CPSC 440 and 540: Advanced Machine Learning

Mark Schmidt

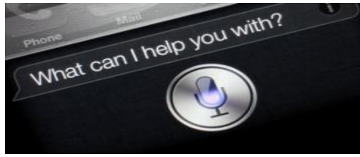
University of British Columbia, Winter 2021

www.cs.ubc.ca/~schmidtm/Courses/440-W21

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

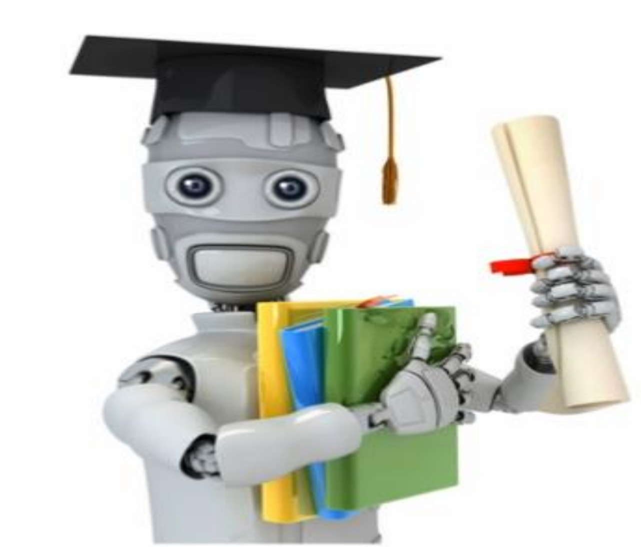
Big Data Phenomenon

- 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.



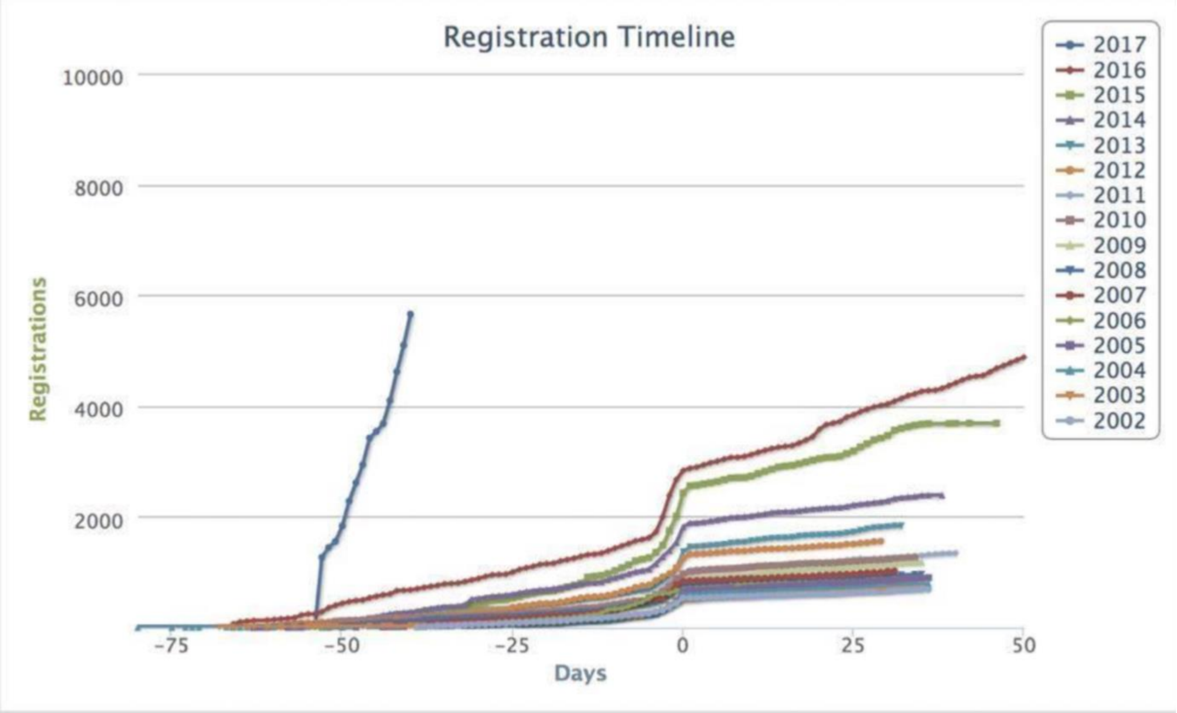
Machine Learning

- 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.



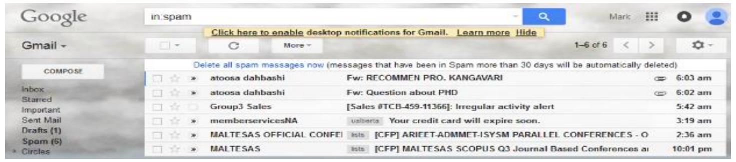
Machine Learning vs. Statistics

- 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:
  - 2018 NeurIPS Sold out in ~11 minutes.
    - 13 thousand registrations in 2020.
  - 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.

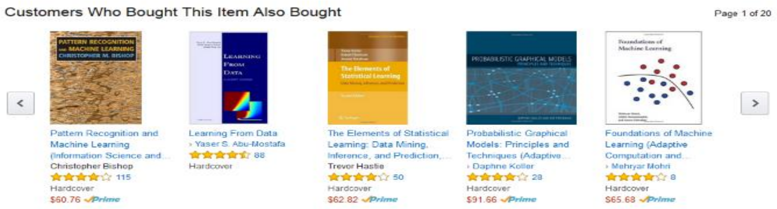


Click here to enable desktop notifications for Gmail. Learn more Hide

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 CONFEI	lists [CFP] ARIEET-ADMMET-ISYSM PARALLEL CONFERENCES - O	2:36 am
MALTESAS	lists [CFP] MALTESAS SCOPUS Q3 Journal Based Conferences ai	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

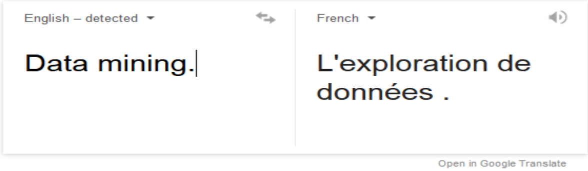
Pattern Recognition and Machine Learning (Information Science and... Christopher Bishop ★★★★☆ 115 Hardcover $60.76 ✓Prime

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

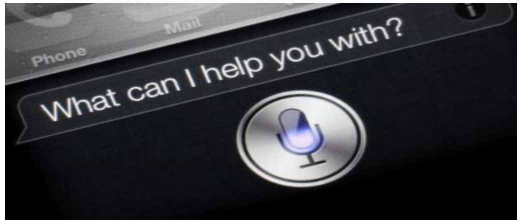
Foundations of Machine Learning (Adaptive Computation and... › Mehryar Mohri ★★★★☆ 8 Hardcover $65.68 ✓Prime

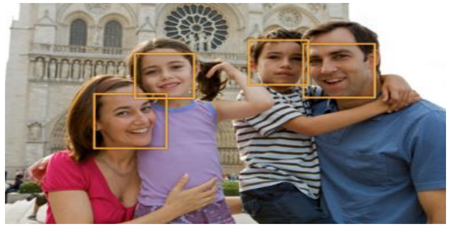


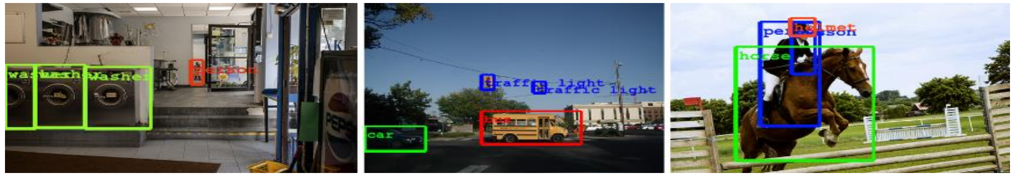
Data mining.

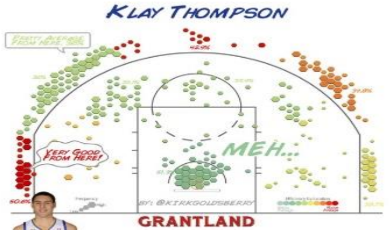
L'exploration de données .

Open in Google Translate



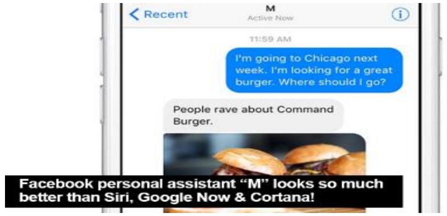






Applications

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

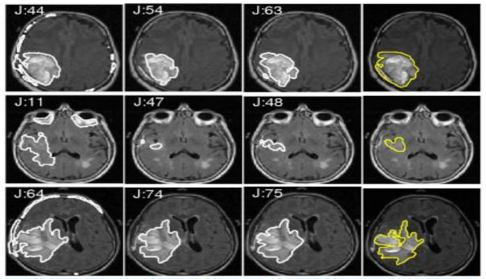


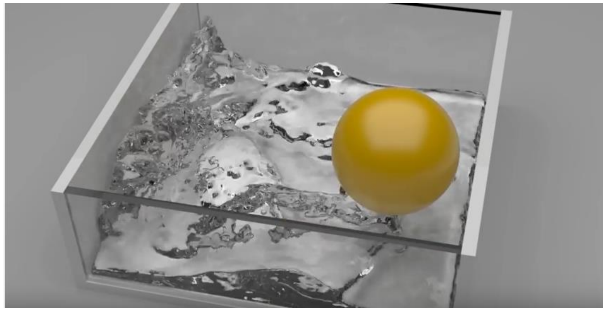
11:59 AM

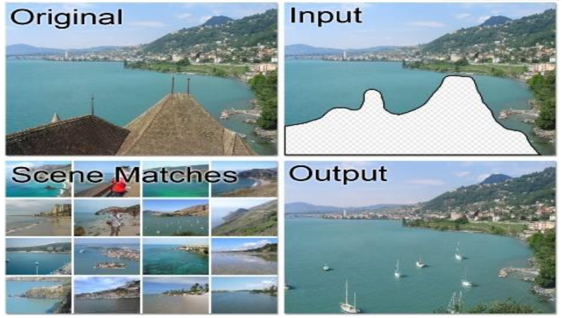
I'm going to Chicago next week. I'm looking for a great burger. Where should I go?

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

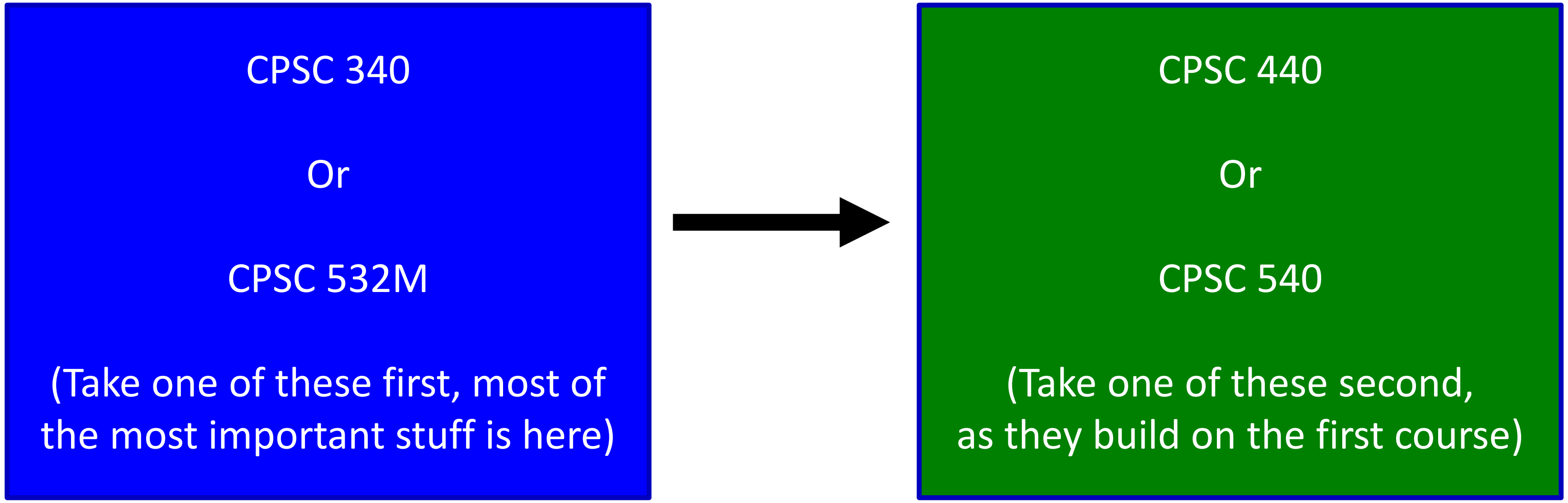
Youngsters, May 1912

Horizon
Trees
Leaves
Towers & Pagodas
Buildings
Birds & Insects

(pause)

CPSC 340 and CPSC 440 (and 532M and 540)

- There are two "core" ML classes: CPSC 340 and CPSC 440.
  - You can take CPSC 340 for grad credit as CPSC 532M (though not this year).
  - You can take CPSC 440 for grad credit as CPSC 540.
  - Structured as one full-year course: 440/540 starts where 340/532M ends.



CPSC 340

Or

CPSC 532M

(Take one of these first, most of the most important stuff is here)

CPSC 440

Or

CPSC 540

(Take one of these second, as they build on the first course)

CPSC 340 and CPSC 440 (and 532M and 540)

- CPSC 340/532M (the other course):
  - Introductory course on data mining and ML.
  - Emphasis on applications and core ideas of ML.
  - Covers implementation of methods based on counting and gradient descent.
  - Most useful techniques that you can apply to your research/work.
- CPSC 440/540 (this course):
  - Research-level ML methods (“what is in papers” vs. “in applications”).
  - More “what might be used in the future” than “what is used now”.
  - Assumes strong background on fundamental ML concepts.
  - Assumes stronger math/CS background

CPSC 340 and CPSC 440 (and 532M and 540)

- If you can only take one class, take the other class (CPSC 340/532M).
  - 340/532M covers the most useful methods and ideas.
  - If you take 440/540 first, you’ll miss half the story and a lot will seem random.
    - This is not an intro course and there is not a lot of review in 440/540.
      - So 440/540 is missing a lot important topics.
  - 440/540 is NOT an “advanced” version of 340/532M.
    - It just covers the methods that require more advanced math/CS background to use.

- It is much better to do CPSC 340/532M first:
  - Many people have taken 340/532M *after* 440/540 (not recommended).
  - A few people took 440/540 then 340/532M then *sat in on 440/540 again, (REALLY not recommended).

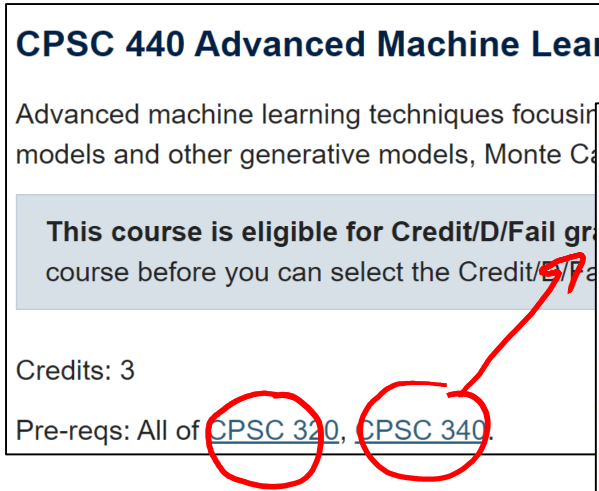
CPSC 340 and CPSC 440 (and 532M and 540)

- I’m not covering any of the below “typical” topics, and assume you already know them:
  - Calculus in matrix notation, including derivation of normal equations for least squares.
  - IID assumption, complexity vs. generalization trade-off, ensemble methods, and cross-validation.
  - Probabilistic classifiers, maximum likelihood, and MAP estimation.
  - Radial basis functions, how to show a function is convex, and the kernel trick.
  - Stochastic gradient descent, softmax loss, and L1-regularization.
  - PCA and collaborative filtering.
  - Deep learning and convolutional neural networks.

- **You will get lost very quickly if you don’t know this material**.
  - You should already be able to write code implementing all of the above ideas.

- CPSC 440/540 Course Outline:
  - Density estimation, graphical models, prediction with structured data, Bayesian methods.

Prerequisites (A lot of Math and CS)



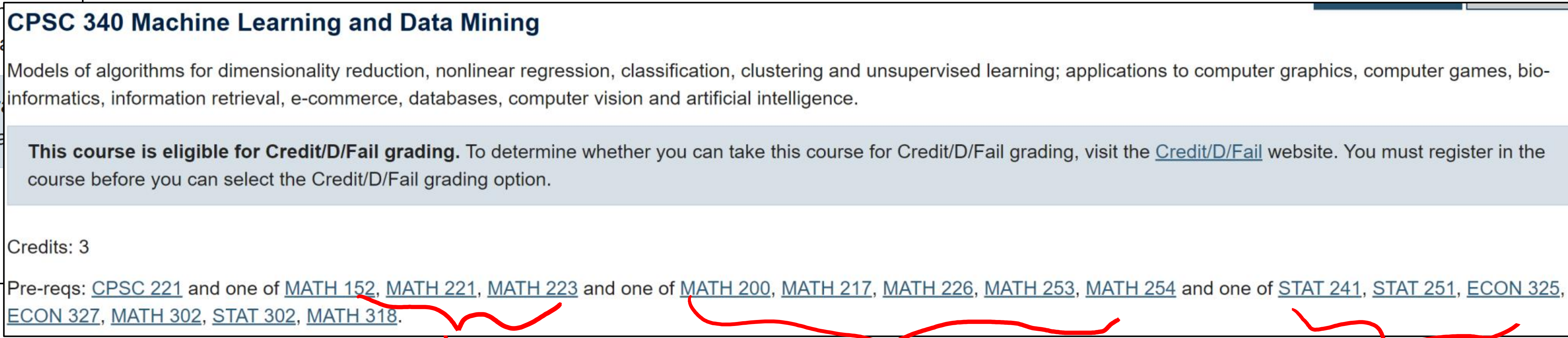
CPSC 440 Advanced Machine Learning

Advanced machine learning techniques focusin
models and other generative models, Monte Ca

This course is eligible for Credit/D/Fail gr
course before you can select the Credit/D/Fa

Credits: 3

Pre-reqs: All of CPSC 320, CPSC 340.



CPSC 340 Machine Learning and Data Mining

Models of algorithms for dimensionality reduction, nonlinear regression, classification, clustering and unsupervised learning; applications to computer graphics, computer games, bio-informatics, information retrieval, e-commerce, databases, computer vision and artificial intelligence.

This course is eligible for Credit/D/Fail grading. To determine whether you can take this course for Credit/D/Fail grading, visit the Credit/D/Fail website. You must register in the course before you can select the Credit/D/Fail grading option.

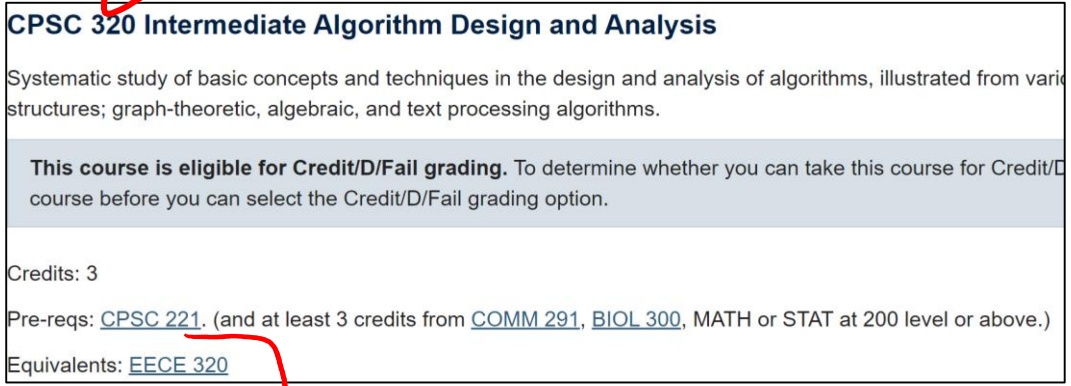
Credits: 3

Pre-reqs: CPSC 221 and one of MATH 152, MATH 221, MATH 223 and one of MATH 200, MATH 217, MATH 226, MATH 253, MATH 254 and one of STAT 241, STAT 251, ECON 325, ECON 327, MATH 302, STAT 302, MATH 318.

linear algebra

multivariate calculus

probability meets calculus



CPSC 320 Intermediate Algorithm Design and Analysis

Systematic study of basic concepts and techniques in the design and analysis of algorithms, illustrated from vari
structures; graph-theoretic, algebraic, and text processing algorithms.

This course is eligible for Credit/D/Fail grading. To determine whether you can take this course for Credit/D
course before you can select the Credit/D/Fail grading option.

Credits: 3

Pre-reqs: CPSC 221. (and at least 3 credits from COMM 291, BIOL 300, MATH or STAT at 200 level or above.)

Equivalents: EECE 320

Basic Algorithms and data structures

Examples of CS concepts you should know:

- writing/debugging complex programs, binary trees, hash functions, graphs, big-O, randomized algorithms, dynamic programming, NP-completeness.

Examples of math concepts you should know:

- matrix algebra, norms, gradients, random variables, expectations, minimizing quadratic functions, random vectors.

Auditing

- 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

- Auditing 440:
  - There are only 120 “seats”.
  - If these are full, we won’t allow auditors.
  - If these are not full, see above for how we’ll deal with 540 auditors.

(pause)

Grading

- 40%: 4 assignments (written, math, and Julia programming).
- 30%: Final (date unknown).
- 30%: Course project (due date near end of exam period).
  - Subject to reasonable changes (last year the final ended up being optional, may not happen again).
  - There will be no post-course grade changes based on grade thresholds:
    - 48% will not be rounded to 50%, and 70% will not be rounded to 72%, and so on.

- Don’t expect a high grade without a high effort.
  - We cover a lot of material and my assignments are LONG. This is not an “easy” class.

- No, you can’t do the assignments in Python, R, Matlab, and so on.
  - Julia is free and way faster than Python/R/Matlab.
  - Assignments have prepared code that we won’t translate to 3 languages.
  - TAs shouldn’t have to know 3 languages to grade.
  - It’s important to know how to learn a new language (probably won’t always use the same language).
- For the course project, you can use any language.

Assignments

- Due at midnight on days where we have lectures:
  - First assignment due Friday of next week.
    - Subsequent assignments due every ~3-4 weeks.

- Start early, the assignments are a lot of work:
  - Previous students estimated that each assignments takes 6-25 hours:
    - This was heavily correlated with satisfying prerequistes.
    - Please look through the assignment in previous offerings to see length/difficulty.

- Assignment 1 should be done on your own.
  - It is mostly review of material from the prereq classes.
  - If you find takes a ton of time,
- Assignments 2-4 can be done 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” for Assignments 2+.
- Example:
  - Assignment 2 is due on a Friday.
  - You can use 1 late class to hand it in the following Monday.
  - You can use 2 late classes to hand it the following Wednesday.

- FAQ:
  - You cannot use “late classes” on Assignment 1 or exams.
  - You can use “late classes” on Assignments 2-4 and the project.
  - You cannot use more than 2 “late classes” on any one assignment (0 after that).
  - You cannot use more than 4 total “late classes” throughout the term (0 after that).
    - Otherwise, there is no penalty for using “late classes”.
  - Number of late classes for a group:
    - If group member ‘i’ has c_i late classes, group can use at most ceil(mean(c_i)).

Assignment Issues

- No extensions will be considered beyond the late days.
  - Also, since you can submit more than once, you have no excuse not to submit something preliminary by the deadline.

- Due to limited TA hours, issues like the below are a 50% penalty:
  - We can’t easily figure out who submitted an assignments.
  - Corrupted submission files or not using correct submission format.
  - Submitting the wrong assignment (year or number).
  - Not including answers in the correct location in the .pdf file.

Cheating and Plagiarism

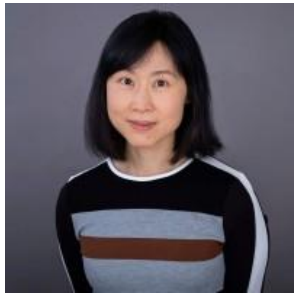
- Read about UBC’s policy on “academic misconduct” (cheating):
  - http://www.calendar.ubc.ca/Vancouver/index.cfm?tree=3,54,111,959

- When submitting assignments, acknowledge all sources:
  - Put “I had help from Lucy on this question” on your submission.
  - Put “I got this from another course’s answer key” on your submission.
  - Put “I copied this from the Coursera website” on your submission.
  - Otherwise, this is plagiarism (course material/textbooks are ok with me).

- At Canadian schools, this is taken very seriously.
  - Could receive 0 in course, be expelled from UBC, or have degree revoked.

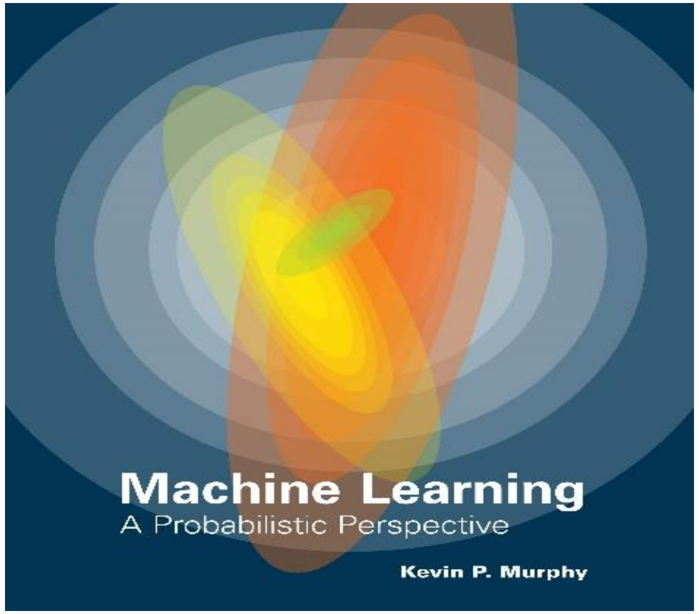
Getting Help

- We will use Piazza for assignment/course questions:
  - Link on Canvas.
  - Private posts asking about general information will be made public without asking.
- Weekly or almost-weekly Tutorials:
  - Run by TAs covering related material, mainly to help with assignments.
  - They are Thursdays, optional, and starting this week?
- Instructor and TA office-hours:
  - Schedule on Canvas (starting this week).
- Teaching Assistants:
  - Setareh Cohen
  - Peyman Gholami
  - Nam Hee Gordon Kim
  - Frederik Kunstner
  - Shahriar Shayesteh
  - Betty Shea



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.
  - We may get access to the new version.



- 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.

Bonus Slides

- I will include a lot of “bonus slides”.
  - May mention advanced variations of methods from lecture.
  - May overview big topics that we don’t have time for.
  - May go over technical details that would derail class.

- You are not expected to learn the material on these slides.
  - But you may find them interesting or useful in the future.

- I’ll use a different colour of background on bonus slides.
  - I often include “post-lecture” bonus slides after the “Summary” slide.

Textbook and Other Optional Reading

- Other good machine learning textbooks:
  - All of Statistics (Wasserman).
  - Elements of Statistical Learning (Hastie et al.).
  - Pattern Recognition and Machine Learning (Bishop).

- Good (online) textbook covering needed mathematical background:
  - Mathematics for Machine Learning (Deisenroth, Faisal, Ong).

- Good textbooks on specialized topics from this course:
  - Probabilistic Graphical Models (Koller and Friedman).
  - Deep Learning (Goodfellow et al.).
  - Bayesian Data Analysis (Gelman).

Final Exam

- Final exam:
  - Will probably be online and open book.
  - No requirement to pass the final (but recommended).
  - Will be scheduled by UBC.
    - Make sure you have internet access through April 29th.

- In the offline world, I used to use two types of questions:
  - 'Technical' questions requiring things like pseudo-code or derivations.
    - On topics covered in assignments (similar to assignment questions).
  - 'Conceptual' questions testing understanding of key concepts.
    - All lecture slide material except "bonus slides" is fair game here.

Course Project

- Course projects can be done in groups of 2-3.
  - More details coming later in the term.
  - Project scope will be smaller than projects in most classes.

Lectures

- All slides will be posted online (before lecture, and final version after).

- Please ask questions: you probably have similar questions to others.
  - I may deflect to the next lecture or Piazza for certain questions.

- Be warned that the course will move fast and cover a lot of topics:
  - Big ideas will be covered slowly and carefully.
  - But a bunch of other topics won’t be covered in a lot of detail.

- Isn’t it wrong to have only have shallow knowledge?
  - In this field, it’s better to know many methods than to know 5 in detail.
    - This is called the “no free lunch” theorem: different problems need different solutions.
    - If you know why something is important, and the core ideas, you can fill in details later.

Warning regarding teaching quality

- First time that CPSC 440 has been offered.
  - And I didn’t have as much time to prepare as a wanted.
    - Got mysterious sickness last March, then lungs hurt until mid-November.
  - So the course isn’t as “put together” as you might like.
    - I have material covering a bunch of relevant topics, so you will definitely learn things.
    - But it won’t be as smooth as a course that has been offered multiple times.
      - Switching from PowerPoint to Beamer and back, assignments may not sync well with lectures, some course material will be underdeveloped, and so on.
  - I’m also not a teaching faculty member.
    - I run one of the typically-largest labs in CS (e.g., I have 5 PhD students and median is 1).
    - I try but may not be as available/good as some the full-time teachers.

- If these things are going to bother you, it might be better to take this course later and/or take a different course this term.

Summary

- Machine learning: automatically detecting patterns in data to help make predictions and/or decisions.
- CPSC 440: advanced/difficult graduate-level 2nd or 3rd+ course on this topic.
  - Also called CPSC 540 for grad students.
- Course admin: these slides are the syllabus!

- Next time: review of 340, and filling in some theory gaps.

- UBC provides resources to support student learning and to maintain healthy lifestyles but recognizes that sometimes crises arise and so there are additional resources to access including those for survivors of sexual violence. UBC values respect for the person and ideas of all members of the academic community. Harassment and discrimination are not tolerated nor is suppression of academic freedom. UBC provides appropriate accommodation for students with disabilities and for religious and cultural observances. UBC values academic honesty and students are expected to acknowledge the ideas generated by others and to uphold the highest academic standards in all of their actions. Details of the policies and how to access support are available here: https://senate.ubc.ca/policies- resources-support-student-success

CPSC 340 and CPSC 440 (and 532M and 540)

- Quotes from people who probably should have taken CPSC 340:
  - "I did Coursera [or other online class] and have have done well in Kaggle competitions."
    - Neither of these cover calculus in matrix notation or MLE and MAP estimation.
  - "I've used SVMs, PCA, and L1-regularization in my work."
    - Sure, but do you know how to implement them from scratch?
  - "I've seen most of the 340 topics before."
    - Sure, but at what level of detail and do you know how to implement them from scratch?
  - "I want to apply machine learning in my research."
    - Great! Take 340 to learn how the most useful tools work and also **what can go wrong**.
  - "I took a machine learning course at my old school."
    - 340 is more broad/advanced than at most schools (talk to me if unsure).
  - "I've already learned about deep learning, so can I skip the basic stuff?"
    - When something goes wrong, you are going to want to understand the fundamentals.
  - " I took CPSC 540 with you. I wish I would have taken CPSC 340 first."
    - From a really-smart person who was working in a machine learning research job at the time.