Syllabus for CSCI 353: Machine Learning

Professor: Susan L. Epstein **Email:** susan.epstein@hunter.cuny.edu **Office:** 1090C Hunter North
Office hours: Mondays and Wednesdays, 3–4 PM and by appointment **Telephone:** 212-772-5210
Department office: 1008 Hunter North **Department telephone:** 212-772-5213
Mode of instruction: P (in person) **Class meets:** Mondays and Wednesdays, 5:35–6:50 PM in C107 HN
Course website: On Blackboard, accessible through the CUNY Portal with Chrome, Firefox, or Safari.

Course description

Machine learning is the subfield of artificial intelligence that learns to predict and classify from data. Course material is organized around three essential questions:

- What does it mean for a machine to learn?
- How can we support and evaluate machine learning?
- How do algorithms help machines learn?

This interdisciplinary course takes a pragmatic, hands-on approach to material that is rigorously grounded in mathematics. Course work blends theory with practice. You will learn to use Weka, a powerful, open-source suite of machine learning tools that addresses the theoretical and computational challenges in machine learning. Anticipated topics are posted on the class website under course information. As you explore data from many different sources, this course will change forever your ideas about computers and learning.

Prerequisites

Making the future is fun, but it also takes knowledge. All students should have *completed* CSCI 235 and CSCI 150 with a grade of C or better. If you do not satisfy this requirement, see the instructor *immediately* after the first lecture, and bring a current transcript. Some background in basic probability theory and statistics will prove useful. No programming is required but a general fondness for mathematics and data is essential.

Required course material

Text: *Data Mining: Practical Machine Learning Tools and Techniques,* 2017, Witten, Frank, and Hall. Morgan Kaufmann, the **fourth edition**, ISBN 978-0-12-804291-5. (Don't let "data mining" rather than "machine learning" worry you…you're in the right place.)
Available through the Hunter bookstore: `https://hunter.textbookx.com/institutional/index.php?action=browse#books/2098266/`

Errata for the text are posted at `http://www.cs.waikato.ac.nz/ml/weka/errata.html`

Additional reading material will be posted on the course website. Some of the course material is copyrighted and therefore **available only at Hunter**. You can access it only If you log into the library while you are on campus or if you go to the library. Plan accordingly.

Software: All work in this course requires Weka 3-8-3. **No other software is acceptable.** (Weka 3-9-3 is likely to still be buggy; avoid it.) All assignments and the project are Weka-oriented, and require you to run Weka, *often for many hours*, on multiple data files. The Weka 3-8-3 manual is posted under Course Materials on the class website. Everything else you want to know should be at:
`http://www.cs.waikato.ac.nz/ml/weka/index.html`

There are two ways you can use Weka:

- **Install it** on your home machine. Weka is available for OS, Linux, and Windows. This is by far the simplest choice, and the one most students opt for. Instructions are available on the course website.
- **Use it in the Linux lab**. If you choose this option, plan on spending *considerable* time there. See the course website for how to find Weka in the lab. To use those machines, you must have a CSCI departmental Linux account. See me immediately if you need one.

Whichever you choose, be sure to check out the Weka link above and the documentation manual.

Learning Weka. If you need more help than the brief demonstrations in class, I strongly recommend that you avoid surfing the Web. Instead go to `https://www.cs.waikato.ac.nz/ml/weka/courses.html`, which is more likely to be accurate and current. Although many of these videos are based on the third edition of the text, most seem to be applicable to version 3-8-3.

Datasets. You are going to run algorithms on data…a lot. A small default collection of datasets is available when you install Weka. Additional datasets are on the machines in the 1001B Linux lab, many of which you can also download from `http://www.cs.waikato.ac.nz/ml/weka/datasets.html`. See the course website for how to find the data files in the lab. You will not need every dataset, but you should familiarize yourself with them, probably at Hunter first. Because most Weka experiments require that all of the data be present in memory, *large datasets can be a problem, on any machine.* You will need substantial free disk space if you choose to work elsewhere. There may also be additional datasets posted for the project. *After discussion with the instructor*, students are also welcome to work with their own datasets *in addition* to the required ones.

Students' responsibilities

Come to all classes on time and well prepared. All reading, audio, and video assignments are to be done *before* class. The text is a necessary resource, but lectures will go far beyond it. You are responsible for all material in the reading, whether or not it is covered during class. Be ready to ask, and to answer, questions on the reading. Detailed notes are highly recommended.

Maintain a Linux account with the Department and abide by the rules for the Department's labs. **If you do not have a CSCI Linux account, contact me immediately.** If you already have one **you must reclaim it by September 16.** To do so, log into any machine (through `ssh` or in person). That should put you in your home directory: `/data/blocs/b/student.accounts/firstname.lastname`. Then enter `touch fall.2019`. That will reclaim it.

Regularly read all email sent by the instructor to your registered Hunter Blackboard address. Changes in assignments, clarifications, and instructions will often come by email.

Keep pace with the course. The course schedule, much of the required reading, all assignments, and the project are posted on the Blackboard website. Be sure to **check it regularly for changes** as the semester progresses.

Acknowledge any help received from other people, and reference in full any material (e.g., book, paper, journal, web site) used to prepare assignments. **Be sure to read "How to avoid plagiarism" on the class website.**

Assignments = reading + written homework + project

All assignments are designed to increase your understanding of the material, and must be done as the course progresses. Often, one will build upon the next, so **skipping an assignment is *not* an option**. *It is not possible to do well in this course without doing the assignments thoroughly and submitting them on time.*

• **Reading assignments require hours of careful study.** Classes are organized by topic; some topics will span several classes. If you do not do the reading *regularly and thoroughly in the current (fourth) edition of the book*, you will find the course extremely difficult. **Material not covered by reading may be included on the exams.** Appendix A in the text is a key reference but can be overwhelming; when you need more math background, look there first.

• **Written assignments** require *hours* of thought and effort. **Plan on spending time on each written assignment *over several days*.** You may discuss your ideas with each other, but you must do your own work **All assignments must be machine-printed.** If part of an assignment is requested by **email, use pdf or a snapshot** of your screen only. **Do not email handwritten answers, or arff, csv, or exe files.**

• **Project.** Your project is a more intensive exploration with Weka. How much fun you have with it and how ambitious you are will determine how much time it will take.

All assignments are due at the beginning of class on their respective due dates. Written assignments must be delivered to the instructor in class. If you must miss class (never a good idea), get your homework *to me* before class. Otherwise, you must have your homework *time-stamped by the Department office and left in my mailbox there*. (Please do not try to put it under my office door.)

Penalty for late assignments:
- ε to 24 hours late: 25% penalty
- 24 to 48 hours late: 50% penalty
- More than 48 hours late: **no credit**

The instructor may grant a brief extension on an individual basis, *if requested in advance*. Repeated or last-minute requests for extensions will be denied in other than extraordinary circumstances.

Grading

This course includes both theory and practice. You must be able to define important terms and to explain ideas in clear English. Course grades are based on written assignments (see above), a project with Weka, two exams, and thoughtful, well-prepared class participation. The project will include an original essay of at least 500 words. **Grades will not be curved.**

• Assignments	35%
• Exams	40%
• Project	20%
• Class participation (asking questions counts!)	5%

Learning goals

This course addresses Departmental learning goals 1d, 3a, and 4.

Intelligent agents communicate

Talk to me

Everyone is expected to participate in class. Ask questions. Express opinions. In return, I am happy to answer questions, listen to concerns, and talk to any student about topics related to the class (or not). I actually *enjoy* student visits during office hours. You can also make an appointment to see me at other times. I also welcome your feedback throughout the semester about how the course is progressing.

Write to me

You can reach me by email almost every day, but not late at night or very early in the morning.

Avoid hunting on the web

Much of what is out there is out of date or simply wrong.

Course website

The course website is available on Blackboard and used in a variety of ways. Check it regularly for updates.

Be clear and correct

Homework and exam answers must be *legible and unambiguous*. If a question is of the "yes or no" type, you must justify your answer.

Share

There are some terrific ML web pages out there. If you find a good resource you'd like to recommend (not a page with broken links), please send it to me and I'll post it on the course website.

Writing

In accordance with Departmental requirements for elective courses, the final project includes (but is not limited to) a written analysis of at least 500 words.

Study groups

Although study groups are not required, students who work together typically learn much more than those who work alone. The ideal group size is three or four. You are encouraged to form study groups to increase your understanding of the material.

Course policies

Attendance: Students are expected to attend all classes.

Lateness: See the section on assignments for the lateness policy.

Missed exams: No makeup exams will be given.

Extra credit: Any extra credit opportunities will be specified in the assignments and/or project and due at the same time as the required material. No late extra credit will be accepted.

Blackboard: Students are expected to check the class website daily.

Email: Students are expected to read their Hunter email daily for clarifications and changes in reading and other assignments.

Time commitment: The amount of time you devote to this course will depend upon your interest, your ability to read technical material, and your skill with Weka. Plan on at least 8 hours per week outside of class for it.

Linux Lab: Access to the Linux lab is through your OneCard.
No eating or drinking in the Linux lab.
You must successfully log in using your account information.
If you need help, *first* look at the Linux lab FAQ at
`http://www.geography.hunter.cuny.edu/tbw/CS.Linux.Lab.FAQ/department_of_computer_science.faq.htm`

Remote login access is a privilege, not a right.
If you *still* need help logging in after that, send email to cstechsp@hunter.cuny.edu. Your request must:

- Originate from your MyHunter.cuny.edu email account
- Include the exact command you are trying to execute
- Include the exact error message(s)
- Include the name of the machine you're trying to log into
- Include your Linux account user name and your full name as it appears in CUNYfirst

Hunter College Policy on Academic Integrity: Hunter College regards acts of **academic dishonesty** (e.g., plagiarism, cheating on examinations, obtaining unfair advantage, and falsification of records and official documents) as serious offenses against the values of intellectual honesty. The College is committed to enforcing the CUNY Policy on Academic Integrity and will pursue cases of academic dishonesty according to the Hunter College Academic Integrity Procedures.
Clarification: Discussion of assignments is fine, but **you must run your own experiments and write your own assignments and project**. *Giving and receiving output or answers are equally reprehensible*

ADA policy: In compliance with the American Disability Act of 1990 (ADA) and with Section 504 of the Rehabilitation Act of 1973, Hunter College is committed to ensuring educational parity and accommodations for all students with documented disabilities and or/or medical conditions. It is recommended that all students with documented disabilities (Emotional, Medical, Physical and/or Learning) consult the Office of AccessABILITY

located in Room E1214B to secure necessary academic accommodations. For further information and assistance please call 212-772-4857 or 212-650-3230.

Hunter College Policy on Sexual Misconduct: In compliance with the CUNY Policy on Sexual Misconduct, Hunter College reaffirms the prohibition of any sexual misconduct, which includes sexual violence, sexual harassment, and gender-based harassment retaliation against students, employees, or visitors, as well as certain intimate relationships. Students who have experienced any form of sexual violence on or off campus (including CUNY-sponsored trips and events) are entitled to the rights outlined in the Bill of Rights for Hunter College.
Sexual Violence: Students are strongly encouraged to immediately report the incident by calling 911, contacting NYPD Special Victims Division Hotline (646-610-7272) or their local police precinct, or contacting the College's Public Safety Office (212-772-4444).
All Other Forms of Sexual Misconduct: Students are also encouraged to contact the College's Title IX Campus Coordinator, Dean John Rose (jtrose@hunter.cuny.edu or 212-650-3262) or Colleen Barry (colleen.barry@hunter.cuny.edu or 212-772-4534) and seek complimentary services through the Counseling and Wellness Services Office, Hunter East 1123.
http://www2.cuny.edu/wp-content/uploads/sites/4/page-assets/about/administration/offices/legal-affairs/POLICY-ON-SEXUAL-MISCONDUCT-10.1.2015-with-links.pdf

Want to learn more?

Students often ask for additional reference material. Here are some suggestions:
Do not go web surfing. Many sites and blogs are inaccurate or shallow or both.
General machine learning:
• Andrew Moore's tutorials `https://www.cs.cmu.edu/~awm/tutorials.html`
• *Bayesian Reasoning and Machine Learning,* David Barber. **Available online** at
`http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/240415.pdf` with errata at
`http://web4.cs.ucl.ac.uk/staff/D.Barber/pmwiki/pmwiki.php?n=Brml.Errata`
• *The Elements of Statistical Learning*, Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2008, a solid mathematical approach, **available online** at
`https://web.stanford.edu/~hastie/Papers/ESLII.pdf`
• *Foundations of Machine Learning*, Mehryar Mohri, Afshin Rostanmizadeh, Ameet Talwalkar, second edition, 2018. Computational learning theory meets deep learning.
• *Introduction to Machine Learning,* Etham Alpaydin, second edition. A theoretically oriented text
• *Machine Learning,* Tom Mitchell. 1997 but a classic and accessible text
• *Machine Learning,* Peter Flach, 2012, mathematical but reasonably accessible with limited topic coverage
• *Pattern Classification,* Richard Duda, Peter Hart, and David Stark, second edition, 2001. The rigorous statistical origins of much research in machine learning
• *Machine Learning: A Probabilistic Perspective,* Kevin P. Murphy, 2012. The probabilistic compendium, with a heavy mathematical bent.
• *Pattern Recognition and Machine Learning*, Christopher M. Bishop, 2007. An excellent graduate level text with a heavy mathematical bent.
• Feeling ambitious? If you want to commit to Jupyter Notebooks, you could learn Tensor Flow and use Google's cloud: `https://ai.google/tools/`
Statistics:
• *Introduction to the New Statistics: Estimation, Open Science, and Beyond*, 2017, Cumming and Calin-Jageman, Taylor & Francis, ISBN 9781138825529, a basic and clear introduction to statistical testing
• *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*, 2012, Cumming, Taylor & Francis, ISBN 9780415879682, a more advanced version that elaborates on the material in *Introduction to the New Statistics*
Deep learning:
• *Deep Learning,* Goodfellow, Bengio, Courville, 2016. Heavily mathematical and an excellent text.

• *Introduction to Deep Learning*, Charniak, 2018. Python-oriented and gentle.
• *Neural Networks and Deep Learning,* Michael A. Nielsen. Brief and elementary, a**vailable online** at
`http://neuralnetworksanddeeplearning.com/`

Mathematics: (and see more resources on our webpage):
* *Probability: The Analysis of Data*, Guy Lebanon. Volume 1 covers most of the math you wish you knew, and is **available online** `http://www.theanalysisofdata.com/probability/0_2.html`

Want to stay current?

The top conferences and journals in the field are
• *International Conference on Machine Learning* (ICML)
• *Conference on Neural Information Processing Systems* (NeurIPS)
• *Annual Conference on Learning Theory* (COLT)
• *Journal of Machine Learning Research* (JMLR) available free on line at `www.jmlr.org`
• *Machine Learning* (MLJ) Published by Springer

How to do well in this course

- **Allot *substantial* time from your life to this course.**
- **Do the assigned reading *before* the lecture, and ask informed questions.**
- **Attend class faithfully and take detailed notes.**
- **Ask questions in class when you don't understand something.**
- **Study your lecture notes, the reading, and the practice problems.**
- **Submit all assignments ON TIME.**
- **Abide by the Department's Policy on Academic Dishonesty.**

Except for changes that substantially affect implementation of the grading policy, this syllabus is a guide for the course and is subject to change. In particular, the course schedule, including topics and all assignments, is subject to change. Be sure to check the course schedule online regularly. Since you have read this far, before 5:35 PM on September 4 email a picture of your favorite beach to me to get 10 extra credit points on your first homework assignment.

Acknowledgements

Scholars acknowledge their sources. Thanks to Jimmy Ba, Yoshua **Bengio,** Carla Brodley, Aaron Courville, Tom Dietterich, Mark Ebden, Tim Finin, Peter Flach, Ian Goodfellow, Geoff Hinton, Marie des Jardins, Thorsten Joachims, Ron Kohavi, Miroslav Kubat, Yann LeCun, Percy Liang, Tom Mitchell, Mehryar Mohri, Andrew Moore, Andrew Ng, Peter Norvig, Smiljana Petrovic, Afshin Rostamizadeh, Lorenzo Rosasco, Stuart Russell, Russ Salakhutdinov, Padhraic Smyth, Ameet Talwalkar, Dan Weld, and Pat Winston, from whose material some of this course was adapted.