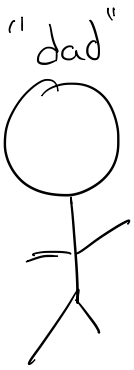
Causal Inference

Is ML useless for making decisions? Possibly



male
68 yrs
history of congenital heart failure
atrial fibrillation
takes aspirin

USe ML to predict stroke within 1 year

$f(x)$ = fctn of (gender, race, age, medical history, drugs, etc.)

$f(x) = 3.4 \cdot \mathbb{1}_{[\text{congenital heart failure}]} + 4.2 \cdot \mathbb{1}_{[\text{afib}]} + \cdots$

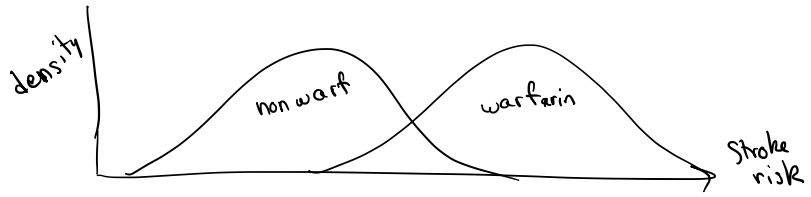
$\cdots + 1.1\,\mathbb{1}_{[\text{female}]} + 2.4 \cdot \mathbb{1}_{[\text{smoking}]}$

If dad stops smoking, does it reduce his risk of stroke?

$f(x) = 2^* \mathbb{1}_{[\text{warfarin}]} - 1 = \begin{cases} \text{predicts stroke if warfarin} \\ \text{predicts no stroke if not} \end{cases}$

Correlation $\not\Rightarrow$ causation

what happened?



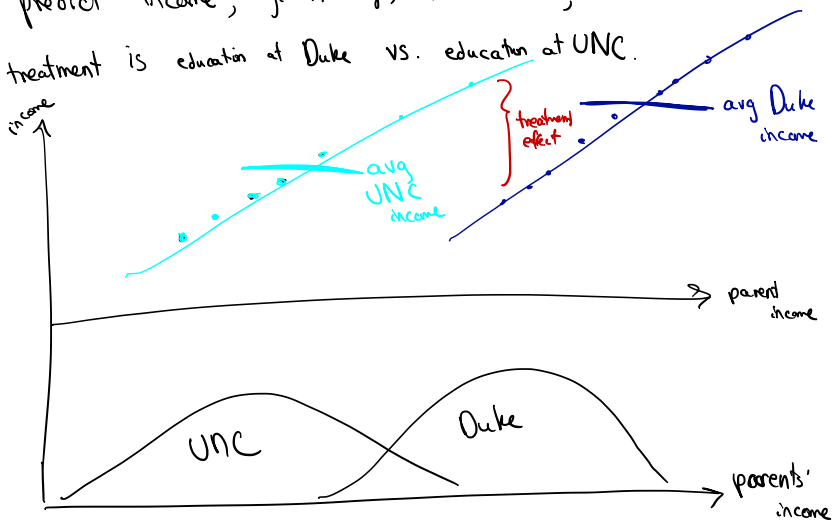
Regular ML: model $Y|X$

"Causal inference" ML: model $Y|X, T=1 - Y|X, T=0$

↑ treatment effect, conditional difference

Example: predict income, given age, parent's income, major field, ...

treatment is education at Duke vs. education at UNC.



What data do you get?

X	T	Y
parents income	UNC/DUKE	income
200K	Duke	80K
⋮	⋮	⋮

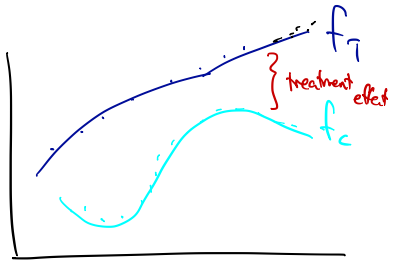
To do supervised learning, need X, treatment effect $Y|X, T=1 - Y|X, T=0$

↑ never have that

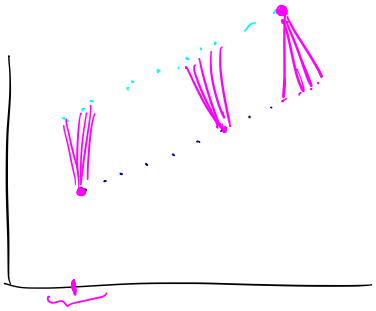
problem is exactly half supervised.

How to use ML to solve it.

- regress separately on treatment & control
  $f_T - f_c$



- matching or nearest neighbor
- ⋮ <tree structures>



beware of MLers who don't know CI!

Course Overview

perceptron alg, winnow alg, convergence proofs

KDD process

How to evaluate an alg/model:

2 kinds of ROC curves, AUC

cross validation

General form of a supervised learning alg

{ loss + C·Regul
{ (in)likelihood + C·Prior

Trees + Ensemble methods

- CART, C4.5
- random forest + variable importance
- boosting

Generative Models: Logistic regression, boosting, Gaussian Mixture Models

Optimization methods: boosting (coord desc)

SVM (convex optimiz + duality)

NN's (back prop)

High dimensional spaces: • kernels + RKHS for SVM + ridge

• margin theory for SVM

Generalization = data + knowledge

Analytical Methods: Least squares, ridge, kernel ridge regression

kernel regression

Clustering: K-means, HAC, GMM

Multi-armed Bandits

Causal Inference