CIS 520 Machine Learning

Lyle Ungar

Install Poll Everywhere from google play or app store or https://pollev.com/lyleungar251

What's your favorite word?

happy
CIS 520 Machine Learning

Lyle Ungar
Computer and information Science
Should I be here?

◆ If you’re waiting to get into this course
  ● Only via https://forms.cis.upenn.edu/waitlist/index.php
  ● It probably won’t happen 😞
  ● But the course will be offered again in the spring

◆ Alternate courses
  ● CIS 419/519 Applied Machine Learning in python
  ● STAT 471/571/701 Modern Data Mining in R
  ● CIS 545: Big Data Analytics in python
Introductions

- Who am I?
- Who are you?
  - Why are you here?
- What will this course look like?
  - Lectures (no real Recitations)
    - Slides, chalkboard, wiki & “clickers”
  - Homework
    - Math (latex) and MATLAB
    - Canvas and turnin
  - Exams
    - Midterm and final
- Questions – before and after class
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
Administrivia

◆ **Course wiki**
  - Lecture notes
  - Resources
    - Grading scheme, academic integrity,
    - office hours, …
  - Readings -- including the Bishop ‘textbook’ – free online
    - Mostly for reading after lectures
    - ”supplemental” really means that

◆ **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” 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) Two of the above
E) None of the above
Matlab

- 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
- \( \min | y^\text{est}(x) - y | \) - 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 canvas and the course wiki
  - [https://alliance.seas.upenn.edu/~cis520/wiki](https://alliance.seas.upenn.edu/~cis520/wiki)

- **Install Poll Everywhere (free)**

- **Install MATLAB (free from Penn)**

- **Visit canvas**
  - [https://canvas.upenn.edu/](https://canvas.upenn.edu/)
  - Do HW 0 (trivial latex)

- **Review probability (see the wiki)**
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
This is your machine learning system?

Yup! You pour the data into this big pile of linear algebra, then collect the answers on the other side.

What if the answers are wrong?

Just stir the pile until they start looking right.