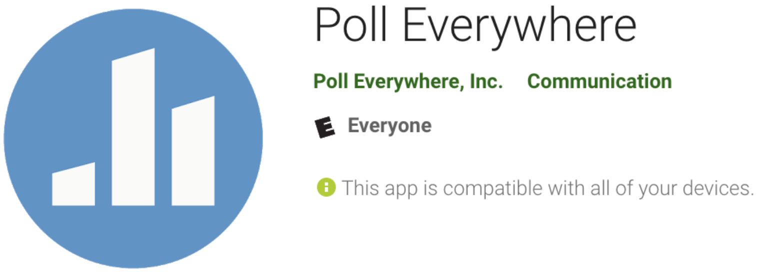
CIS 5200 Machine Learning

Lyle Ungar

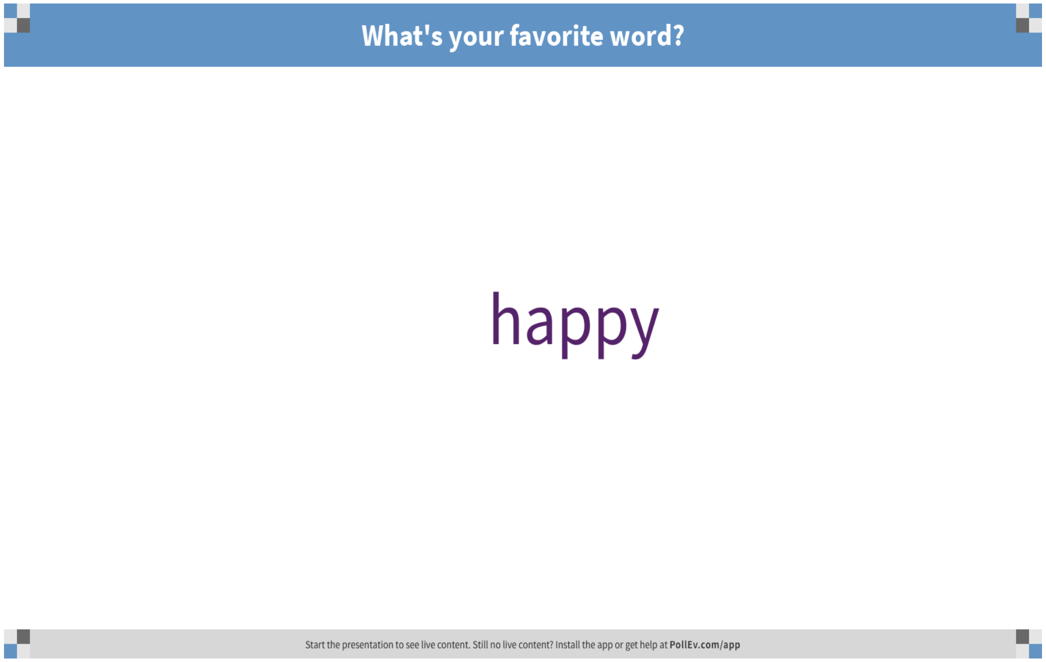


Poll Everywhere

Poll Everywhere, Inc. Communication

Everyone

This app is compatible with all of your devices.



What's your favorite word?

happy

Start the presentation to see live content. Still no live content? Install the app or get help at **PollEv.com/app**

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

CIS 5200 Machine Learning

Lyle Ungar

Computer and information Science

Learning Objectives

Is CIS5200 for you?

What you need to know: Administrivia and Course Goals

Types of machine learning

Should I be here?

- **You should know probability and linear algebra**
  - See prequiz from wiki
- **If you're waiting to get into this course**
  - The course will be offered again in the spring
- **Alternate courses**
  - CIS 4190/5190 **Applied Machine Learning** less math
  - STAT 4710/5710/7010 **Modern Data Mining** in R
  - CIS 5450 **Big Data Analytics:** more data handling
  - ESE 5450 **Data Mining** more math?

Introductions

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

What will this course look like?

- **Lectures (MW) Review (F)** – live, livestreamed, and recorded on canvas
  - Slides, poll-everywhere, wiki
- **Pods (Wθ)** - start next week
  - Mandatory attendance;
- **Office hours**: see “people” on the wiki
- **Ed** – first stop for questions
- **Worksheets** – Jupyter notebooks for code
- **Homework**
  - **Conceptual** (math in latex - **overleaf**) and
  - **Coding** (python/numpy/sklearn/pytorch/jupyter - **colab**)
  - Submit via Gradescope
- **Exams**
  - Midterm and final – multiple choice with “cheat sheet”
- ***Quizzes, Surveys– each week on canvas***
- ***Evolving over the semester, so lots of feedback to me!!!***

The Course Cadence

- **MW Lecture:** new material
- **Wθ Pods:** discuss
- **F: Review**
- **θFSSMT:** quiz, survey, Worksheets, HW for preceding week

The course moves fast;
you need to keep up!

Pods

- Meet weekly, mandatory attendance
  - Get to know people!
- How do I sign up?
  - Coming this weekend
- What do I do if I can't make my pod?
  - Let your pod leader know
  - Come to make-up

The Course Philosophy

- **Understand the huge number of math concepts behind machine learning**
  - Lecture/quiz/midterm/final
- **Be familiar with the standard ML coding platforms**
  - Worksheets/HW

If Worksheets are taking more than 5 hours/week,
then you should be doing them during special “pod
hours” on the weekend.

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, 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

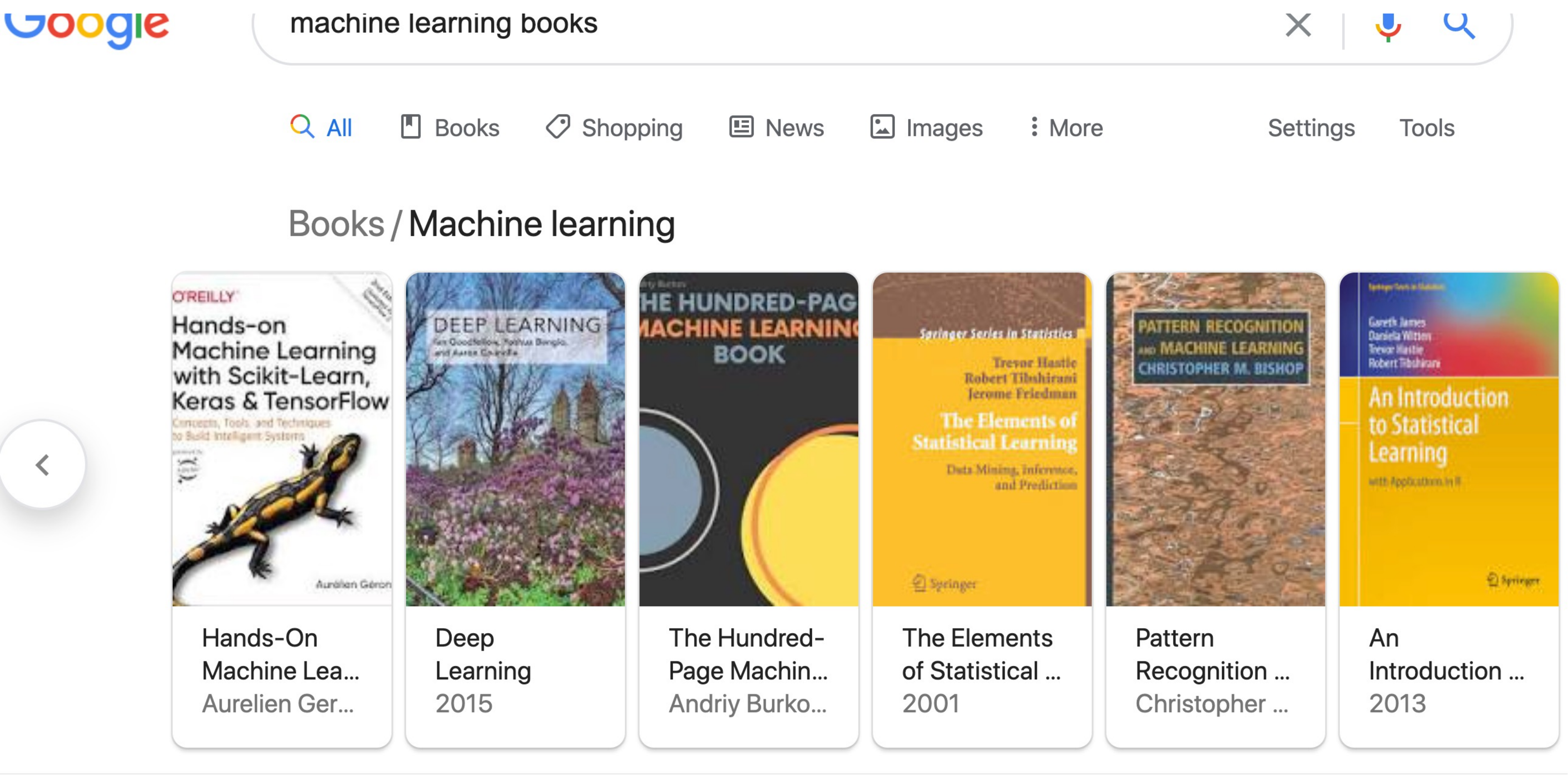
Course goals

- **Be familiar with math behind all major ML methods**
  - Information theory/entropy/KL divergence
  - Norms and distances
  - Likelihood: MLE/MAP
  - Optimization via gradient descent
  - EM
  - RL

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
- **Ed**
  - *look here first for answers!*

Textbooks



machine learning books

Google
All
Books
Shopping
News
Images
More
Settings
Tools
O'REILLY
Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow
Concepts, Tools, and Techniques to Build Intelligent Systems
Aurélien Géron
DEEP LEARNING
THE HUNDRED-PAGE MACHINE LEARNING BOOK
Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Springer
PATTERN RECOGNITION AND MACHINE LEARNING
CHRISTOPHER M. BISHOP
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Springer

Books / Machine learning

Hands-On Machine Lea...	Deep Learning	The Hundred-Page Machin...	The Elements of Statistical ...	Pattern Recognition ...	An Introduction ...
Aurelien Ger...	2015	Andriy Burko...	2001	Christopher ...	2013

Learning in the time of post?-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

Yes or no?

Start the presentation to activate live content

If you see this message in presentation mode, install the add-in or get help at PollEv.com/app

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

Working Together

Homework is mostly “pair programming” and “pair problem solving”

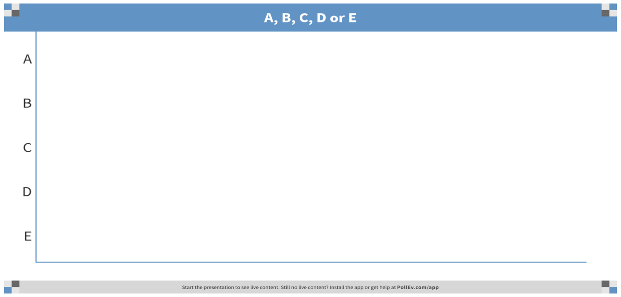
If it is determined that code submitted by two students might have been copied

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



A, B, C, D or E

Start the presentation to see live content. Still no live content? Install the app or get help at **PollEv.com/app**

Asking Questions

- **Questions about homework should be**
  - A) Asked during office hours
  - B) Emailed to the instructor or a TA
  - C) Asked on Ed
  - D) A or C
  - E) A, B or C

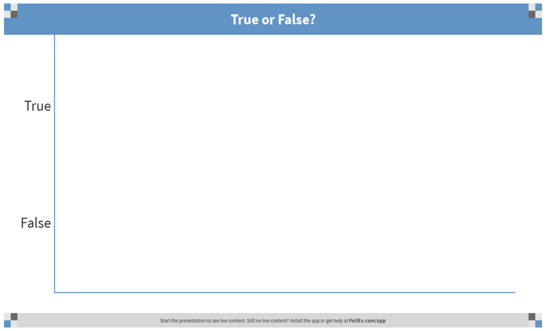
Start the presentation to see live content. Still no live content? Install the app or get help at **PollEv.com/app**

Python

- **Python is a better ML language than matlab**

  A) True

  B) False



Start the presentation to see live content. Still no live content? Install the app or get help at **PollEv.com/app**

Where is Machine Learning used?

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

Google
f
amazon.com
Baidu 百度 du
腾讯 Tencent
Alibaba Group 阿里巴巴集团

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/

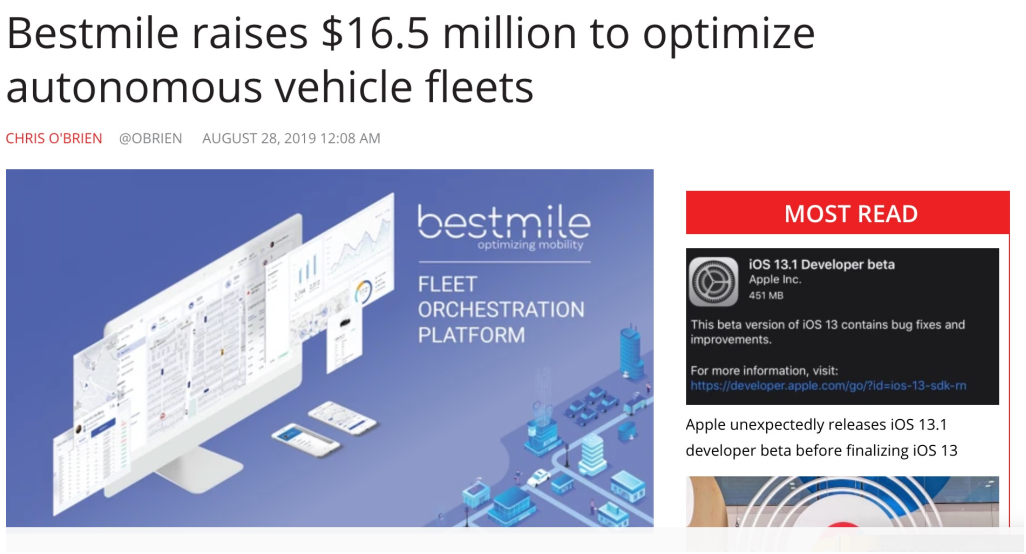
ML: manufacturing

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/

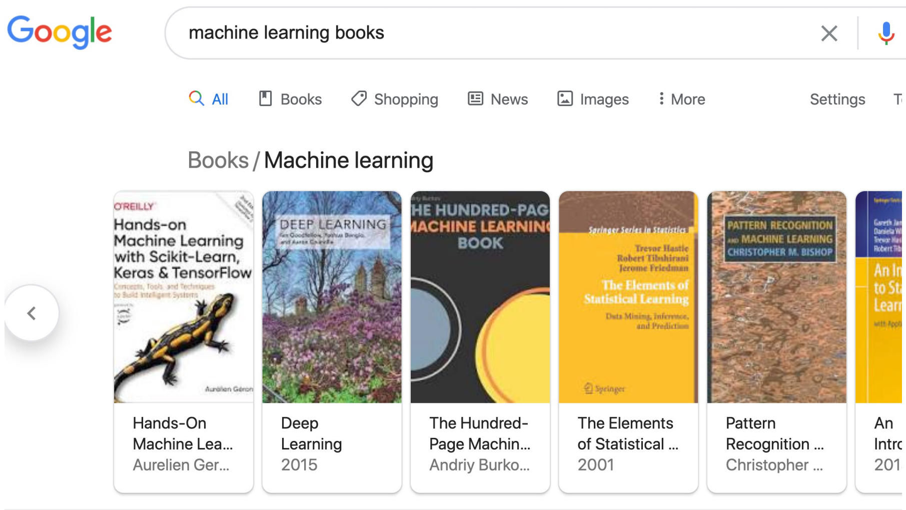
Components of ML

- ***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***
  - $L(\mathbf{y}, \hat{\mathbf{y}}) = ||\mathbf{y} - \hat{\mathbf{y}}||_2$
- ***Optimization method***
  - $\mathit{argmin}_w\, L(y, \hat{y}(w))$
  - gradient descent

Google ads as machine learning



machine learning books

Books / Machine learning

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

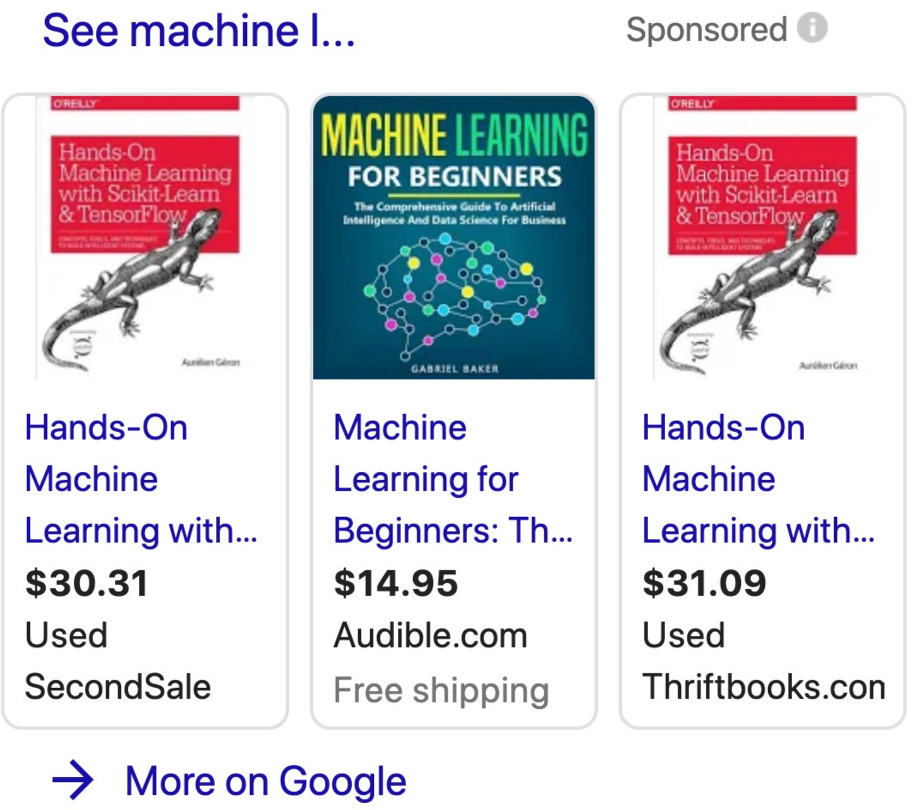
Deep Learning 2015

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

The Elements of Statistical ... 2001

Pattern Recognition ... Christopher ...

An Intro 201



See machine l...

Hands-On Machine Learning with...
$30.31
Used
SecondSale

Machine Learning for Beginners: Th...
$14.95
Audible.com
Free shipping

Hands-On Machine Learning with...
$31.09
Used
Thriftbooks.con

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 || ŷ(x) – y ||* - optimization
- **Unsupervised** **X**
  - *p(x)* - "generative" model

Types of models

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

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

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

TODO

- **Visit canvas** https://canvas.upenn.edu/
  - Take the self-test in canvas
  - Do HW 0 (trivial latex; be able to run numpy in jupyter)
- **Join Ed**
  - Linked to from canvas
- **Look at the wiki** https://alliance.seas.upenn.edu/~cis520/wiki
- **Get up to speed on python, numpy**
  - By doing the worksheets

What you should know

- **Turning a real-world problem into a well-posed ML problem is often hard**
  - pick features/predictors, $\mathbf{x}$
  - output/response, y
  - loss function L(y, $f(\mathbf{x}; \theta)$)
- **Unsupervised vs. supervised vs. reinforcement**
  - generative $p(\mathbf{x})$ vs. conditional $p(y|\mathbf{x})$ models
- **Parametric, non-parametric, semi-parametric**
- **Canvas, Ed, wiki**

What questions do you have on today's class?

Top

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**