CIS 520 Machine Learning

Lyle Ungar

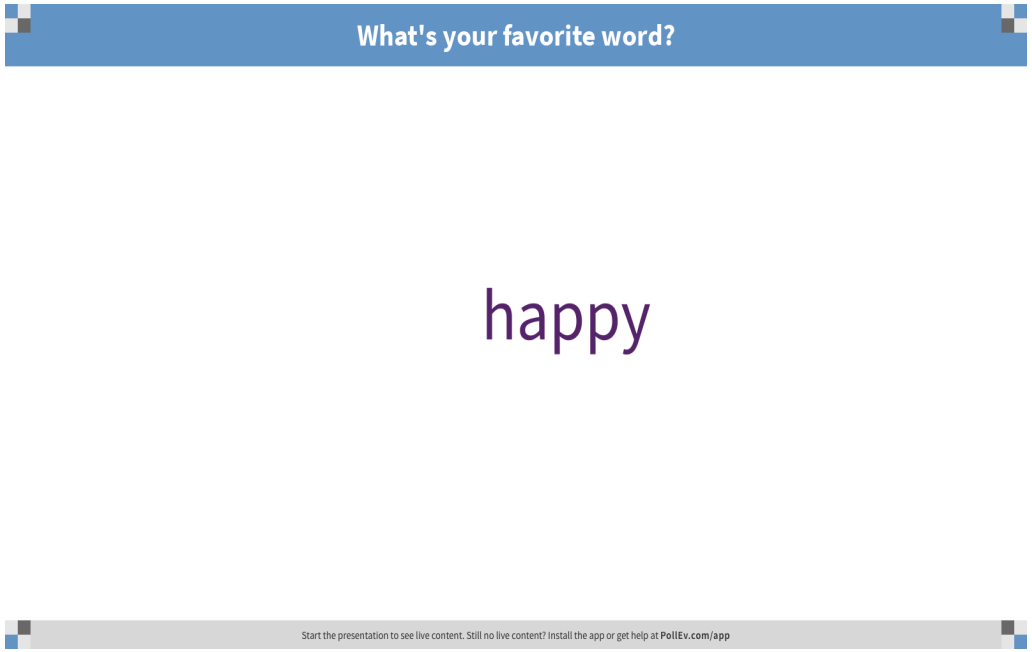
Poll Everywhere

Poll Everywhere, Inc. Communication

Everyone

This app is compatible with all of your devices.

Install *Poll Everywhere* from app store
or go to
https://pollev.com/lyleungar251



CIS 520 Machine Learning

Lyle Ungar

Computer and information Science

Learning Objectives

Is CIS520 for you?

What you need to know for 520

Types of machine learning

Should I be here?

- **You should know probability and linear algebra**
  - See prequiz on canvas
- **If you're waiting to get into this course**
  - Only via https://forms.cis.upenn.edu/waitlist/
  - The course will be offered again in the spring (not by me)
- **Alternate courses**
  - CIS 419/519 **Applied Machine Learning** less math
  - STAT 471/571/701 **Modern Data Mining** in R
  - CIS 545: **Big Data Analytics** more data handling
  - ESE 545: **Data Mining** more math?

Introductions

- **Who am I?**
- **Who are you?**
  - Why are you here?

Breakout room

Introductions

Why are you taking this course? What do you want from it?

What will this course look like?

- **Lectures (MWF)** - synchronous and recorded on canvas
  - Slides, poll-everywhere, wiki
- **Office hours: see "people" on the wiki**
- **Worksheets**
- **Homework**
  - **Conceptual** (math in latex - **overleaf**) and
  - **Coding** (python/numpy/sklearn/tensorflow/jupyter - **colab**)
  - Canvas (out) and Gradescope (in)
- **Exams**
  - Midterm and final
- ***Quiz, Feedback – each week on canvas***
- ***Evolving over the semester, so lots of feedback to me!!!***

Course goals

- **Be familiar with all major ML methods**
  - Regression (linear, logistic), regularization, feature selection
  - K-NN, Decision trees, random forests, SVM
  - PCA, K-means, GMM
  - Naive Bayes, Bayes Nets, Markov Nets, HMMs
  - Online learning: boosting, perceptrons, LMS
  - Deep learning
- **Know their strengths and weaknesses**
  - know jargon, concepts, theory
  - be able to modify and code algorithms
  - be able to read current literature

Administrivia

- **Canvas**
  - Homework, Lecture recordings, quizzes
- **Gradescope**
- **Course wiki**
  - Lecture notes, slides
  - Resources
    - Grading scheme, academic integrity,
    - office hours, …
  - Readings -- including the Bishop ‘textbook’ – free online
    - Mostly for reading after lectures
    - ”supplemental” really means that
- **Piazza**
  - *look here first for answers!*

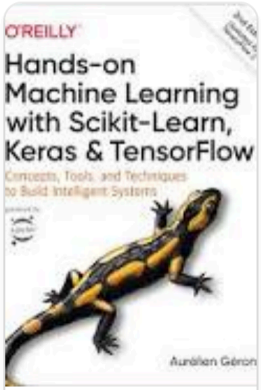
Textbooks

Google

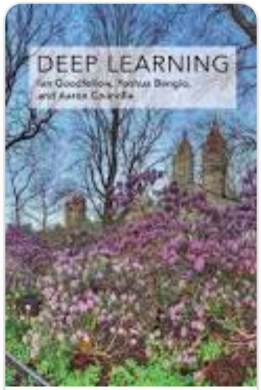
machine learning books

All Books Shopping News Images More Settings Tools

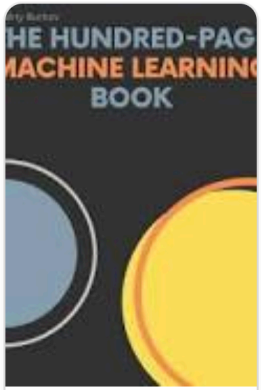
Books / Machine learning



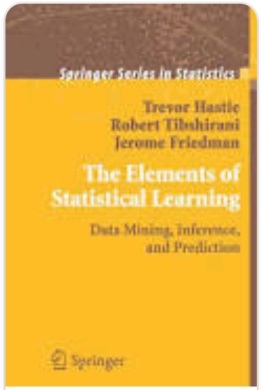
Hands-On Machine Lea...
Aurelien Ger...



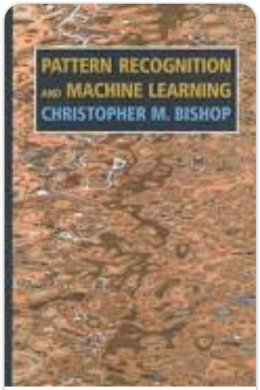
Deep Learning
2015



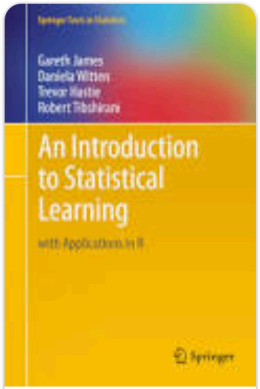
The Hundred-Page Machin...
Andriy Burko...



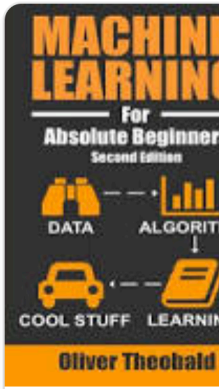
The Elements of Statistical ...
2001



Pattern Recognition ...
Christopher ...



An Introduction ...
2013



Machine Learning for ...
O. Theobald,...

Learning in the time of COVID

- **This course is in *beta***
  - Mix of synchronous and asynchronous.
  - Give me lots of feedback!!!!

- **Let me know if you experience challenges**

I care!!!

Do you have Poll Everywhere?

A) Yes

B) No

Install *Poll Everywhere* from app store
or go to
https://pollev.com/lyleungar251

Also remember the zoom chat window

Working Together

Homework is mostly "pair programming" and "pair problem solving"

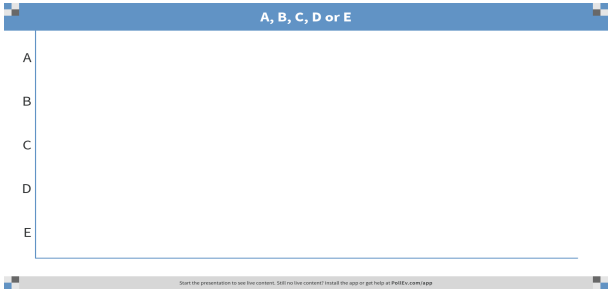
If it is determined that code submitted by two students might have been copied

A) Both will receive half credit

B) The person who copied will be referred to the Office of Student Conduct (OSC)

C) Both students will be referred to the Office of Student Conduct (OSC)

D) None of the above



Asking Questions

- **Questions about homework should be**
  - A) Asked during office hours
  - B) Emailed to the instructor or a TA
  - C) Asked on piazza
  - D) A or C
  - E) None of the above

Python

- **Python is a better ML language than matlab**
  - A) True
  - B) False

Where is Machine Learning used?

https://alliance.seas.upenn.edu/~cis520/wiki/

EMC, Teradata, Oracle, SAP, Vmware, Splunk, MemSQL, Palantir, Trifacta, Datameer, Neo,, Infobright, Fractal Analytics

http://www.datamation.com/applications/30-big-data-companies-leading-the-way-1.html

ML unicorns: business

- 4Paradigm — Anti-fraud for insurance & banking — China
- Dataminr — Business intelligence — US
- Afiniti — Behavior analytics — US
- InsideSales.com — Platform for sales teams — US
- Avant — Credit scores — US
- ZipRecruiter — Recruitment platform — US
- SoundHound — Voice-enabled AI assistants — US
- Momenta — AV perception software — China
- Bytedance — Personalized news curation — China

https://www.cbinsights.com/research/ai-unicorn-club/

ML: cybersecurity, surveillance

- CrowdStrike — Cybersecurity US
- Darktrace — Cybersecurity UK
- Tanium — Cybersecurity US
- Face++ — Facial recognition China
- SenseTime — Facial recognition China
- Cloudwalk — Facial recognition China
- YITU Technology — Facial recognition China
  medical imaging & diagnostics

https://www.cbinsights.com/research/ai-unicorn-club/

ML: healthcare, drugs

- iCarbonX Personalized healthcare China
- Tempus Labs Drug R&D US
- BenevolentAI Drug R&D UK
- Butterfly Network Portable ultrasound US
- OrCam Technologies Wearables for visually impaired Israel

https://www.cbinsights.com/research/ai-unicorn-club/

ML: manufacturing

- Preferred Networks Mfg, medical imaging & diagnostics, auto Japan
- Automation Anywhere Robotic process automation US
- UiPath Robotic process automation US
- C3 IIoT platform US
- Uptake Technologies IIoT platform US

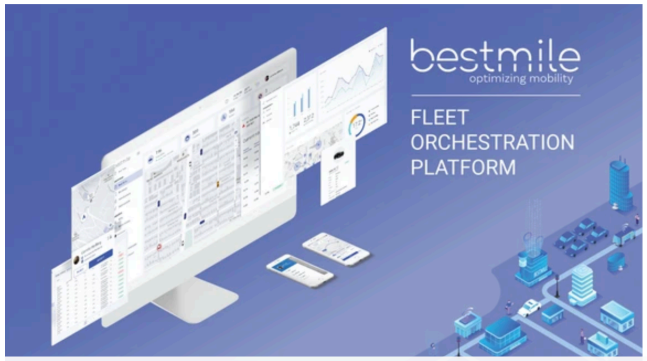
https://www.cbinsights.com/research/ai-unicorn-club/

ML: Automomous vehicles

- Pony.ai Autonomous vehicles US
- Zoox Autonomous vehicles US

Bestmile raises $16.5 million to optimize autonomous vehicle fleets

CHRIS O'BRIEN @OBRIEN AUGUST 28, 2019 12:08 AM



MOST READ

iOS 13.1 Developer beta
Apple Inc.
451 MB

This beta version of iOS 13 contains bug fixes and improvements.

For more information, visit:
https://developer.apple.com/go/?id=ios-13-sdk-rn

Apple unexpectedly releases iOS 13.1 developer beta before finalizing iOS 13

https://www.cbinsights.com/research/ai-unicorn-club/

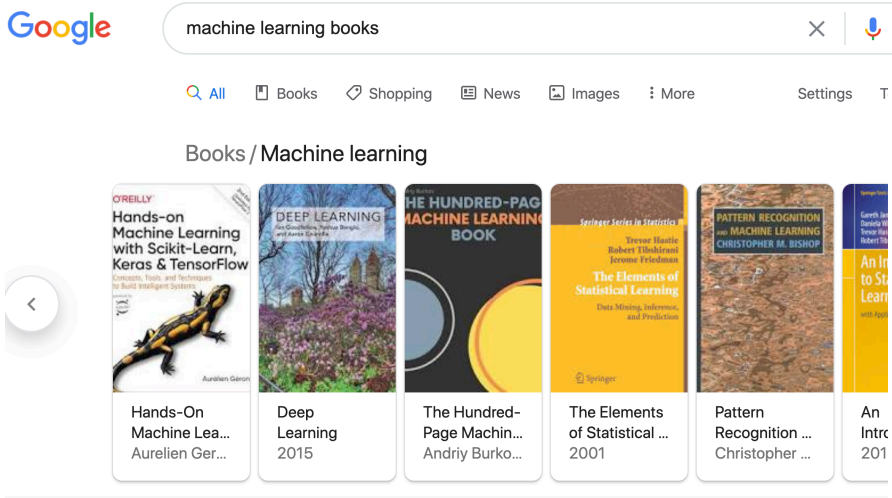
Components of ML

- ***Representation***
  - feature set
  - model form
- ***Loss function***
- ***Optimization method***
  - For parameter estimation
  - For model selection and hyperparameter tuning

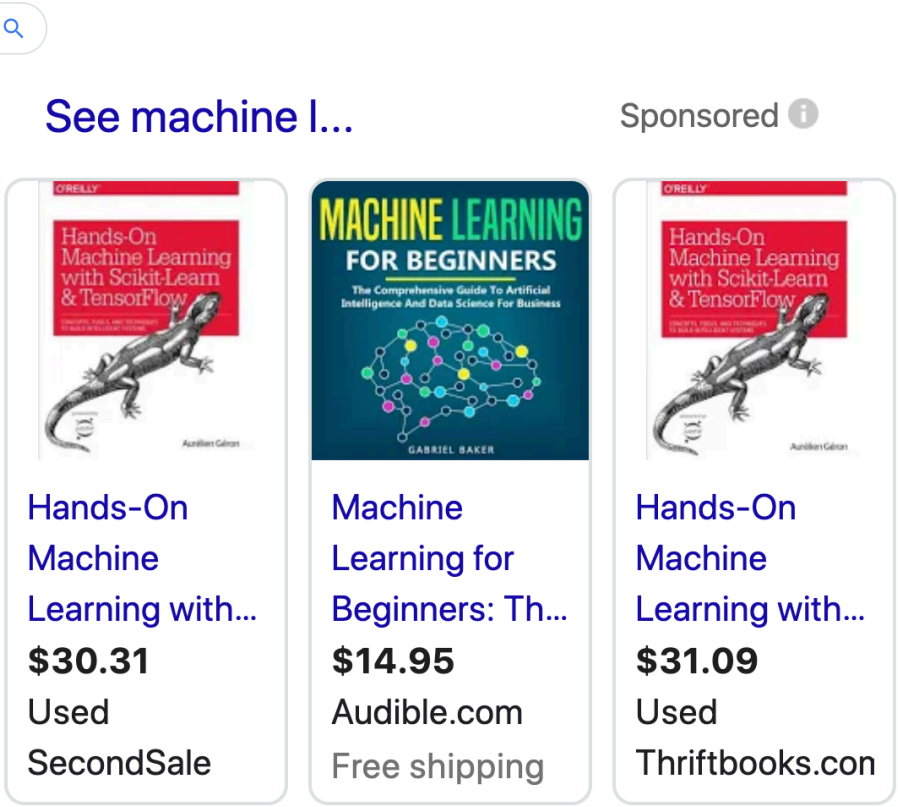
Components of ML

- ***Representation***
  - $\hat{y} = f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T\mathbf{x}$
- ***Loss function***
  - $\mathrm{L}(\mathbf{y}, \hat{\mathbf{y}}) = \|\mathbf{y} - \hat{\mathbf{y}}\|_2$
- ***Optimization method***
  - $\mathrm{argmin}_w L(y, \hat{y}(w))$
  - gradient descent

Google ads as machine learning



What features?
What model?
What loss function?



Types of Learning

- **supervised** **X, y**
  - Given an observation ***x***, what is the best label *y*?
- **unsupervised** **X**
  - Given a set of ***x***'s, cluster or summarize them
- **reinforcement**
  - Given a sequence of states ***x*** and possible actions ***a***, learn which actions maximize reward.

Types of Learning as Probabilities

- **supervised** ***X, y***
  - $p(y|x)$ - conditional probability estimation
  - $min \; || \hat{y}(x) - y ||$ - optimization
- **unsupervised** **X**
  - $p(x)$ - "generative" model

Types of models

- **Generative**
  - $p(\boldsymbol{x})$
- **Discriminative**
  - $p(y|\boldsymbol{x})$

X: features, predictors, design matrix, input

y: response, label, output

Types of models

- **Parametric**
  - $\hat{y} = \boldsymbol{w} \cdot \boldsymbol{x}$
  - $\hat{y} = f(\boldsymbol{x}; \theta)$
  - $\boldsymbol{w}$ and θ are parameters
- **Non-parametric**
  - k-nn, decision trees
- **"Semi-parametric"**
  - Deep learning

Consider the Netflix problem

- Given a list of people and the ratings they have given movies, predict their ratings on other movies
- What type of learning is this?
  - A) supervised
  - B) unsupervised
  - C) something else

Breakout room

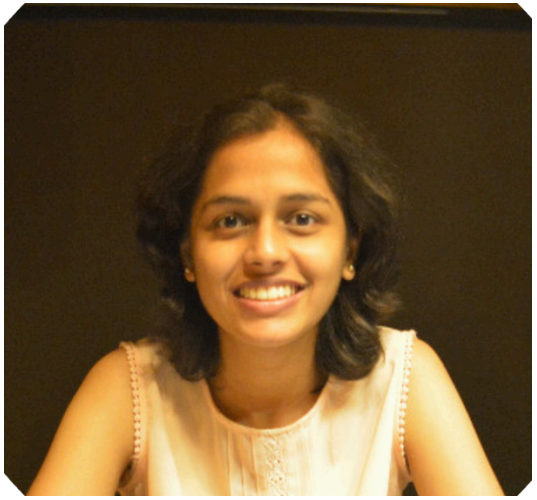
- How might you go about solving it?

ML vs. Statistics vs. Data Science

- **Statistics**
  - more modeling, especially of the noise
  - more hypothesis testing
- **ML**
  - more predictive accuracy
  - more flexible model forms
- **Data Science**
  - Includes data collection and cleaning
  - More interpretation, less math

A few words from a former student

- Pooja Consul



TODO

- **Visit canvas**
  - https://canvas.upenn.edu/
  - Do HW 0 (trivial latex; be able to run numpy in jupyter)
- **Join piazza**
  - Linked to from canvas and the course wiki
  - https://alliance.seas.upenn.edu/~cis520/wiki
- **Take the self-test in canvas**
  - Make sure you know enough linear algebra and probability
- **Get up to speed on python, numpy (for Friday!)**

What you should know

- **Turning a real-world problem into a well-posed ML problem is often hard**
  - pick features/predictors (**x**, y) and loss function
- **Unsupervised vs. supervised vs. reinforcement**
  - generative $p(\boldsymbol{x})$ vs. conditional $p(y|\boldsymbol{x})$ models
- **Parametric, non-parametric, semi-parametric**
  - Parameters vs. hyper-parameters
- **Canvas, piazza, wiki**

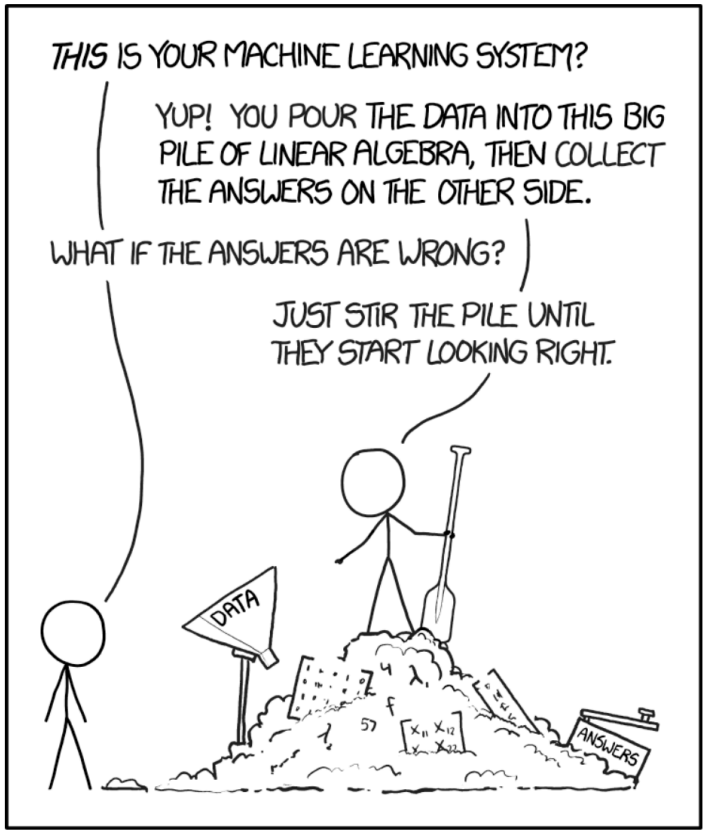
What questions do you have on today's class?

Top

What could we have done better? (including technology)

Top

Machine Learning



Permanent link to this comic: https://xkcd.com/1838/