant problems with your current level of progress. Accordingly, this is an N-2 letter: you are in danger of receiving an N-1 letter next Black Friday unless you improve your rate of progress toward a Ph. D. In particular:

N-1 = *This is an N-1 letter. You may not be allowed to continue in the PhD program past the next Black Friday meeting unless you satisfy the following conditions:*

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.