Lecture 15.1: Special topics

Roadmap

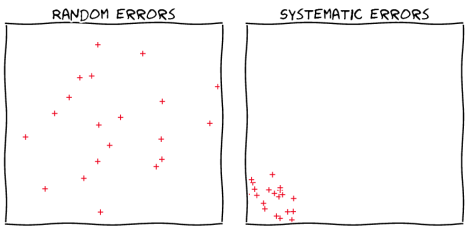
Algorithmic fairness

Causality

Figure from Moritz Hardt

Fairness

Two classifiers with 5% error:

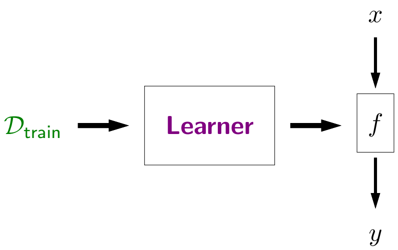


Fairness in criminal risk assessment

- Northpointe: COMPAS predicts criminal risk score (1-10)
- ProPublica: given that an individual did not reoffend, black individuals 2x likely to be (wrongly) classified 5 or above
- Northpointe: given a risk score of 7, 60% of white individuals reoffended, 60% of black individuals reoffended

[whiteboard: different fairness criteria]

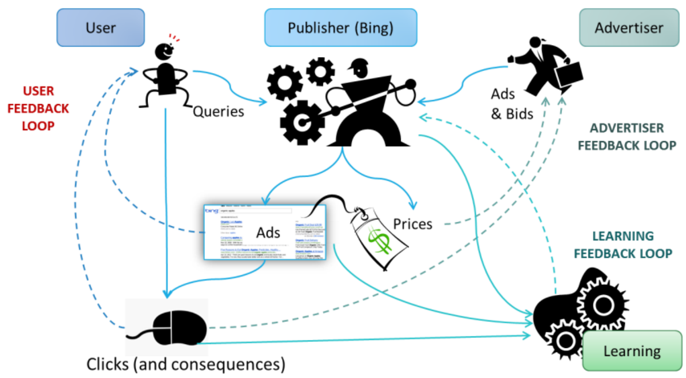
Are algorithms neutral?



By design: picks up patterns in training data, including biases

[Leon Bottou]

Feedback loops



Roadmap

Algorithmic fairness

Causality

Causality

Goal: figure out the effect of a treatment on survival

Data:

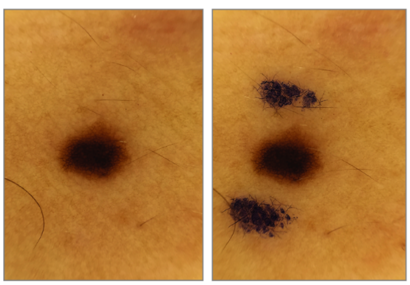
For untreated patients, 80% survive
For treated patients, 30% survive

Does the treatment help?

Who knows? Sick people are more likely to undergo treatment...

[Winkler et al., JAMA 2019]

Non-causal features

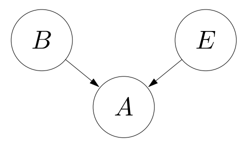


Same benign mole, with and without ink markings

Existing CNN model trained to predict if benign or malignant

- No ink: 16% false positive rate
- With ink: **54%** false positive rate

Bayesian network (alarm)



$$\mathbb{P}(B = b, E = e, A = a) = p(b)p(e)p(a \mid b, e)$$

Probabilistic inference (alarm)

Joint distribution:

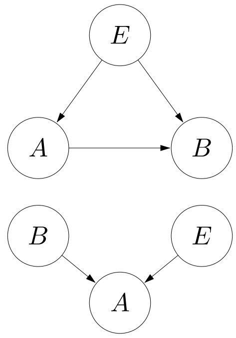
b	e	a	$\mathbb{P}(B = b, E = e, A = a)$
0	0	0	$(1-\epsilon)^2$
0	0	1	0
0	1	0	0
0	1	1	$(1-\epsilon)\epsilon$
1	0	0	0
1	0	1	$\epsilon(1-\epsilon)$
1	1	0	0
1	1	1	ϵ^2

Queries: $\mathbb{P}(B)$? $\mathbb{P}(B \mid A = 1)$? $\mathbb{P}(B \mid A = 1, E = 1)$?

Alternative probabilistic program

- Earthquake happens with probability ϵ
- If earthquake, alarm activates.
- If no earthquake, **alarm accidentally activates with probability** ϵ.
- If alarm and no earthquake, alarm company tries to hide their mistake by **hiring a someone to burglarize your home**.
- If alarm and earthquake, a burglar visits with probability ϵ.

Alternative Bayesian network



Can express the same joint probability distribution

Interventions

Intervention: what happens if I **set** a variable to a particular value?

Not the same as **observing** that a variable has a particular value!

[whiteboard: do calculus]

Summary

- Fairness: care about **which** errors, not just how many errors

- Causality: needed to understand effects of interventions

- Both important in real systems