CIS 520 Machine Learning

Shivani Agarwal & Lyle Ungar
Computer and information Science

Introductions

- **Who am I?**
- **Who are you?**
  - Why are you here?
- **What will this course look like?**
  - **Lectures & "Recitations"**
    - Slides, chalkboard, wiki & "clickers"
  - **Homework**
    - Math and MATLAB
    - Canvas and turnin
  - **Exams**
    - Midterm and final

Course goals

- **Be familiar with all major ML methods**
  - Regression (linear, logistic) & feature selection
  - Decision trees & random forests
  - Naive Bayes, Bayes Nets, Markov Nets, HMMs
  - SVM, kernels, PCA, CCA
  - online learning: boosting …
  - deep learning
- **Know their strengths and weaknesses**
  - know jargon, concepts, theory
  - be able to modify and code algorithms
  - be able to read current literature

Introductions (2)

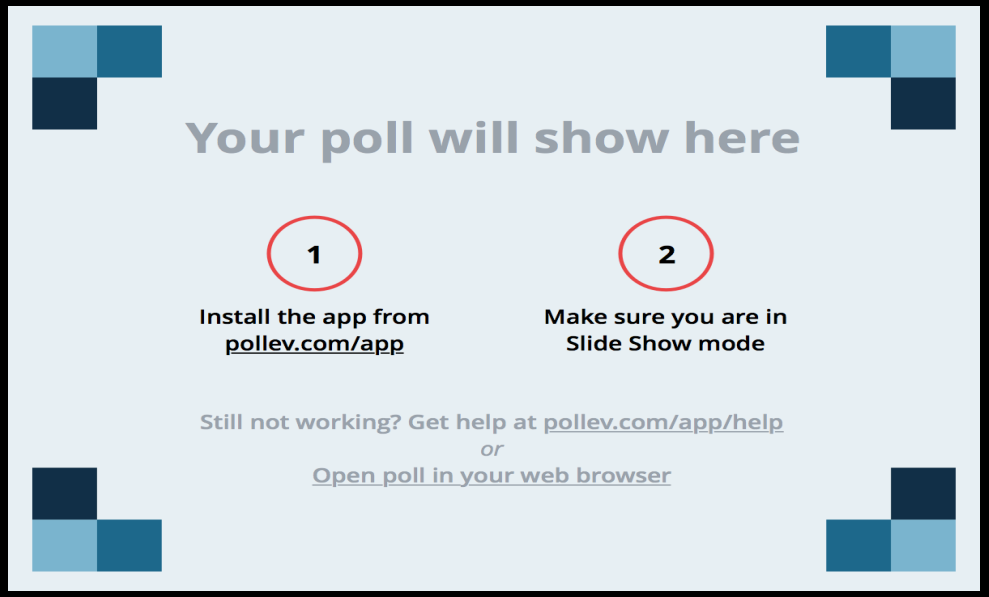
- **If you're waiting to get into this course**
  - It won't happen ☹
  - But the course will be offered again in the spring
- **Alternate courses**
  - CIS 419/519 Intro to Machine Learning
  - STAT 471/571/701 Modern Data Mining
  - CIS 545: Big Data Analytics

Administrivia

- **Course wiki**
  - Lecture notes
  - Resources
    - Grading scheme, academic integrity,
    - office hours, …
  - Reading (including the Bishop 'textbook' – free online)
    - Mostly for reading after lectures
    - But will sometimes add background info
- **Canvas**
  - Homework, grades
  - Lecture recordings
    - But don't count on them being useful
- **Piazza**
  - *look here first for answers!*

Do you have Polleverywhere?

A) Yes
B) No



Working Together

Homework is mostly "pair programming" or "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) Two of the above
  - E) None of the above

Matlab

- **We will use MATLAB**
  - **Free**
- **Matlab is a better language than python**
  - A) True
  - B) False
- **Matlab and Octave are**
  - A) Very different languages
  - B) Almost identical
  - C) Fully interchangeable except for the user interface
  - D) None of the above

Where is Machine Learning used?

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

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

What kinds of learning are missing here?

Types of Learning

- **supervised** **X, y**
  - P(y|x) - conditional probability estimation
  - min | $y^{est}(x) - y$ | - optimization
- **unsupervised** **X**
  - P(x) - "generative" model

Are you familiar with regression as a conditional probability?
A) Yes B) No

Are you familiar with regression as a minimization problem?
A) Yes B) No

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
- How might you go about solving it?

If you have questions, raise your hand and I'll come around.

Assessing code quality

- Given a bunch of student homework solutions and the ratings that graders gave them for 'coding style' , estimate the ratings for future code.
- What type of learning is this?
  - A) supervised
  - B) unsupervised
  - C) something else
- How might you go about solving it?

ML vs. Statistics

TODO

- **Join piazza**
  - Linked to from the course wiki
  - https://alliance.seas.upenn.edu/~cis520/wiki
- **Install Polleverywhere (free)**
- **Install matlab (free from Penn)**
- **Go to canvas**
  - Do HW 0 (trivial latex)

What you should know

- **Turning a real-world problem into a well-posed ML problem is often hard**
  - E.g. generate features/predictors, pick X and y
- **Unsupervised vs. supervised**
  - Generative P(x) vs. conditional P(y|x) models
- **Canvas, piazza, course wiki**