CPSC 540: Machine Learning

Mark Schmidt

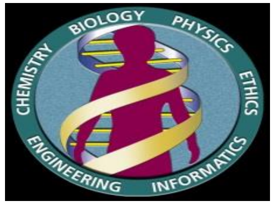
University of British Columbia, Winter 2017

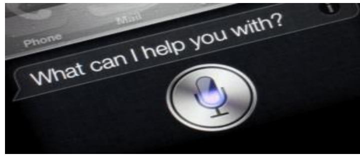
www.cs.ubc.ca/~schmidtm/Courses/540-W17

Some images from this lecture are taken from Google Image Search.

Big Data Phenomenon

- We are collecting and storing data at an unprecedented rate.
- Examples:
  - News articles and blog posts.
  - YouTube, Facebook, and WWW.
  - Credit cards transactions and Amazon purchases.
  - Gene expression data and protein interaction assays.
  - Maps and satellite data.
  - Large hadron collider and surveying the sky.
  - Phone call records and speech recognition results.
  - Video game worlds and user actions.

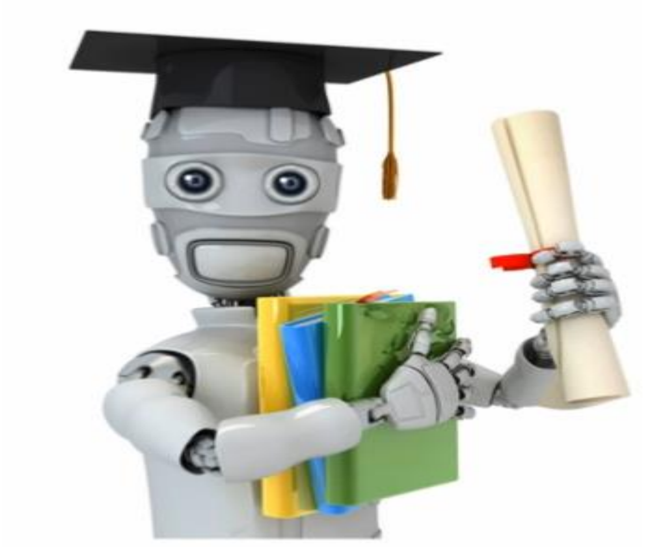




What can I help you with?

Machine Learning

- What do you do with all this data?
  - Too much data to search through it manually.
- But there is valuable information in the data.
  - Can we use it for fun, profit, and/or the greater good?
- Machine learning: use computers to automatically detect patterns in data and make predictions or decisions.
- Most useful when:
  - Don’t have a human expert.
  - Humans can’t explain patterns.
  - Problem is too complicated.

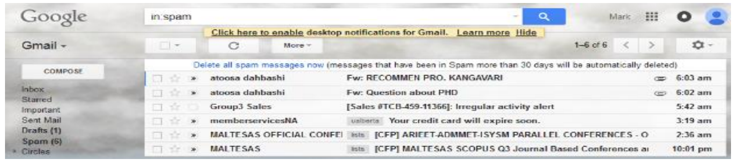


Machine Learning vs. Statistics

- Machine learning (ML) is very similar to statistics.
  - A lot of topics overlap.
- But ML places more emphasis on:
  1. Computation and large datasets.
  2. Predictions rather than descriptions.
  3. Non-asymptotic performance.
  4. Models that work across domains.
- The field is growing very fast:
  - ~2500 attendees at NIPS 2014, ~4000 at NIPS 2015, ~6000 at NIPS 2016.
  - Influence of $$$, too.

Applications

- Spam filtering.
- Credit card fraud detection.
- Product recommendation.
- Motion capture.
- Machine translation.
- Speech recognition.
- Face detection.
- Object detection.
- Sports analytics.
- Cancer subtype discovery.

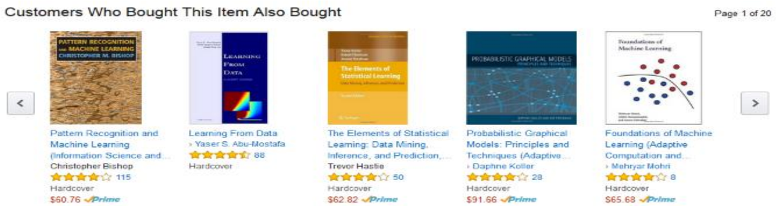


Click here to enable desktop notifications for Gmail. Learn more Hide

Delete all spam messages now (messages that have been in Spam more than 30 days will be automatically deleted)

atoosa dahbashi	Fw: RECOMMEN PRO. KANGAVARI	6:03 am
atoosa dahbashi	Fw: Question about PHD	6:02 am
Group3 Sales	[Sales #TCB-459-11366]: Irregular activity alert	5:42 am
memberservicesNA	ualberta Your credit card will expire soon.	3:19 am
MALTESAS OFFICIAL CONFE	lists [CFP] ARIEET-ADMMET-ISYSM PARALLEL CONFERENCES - O	2:36 am
MALTESAS	lists [CFP] MALTESAS SCOPUS Q3 Journal Based Conferences a	10:01 pm

Transaction Date	Posted Date	Transaction Details	Debit	Credit
Aug. 27, 2015	Aug. 28, 2015	BEAN AROUND THE WORLD VANCOUVER, BC	$10.95	



Customers Who Bought This Item Also Bought

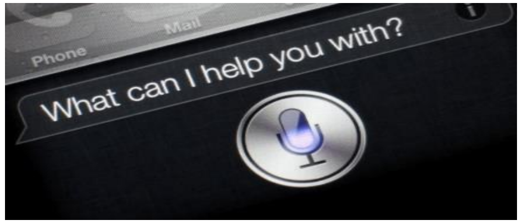
Page 1 of 20

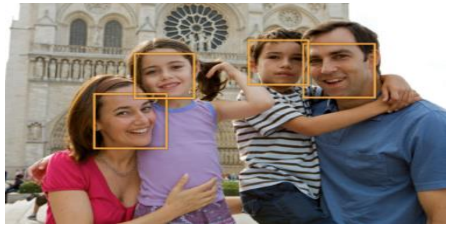
- Pattern Recognition and Machine Learning (Information Science and... Christopher Bishop, 115, Hardcover, $60.76 Prime
- Learning From Data › Yaser S. Abu-Mostafa, 88, Hardcover
- The Elements of Statistical Learning: Data Mining, Inference, and Prediction,... Trevor Hastie, 50, Hardcover, $62.82 Prime
- Probabilistic Graphical Models: Principles and Techniques (Adaptive... › Daphne Koller, 28, Hardcover, $91.66 Prime
- Foundations of Machine Learning (Adaptive Computation and... › Mehryar Mohri, 8, Hardcover, $65.68 Prime

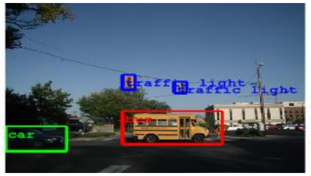
Data mining.

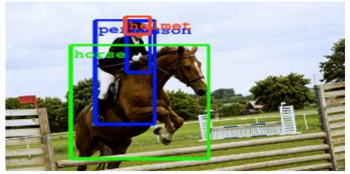
L'exploration de données .

Open in Google Translate









KLAY THOMPSON
VERY GOOD FROM HERE!
MEH...
GRANTLAND

Applications

- Gene localization/functions/editing.
- Personal Assistants.
- Medical imaging.
- Self-driving cars.
- Scene completion.
- Image search and annotation.
- Artistic rendering.
- Physical simulations.
- Image colourization.

Recent

M

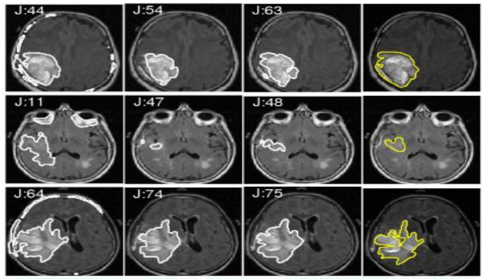
Active Now

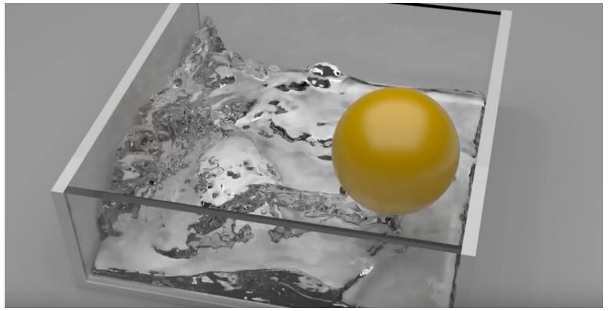
11:59 AM

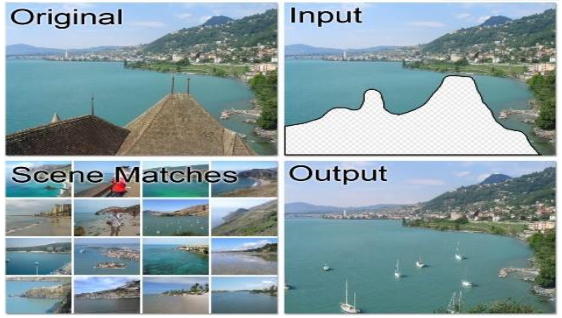
I'm going to Chicago next week. I'm looking for a great burger. Where should I go?

People rave about Command Burger.

Facebook personal assistant “M” looks so much better than Siri, Google Now & Cortana!







a cat is sitting on a toilet seat
logprob: -7.79

a display case filled with lots of different types of donuts
logprob: -7.78

a group of people sitting at a table with wine glasses
logprob: -6.71

Youngsters, May 1912

CPSC 340 and CPSC 540

- There are two ML classes: CPSC 340 and 540.
  - They are structured as one full-year course: 540 starts where 340 ends.
- CPSC 340:
  - Introductory course on data mining and ML.
  - Emphasis on applications of ML.
  - Covers implementation of methods based on counting and gradient descent.
  - Most useful techniques that you can apply to your research/work.
- CPSC 540:
  - Research-level ML methods and theory.
  - Not an introductory course:
    - Assumes familiarity with basic ML concepts.
    - Stronger math/CS background
    - Much more work.

CPSC 340 and CPSC 540

- If you can only take one class, take CPSC 340:
  - 340 covers the most useful methods.

- If want to work in ML you should take both courses:
  - There is not a lot of overlap between the topics, 540 is missing a lot important topics:
    - Learning theory, random forests, clustering, collaborative filtering, data visualization, and so on.
  - 540 is NOT an “advanced” version of 340.
    - It just covers the methods that require more advanced math/CS background.

- It is much better to do CPSC 340 first:
  - Many people have taken CPSC 340 after CPSC 540 (not recommended).

- There will be less overlap between 340 and 540 this year:
  - 340 now requires multivariate calculus, so many topics were moved from 540 to 340.
  - 540 will only cover the “diff” between 340 in 2015 and 2016.
    - If you took 340 before 2015, you should consider re-taking it – it is much more advanced now.

Course Outline

- 2-4 lectures on each of the following:
  - Large-scale machine learning.
  - Density estimation.
  - Graphical models.
  - Bayesian methods.
  - Recurrent neural networks.
  - Causal, active, and online learning (time permitting).
  - Reinforcement learning (time permitting).
- For an overview of topics covered in 340 and 540 see here:
  - http://www.cs.ubc.ca/~schmidtm/Courses/340-F16/L35.pdf

Math Prerequisites

- Research-level ML involves a lot of math.
- You should be comfortable with:
  - Linear algebra: vectors, matrices, eigenvalues.
  - **Probability**: conditional probability, expectations.
  - **Multivariate calculus**: gradients, optima.
  - Proof strategies and filling in derivation details.
- Suggested courses: Math 200, 220, 221, and 302.

- “I didn't really feel prepared for this course. I had never really done vector calculus before.”

Computer Science Prerequisites

- ML places a big emphasis on computation.
- You should be comfortable with:
  - Data structures: pointers, trees, heaps, hashes, graphs.
  - **Algorithms and complexity**:
    - Big-O, divide + conquer, randomized algorithms, dynamic programming, NP-completeness.
  - Scientific computing: matrix factorization, gradient descent, condition number.
- Suggested courses: CPSC 221, 302, and 320:
  - “I have programming experience in my work/research/courses” is not enough.

- “It is taught in a manner very hard and intimidating for those who are not in computer science.”

Stat/ML Prerequisites

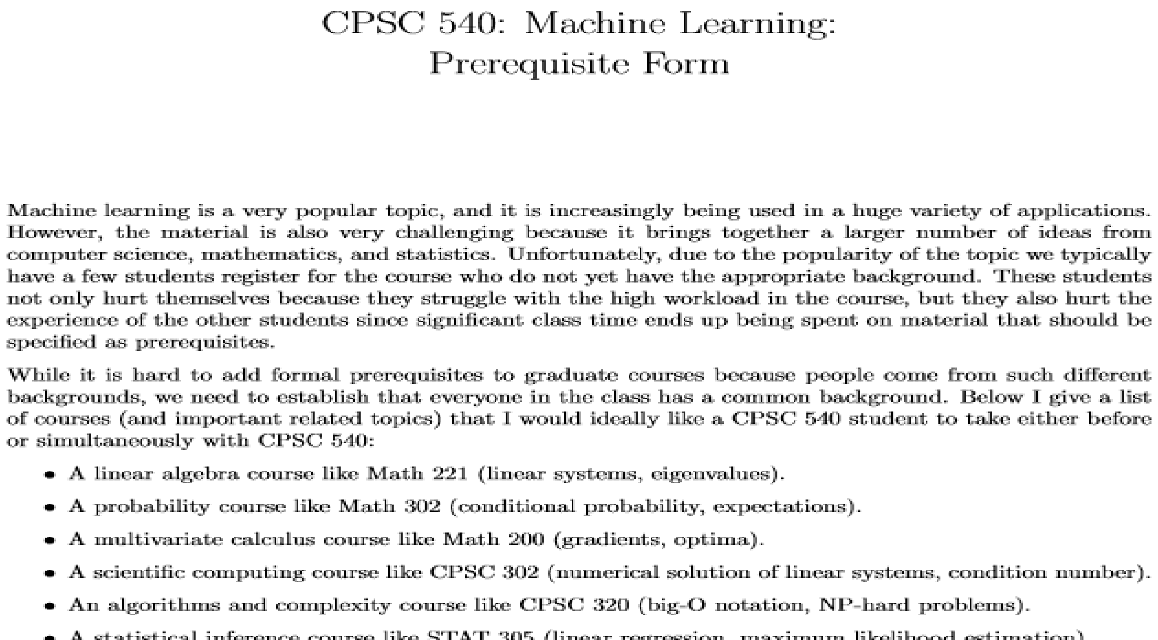
- This is not an introductory ML course.
  - CPSC 340 is a fast-paced 35-lecture course that skips a few details in order to cover the most fundamental and practically-useful topics.

- You should be comfortable with all topics in CPSC 340.
  - Cross-validation, generative models, non-parametric models, ensemble methods, non-parametric bases, stochastic gradient, kernel methods, maximum likelihood estimation, L1-regularization, softmax loss, PCA, sparse matrix factorization, collaborative filtering, multi-dimensional scaling, neural networks, deep learning, and so on.
- This course starts where CPSC 340 ends:
  - I'm not covering any of the above, and assume you already know these concepts.
  - If you don't know all the above , you will fall behind quickly and should instead take 340.

- Quotes from people who probably should have taken CPSC 340 first:
  - "I did Coursera and have have done well in Kaggle competitions."
  - "I've used SVMs, PCA, and L1-regularization in my work."
  - "I want to apply machine learning in my research."
  - "I took a machine learning course at my old school."

Prerequisite Form

- All students must submit the prerequisite form.
  - CS and ECE grad students: submit in class/tutorial by January 13.
  - All others: submit to enroll in course.
    - I’ll sign enrollment forms between lectures once I have this form.



CPSC 540: Machine Learning:
Prerequisite Form

Machine learning is a very popular topic, and it is increasingly being used in a huge variety of applications. However, the material is also very challenging because it brings together a larger number of ideas from computer science, mathematics, and statistics. Unfortunately, due to the popularity of the topic we typically have a few students register for the course who do not yet have the appropriate background. These students not only hurt themselves because they struggle with the high workload in the course, but they also hurt the experience of the other students since significant class time ends up being spent on material that should be specified as prerequisites.

While it is hard to add formal prerequisites to graduate courses because people come from such different backgrounds, we need to establish that everyone in the class has a common background. Below I give a list of courses (and important related topics) that I would ideally like a CPSC 540 student to take either before or simultaneously with CPSC 540:

- A linear algebra course like Math 221 (linear systems, eigenvalues).
- A probability course like Math 302 (conditional probability, expectations).
- A multivariate calculus course like Math 200 (gradients, optima).
- A scientific computing course like CPSC 302 (numerical solution of linear systems, condition number).
- An algorithms and complexity course like CPSC 320 (big-O notation, NP-hard problems).
- A statistical inference course like STAT 305 (linear regression, maximum likelihood estimation).

Reasons Not to Take This Course

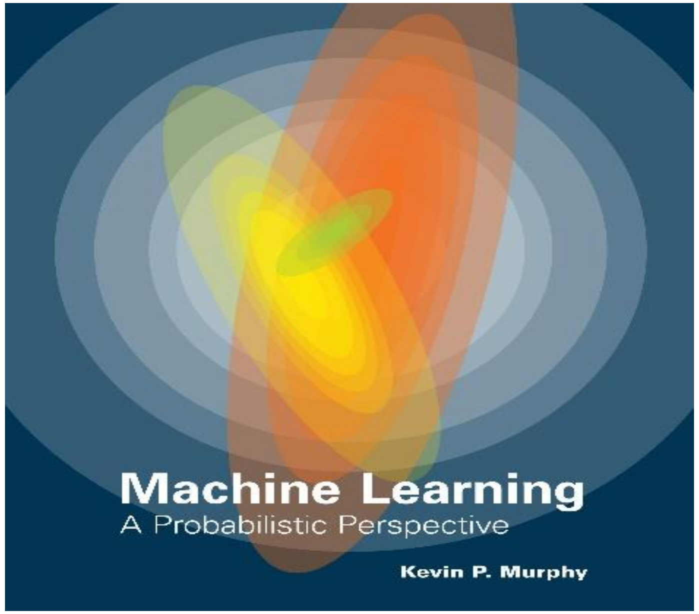
- High workload:
  - “This course's workload was a bit more than I would have liked. It seems like this course takes twice the amount of time as another course.”
- Inexperienced instructor:
  - Teachers improve the most over their first 3 years, I’m not there yet.
- Haven’t taken CPSC 340:
  - You’ll be missing half of the story, you won’t know many of the most important methods, and a lot of stuff will seem random.
- Missing prerequisites:
  - If you are missing MATH or CPSC prerequisites, it’s probably better to fill-in/strengthen your background first and then take this course.
  - “I know too much math” said nobody ever.

Auditing and Recording

- Auditing 540, an excellent option:
  - Pass/fail on transcript rather than grade.
  - Do 1 assignment or write a 2-page report on one technique from class or attend > 90% of classes.
  - But please do this officially:
    - http://students.ubc.ca/enrolment/courses/academic-planning/audit
- About recording lectures:
  - Do not record without permission.
  - All class material will be available online.
  - Videos of material from first month of last year's course are here:
    - https://www.youtube.com/watch?v=p4EnVHSml4U

Textbook and Other Optional Reading

- No textbook covers all course topics.
- The closest is Kevin Murphy’s “Machine Learning”.
  - But we’re using a very different order.



- For each lecture:
  - I’ll give relevant sections from this book.
  - I’ll give other related online material.
- There is a list of related courses on the webpage.

Grading

- Course grades will be split evenly between:
  - 5 assignments (written and Matlab programming).
  - Final (date will be placed here when known).
  - Course project (date will be placed here when known).

- A Matlab license is available for all UBC students:
  - https://it.ubc.ca/services/desktop-print-services/software-licensing/matlab

- No, you can’t do the assignments in Python, R, and so on.
  - You might be able to do them in Octave/Julia, but no guarantees.

Assignments

- Due any time on days where we have lectures:
  - A1: January 16 (1.5 weeks), February 6, February 27, March 15, April 3. (Due dates might be pushed back.)
- Start early, the assignments are a lot of work:
  - Previous students estimated that each assignments takes 6-25 hours:
    - The was heavily correlated with satisfying prereqs.
    - Please look through the assignment in previous offerings to see length/difficulty.
- You can do assignments in groups of 1 to 3.
  - Hand in one assignment for the group.
  - But each member should still know the material.

Late Assignment Policy

- You have up to 4 total “late classes”.
  - Cannot use more than 2 “late classes” on any one assignment.
  - Beyond 2 late classes for one assignment, or 4 total, you receive a 0.
  - You can use late days on the assignments/project, but not the exam.
- Number of late classes for a group:
  - If each member has c_i late classes, group can use at most ceil(mean(c_i)).
- Example:
  - Assignment 1 is due Monday January 16.
  - You can use 1 late class to hand it in January 18.
  - You can use 2 late classes to hand it in January 23.
  - If you need late days for Assignment 1, consider dropping the course.

Getting Help

- Piazza for assignment/course questions:
  - https://piazza.com/ubc.ca/winterterm22016/cpsc540
- Instructor office-hours/help-sessions:
  - Fridays 1:00-2:30 (ICICS 238) or by appointment (starting this week).
- Weekly tutorials:
  - Run by TAs covering related material.
  - Fridays 4:00-5:30 (DMP 101, starting next week).
- Teaching Assistants:
  - Jason Hartford.
  - Robbie Rolin.
  - Sharan Vaswani.

Exam and Course Project

- Final exam details:
  - Date will be written here, hopefully during exam period.
  - Closed book, four-page double-sided “cheat sheet”.
  - Given a list of things you need to know how to do.
  - Mostly minor variants on assignment questions.
  - No requirement to pass the final.
- Do not miss the final.

- Course projects can be done in groups of 2-3 and have 3 parts:
  1. Project proposal (due with Assignment 4).
  2. Literature review (due with Assignment 5).
  3. Coding, experiments, application, or theory (due late April).
     - More details coming later in term.