CSE 446: Week 1
Decision Trees

Administrative Details

- Reminder: sign up for Piazza
- Only a few sections this quarter
  - Please come to office hours! They're there to help
    - Naozumi: Wed 13:30 CSE 220
    - Akshay: Fri 11:00 CSE 218
    - Isaac: Tue 12:00 CSE 220
    - Sergey: Mon 10:30 CSE 528
  - Start on homeworks early and get help as needed
- Should we have a Python tutorial?

Section Schedule (+ Python)

Week 1: Probability review

Week 2: Python tutorial

Week 3: Linear algebra review

Week 6: Midterm review

Week 10: Final exam review

Section Schedule

Week 2: Probability review

Week 3: Linear algebra review

Week 6: Midterm review

Week 10: Final exam review

Recap

- Parts of a machine learning algorithm
  - Data (input x and output y)
  - Hypothesis space (e.g. all boolean functions)
  - Objective (what makes one “incorrect” answer better/worse than another “incorrect” answer)
  - Algorithm (how do we get the least “incorrect” answer)

Recap

- Consider a simple, Boolean dataset:
  - $f : X \rightarrow Y$
  - $X = \{0,1\}^4$
  - $Y = \{0,1\}$

- Question 1: How should we pick the *hypothesis space*, the set of possible functions f?
- Question 2: How do we find the best f in the hypothesis space?

Dataset:

Example	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

Recap: All Boolean Functions?

Consider all possible boolean functions over four input features!

- 2^{16} possible hypotheses
- 2^{9} are consistent with our dataset
- How do we choose the best one?

x_1	x_2	x_3	x_4	y
0	0	0	0	?
0	0	0	1	?
0	0	1	0	0
0	0	1	1	1
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	?
1	0	0	0	?
1	0	0	1	1
1	0	1	0	?
1	0	1	1	?
1	1	0	0	0
1	1	0	1	?
1	1	1	0	?
1	1	1	1	?

Dataset:

Example	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

Decision Trees

- Popular hypothesis class
- Very widely used in practice, from images to stock prices to everything in between
- With a few modifications (model ensambles) decision trees are one of the most successful models around, despite being 30+ years old
- Together with neural networks (which we will cover later in the course) they win the most Kaggle competitions…

kaggle

Decision Trees

[tutorial on the board]

[see lecture notes for details]

I. How to make one decision: designing our first machine learning algorithm
II. How to make multiple decisions
III. Approximate fitting
IV. Greedy vs exact fitting
V. Define the hypothesis space