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

Books / Machine learning

Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow
Aurelien Geron
2015

Deep Learning
Ian Goodfellow, Yoshua Bengio, and Aaron Courville

The Hundred-Page Machine Learning Book
Andriy Burkov

The Elements of Statistical Learning
Trevor Hastie, Robert Tibshirani, and Jerome Friedman
2001

Pattern Recognition and Machine Learning
Christopher Bishop

An Introduction to Statistical Learning
Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani
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

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
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
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: Autonomous vehicles

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

Bestmile raises $16.5 million to optimize autonomous vehicle fleets

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**
  - \( \hat{y} = f(x; w) = w^T x \)

- **Loss function**
  - \( L(y, \hat{y}) = \|y - \hat{y}\|_2 \)

- **Optimization method**
  - \( \text{argmin}_w L(y, \hat{y}(w)) \)
  - gradient descent
Google ads as machine learning

What features?
What model?
What loss function?
Types of Learning

- **Supervised**  
  - Given an observation $x$, what is the best label $y$?

- **Unsupervised**  
  - 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
  - $p(x)$
- Discriminative
  - $p(y|x)$

$X$: features, predictors, design matrix, input
$y$: response, label, output
Types of models

- **Parametric**
  - $\hat{y} = w \cdot x$
  - $\hat{y} = f(x; \theta)$
  - $w$ and $\theta$ 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/](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](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, \( x \)
  - output/response, \( y \)
  - loss function \( L(y, f(x; \theta)) \)

◆ Unsupervised vs. supervised vs. reinforcement
  - generative \( p(x) \) vs. conditional \( p(y|x) \) models

◆ Parametric, non-parametric, semi-parametric

◆ Canvas, Ed, wiki
What questions do you have on today's class?