Elements of Machine Learning

https://www.cs.duke.edu/courses/fall22/compsci371/

Introduction and Logistics

Machine Learning Applications

- **Data Security**: Is this file malware?
- **Fraud Detection**: Is this transaction money laundering?
- **Personal Security**: What's in your bag? Is that you?
- **Photo Collections**: Here are all photos of Jenny playing tennis
- **Financial Trading**: Is this trade likely to profit me?
- **Healthcare**: Does this scan have a tumor? Do these symptoms suggest diabetes?
- **Marketing Personalization**: What can I sell *you*? What movies do you like?
- **Online Search**: Why did/didn't you like this search result?
- **Speech Processing**: What did you say? Let me transfer your call
- **Natural Language Processing**: Here is the information you need
- **Chatbots**: I can help you with your order. Tell me more about your symptoms
- **Smart Cars**: Are you comfortable? Are you alert? Stay in lane! Let me drive…
- …

Machine Learning in One Slide

- Identify a function $y = f(x)$:

$$x = \textbf{email}, \quad y = \textbf{SPAM/NO SPAM}$$

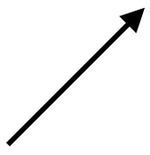
- Give lots of examples (a training set):

$$T = \{(x_1, y_1), \ldots, (x_N, y_N)\}$$

- A learner is **another function** λ:
  It takes T as input and outputs an approximation to f:

$$h = \lambda(T)$$

- Hopefully, f and h behave about the same
  even for previously unseen data:



$$h(x) \approx f(x)$$

- **That's** the big problem!

- **ML is not (just) data fitting**

Logistics

Academic Integrity

- *Short version: Cheating will be prosecuted*
- Cheating: Using someone else's material in your work without giving credit [Lone exception: class materials need not be cited]
- Ditto for making materials available to others
- Giver/receiver are treated the same
- Format for using/making available is immaterial
- Only communication allowed during homework is with your group peers, if any, and with the teaching staff

Notes, Slides, and Videos

- ***Notes on the class Syllabus web page are required reading, and are your main source of information***
- *All appendices in the notes are optional reading*
- Prerecorded videos are for backup, and are *optional*
- In-class lectures are recorded on Panopto, accessible through Sakai
- Feel free to integrate with other sources. See *Resources* web page

Questions for Discussion

- When possible, some lecture time is reserved for questions on the current topic
- You may submit questions for discussion any time on Ed Discussion
- I will address a sample of the questions submitted by noon of the day before lecture
- If your question is not addressed in class, please come to office hours

Homework

- One per topic
- Some math, some text, some programming
- OK to work in groups of one, two, three
  [but no division of labor!]
- Jupyter notebooks → HTML → PDF
- Keep Jupyter cells small
- Two submissions on Gradescope: PDF, Notebook
- Log in to Gradescope through Sakai!
- **One pair of submissions (two files, one PDF, one Notebook) per group. Remember to list all names in the files and through the Gradescope interface!**
- No late homework accepted
- Two worst homework scores (including 0s for no homework) are dropped

Your Weekly Schedule

- Tuesday: Attend lecture
- Thursday:
  - **Homework** about *previous* topic due by 8:30am EDT
  - Attend lecture
- Any Day: Submit **questions** on Ed Discussion. A sample of the questions submitted by noon the day before class will be answered in class when feasible

Exams and Grades

- Exams:
  - Midterm on October 27, in person, during class period
  - Final on December 19 at 7pm, in person, not cumulative
- Grades:
  - Homework 50%
  - Midterm 25%
  - Final 25%

Programming

- All programming will be in **Python 3** (not 2!)
- If you know how to program, picking up Python takes a few hours and Google while you program
- If you don't know how to program, this class may not be for you
- You will write **Jupyter Notebooks** for homework. They are easy to get used to, and let you intersperse text, math, figures, and code
- *Program in an IDE (**PyCharm recommended**) for easier debugging, then copy/paste into notebook*
- A first homework assignment helps you ease into these tools
- The **Anaconda** distribution for everything you need is very strongly recommended
- See the *Resources* web page for tutorials on Python 3, Jupyter, Anaconda

Teaching Staff

- *Graduate TAs*: Rajiv Nagipogu, Zonghao Huang
- *Undergraduate TAs*: Aakash Kothapally, Colin Zhu, Dennis Tang, Ishaan Maitra, Jingtong Zhu, Luke Triplett, Min Woo Kang, Morgan Chen, Nguyen Nguyen, Seojin Park, Yingjie Xu, Zimeng Fang

- If you like this course, please volunteer to TA next year!
- Each of us will have two office hours per week, times and venues TBA
- **Check the online calendar before attending office hours**
- We'll keep listening to Ed Discussion (at reasonable hours)
- **Talk to us!** We are here to help you learn