10-601 Introduction to Machine Learning

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Decision Trees (Part I)

Matt Gormley

Lecture 2
Jan. 15, 2020

Q&A

Q: In Lecture 1, why did we use the term **experience** instead of just **data**?

A: Because our concern isn't just the data itself, but also where the data comes from (e.g. an agent interacting with the world vs. knowledge from a book).

As well, the word *experience* better aligns with the notion of what humans require in order to learn.

Q&A

Q: Who is the single person that will most ensure that this course runs smoothly this semester?

A: Brynn Edmunds.

Q&A

Q: Are we using Canvas?

A: No.

Q&A

Q: How will I earn the 5% Participation points?

A: Good question! One way is by filling out the **required** poll on what WIFI enabled devices you have. (You must sign in to Google using your **Andrew ID** to fill out the form.)

https://forms.gle/UeiXzGrGgGujZKKM6

Other points will be earned through in-class polls, some “free poll points”, out-of-class surveys, and other opportunities to gain participation points.

Starting next week, please come to class with a WIFI enabled smartphone or tablet. We’ll announce on Piazza what to do if you don’t have such a device.

Q&A

Q: Can we have the handwritten notes from lectures?

A: Okay fine…

https://1drv.ms/o/s!Aqk9RupCw3gqhnEVySsGVwiAwMI6

… but just be warned that lots of education research suggests that taking your own notes is the best way to learn!

Reminders

- **Homework 1: Background**
  - **Out: Wed, Jan 15 (2nd lecture)**
  - **Due: Wed, Jan 22 at 11:59pm**
  - Two parts:
    1. written part to Gradescope
    2. programming part to Gradescope
  - unique policy for this assignment:
    1. **two submissions** for written (see writeup for details)
    2. **unlimited submissions** for programming (i.e. keep submitting until you get 100%)
  - unique policy for this assignment: we will grant (essentially) any and all extension requests

Machine Learning & Ethics

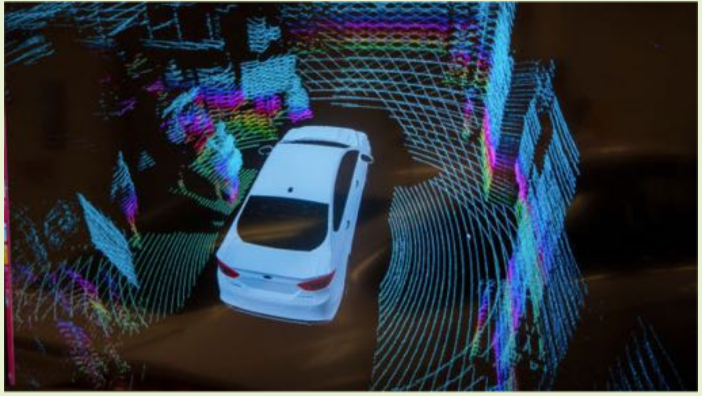
What ethical responsibilities do we have as machine learning experts?

Some topics that we won't cover are probably deserve an entire course

If our search results for news are optimized for ad revenue, might they reflect gender / racial / socio-economic biases?

http://bing.com/

http://arstechnica.com/



How do autonomous vehicles make decisions when all of the outcomes are likely to be negative?

Should restrictions be placed on intelligent agents that are capable of interacting with the world?

http://vizdoom.cs.put.edu.pl/

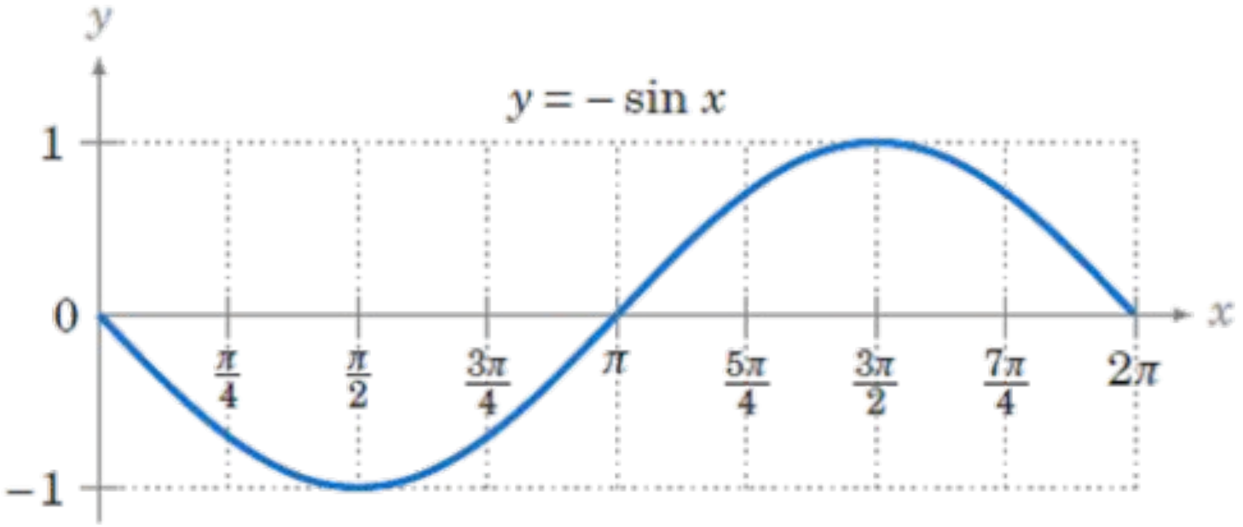
Big Ideas

1. How to formalize a learning problem
2. How to learn an expert system (i.e. Decision Tree)
3. Importance of inductive bias for generalization
4. Overfitting

FUNCTION APPROXIMATION

Function Approximation

Quiz: Implement a simple function which returns sin(x).



A few constraints are imposed:

1. You can’t call any other trigonometric functions
2. You *can* call an existing implementation of sin(x) a few times (e.g. 100) to test your solution
3. You only need to evaluate it for x in [0, 2*pi]

Medical Diagnosis

- Setting:
  - Doctor must decide whether or not patient is sick
  - Looks at attributes of a patient to make a medical diagnosis
  - (Prescribes treatment if diagnosis is positive)
- Key problem area for Machine Learning
- Potential to reshape health care

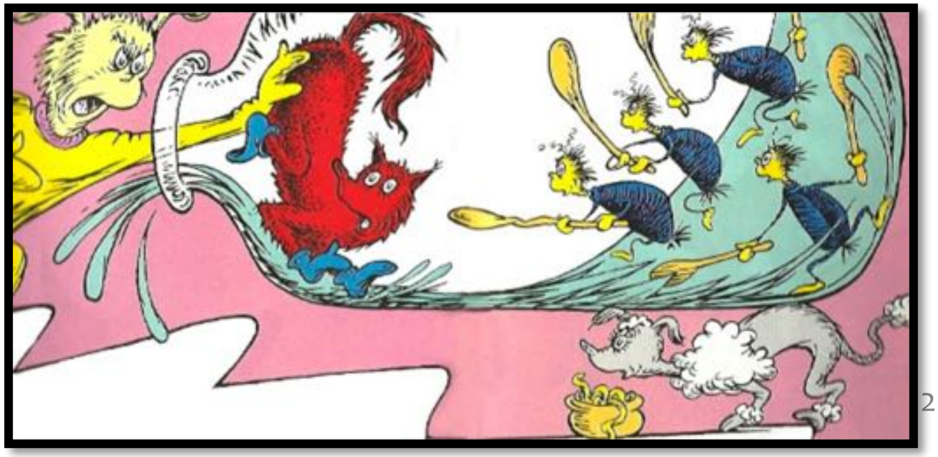
Medical Diagnosis

Interview Transcript
Date: Jan. 15, 2020.
Parties: Matt Gormley and Doctor E.
Topic: Medical decision making

Medical Diagnosis

Interview Transcript
Date: Jan. 15, 2020.
Parties: Matt Gormley and Doctor E.
Topic: Medical decision making

- Matt: Welcome. Thanks for interviewing with me today.
- Dr. E: Interviewing…?
- Matt: Yes. For the record, what type of doctor are you?
- Dr. E: Who said I'm a doctor?
- Matt: I thought when we set up this interview you said—
- Dr. E: I'm a preschooler.
- Matt: Good enough. Today, I'd like to learn how you would determine whether or not your little brother is sick given his symptoms.
- Dr. E: He's not sick.
- Matt: We haven't started yet. Now, suppose he is sneezing. Is he sick?
- Dr. E: No, that's just the sniffles.
- Matt: What if he is coughing; Is he sick?
- Dr. E: No, he just has a cough.
- [Editor's note: preschoolers unilaterally agree that having the sniffles or a cough is not the same as being sick.]
- Matt: What if he's both sneezing and coughing?
- Dr. E: Then he's sick.
- Matt: Got it. What if your little brother is sneezing and coughing, plus he's a doctor.
- Dr. E: Then he's not sick.
- Matt: How do you know?
- Dr. E: Doctors don't get sick.
- Matt: What if he is not sneezing, but is coughing, and he is a fox….
- Matt: …and the fox is in the bottle where the tweetle beetles battle with their paddles in a puddle on a noodle-eating poodle.
- Dr. E: Then he is must be a tweetle beetle noodle poodle bottled paddled muddled duddled fuddled wuddled fox in socks, sir. That means he's definitely sick.
- Matt: Got it. Can I use this conversation in my lecture?
- Dr. E: Yes



ML as Function Approximation

Chalkboard

- Example: Medical Diagnosis
- ML as Function Approximation
  - Problem setting
  - Input space
  - Output space
  - Unknown target function
  - Hypothesis space
  - Training examples
- Error Rate

ML as Function Approximation

Chalkboard

- Algorithm 0: Memorizer
- Aside: Does memorization = learning?
- Algorithm 1: Majority Vote

Majority Vote Classifier Example

Dataset:

Output Y, Attributes A and B

Y	A	B
-	1	0
-	1	0
+	1	0
+	1	0
+	1	1
+	1	1
+	1	1
+	1	1

In-Class Exercise

What is the **training error** (i.e. *error rate on the training data*) of the **majority vote classifier** on this dataset?

Choose one of:
{0/8, 1/8, 2/8, …, 8/8}

ML as Function Approximation

Chalkboard

- Algorithm 2: Decision Stump
- Algorithm 3 (preview): Decision Tree