CIS 520 Machine Learning

Lyle Ungar

Computer and information Science
Introductions

◆ Who am I?
◆ Who are you?
  ● And why are you here?
  ● Prerequisites: linear algebra and time
◆ What will this course look like?
  ● Lectures & “Recitations”
    ■ Slides, chalkboard, wiki & “clickers”
  ● Homework
    ■ Math and MATLAB
    ■ Canvas and turnin
  ● Exams
    ■ Midterm and final  Note: the final is late this year 😞
Course goals

◆ Be familiar with all major ML methods
  ● Regression (linear, logistic) & feature selection
  ● Bayes Nets/Naive Bayes/HMM
  ● SVM, PCA, CCA, LDA
  ● online learning
  ● deep learning

◆ Know their strengths and weaknesses
  ● know jargon, concepts
  ● be able to modify and code algorithms
  ● be able to read current literature
Introductions (2)

◆ The TAs
  ● For office hours, see the wiki!

◆ Alternate machine learning: CIS 419/519
  ● Eric Eaton teaches another ML course this term
  ● And lots of statistics courses

◆ If you’re waiting to get into this course
  ● It won’t happen 😞
Adminstrivia

łam Course wiki https://alliance.seas.upenn.edu/~cis520/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
  ● Lecture recordings
    ■ But don’t count on them being useful

◆ Canvas https://canvas.upenn.edu/
  ● Homework, grades

◆ Piazza https://piazza.com/
  ● Questions - look here first!
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
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
  - \( P(y|x) \) - probability estimation
  - \( \min | y^{est}(x) - y | \) - optimization

- unsupervised
  - \( P(x) \)

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?
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)
What you should know

◆ Example applications of ML
◆ How hard it is to cast a real-world problem in terms of a well-posed ML problem
  ● E.g. generate features/predictors, pick X and y
◆ Generative vs. conditional models
  ● P(x) vs P(y|x)
  ● Unsupervised vs. supervised
◆ Free MATLAB to registered students
◆ Sign up on Piazza!!!