CAUSE: A Data Repository for Causal Inference

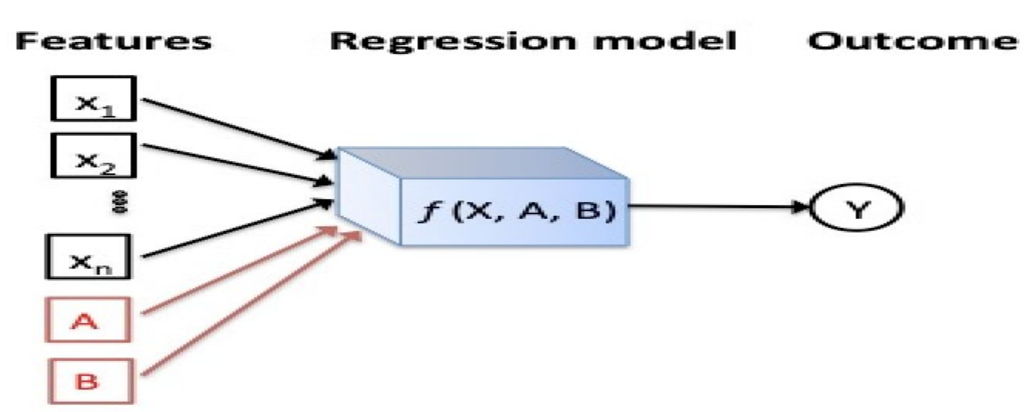
Raha Moraffah, Lu Cheng, Ruocheng Guo, K Selcuk Candan, Adrienne J Raglin, Huan Liu

Arizona State University

{rmoraffa, huan.liu}@asu.edu

Traditional Supervised Learning vs Causal Inference

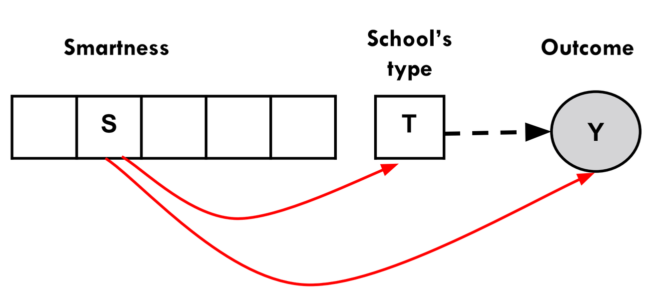
- How does type of school affect child's achievements in feature?
- Traditional supervised learning

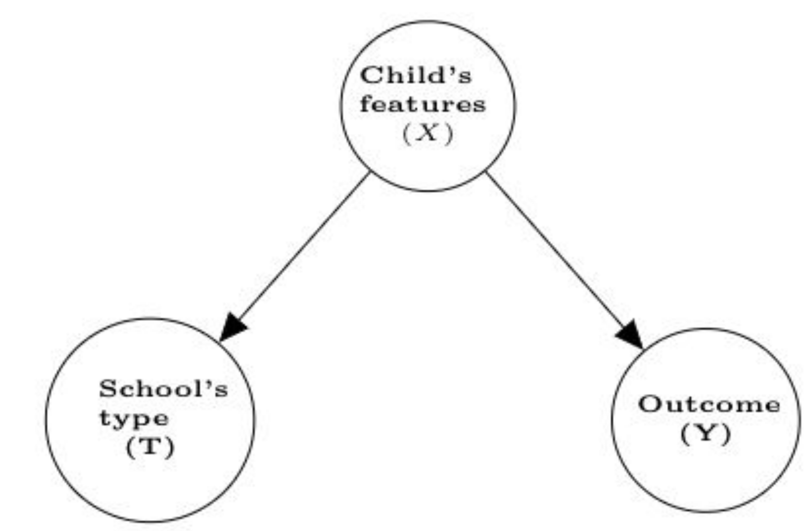


- These methods are designed to predict the outcome and not the effect of types of schools

Correlation vs Causation

- From observational data, School's type (T) and outcome (Y) are correlated, But neither of them might cause the other one
- Observed correlation could be due to the child's features





Challenges and desiderata of causal inference

Challenges:

- Causal structure of variables is unknown
- Not all variables are confounders
- Counterfactuals **can not** be observed in observational data

Example:

Factual: John Went to X school and his college gpa is A.

Counterfactual: What would have been John's gpa, had he not gone to X school?

Desiderata:

- Overlap assumption: Common support
- Ignorability assumption: No unmeasured confounder

Causal feature evaluation; datasets and methodologies

Methodologies:

- Available datasets for causal discovery:
  - Not available for all tasks
  - Existing datasets are small
- Mapping the problem to another domain and use existing datasets :

- Relaxing the problem by enforcing strong assumptions:
  - Based on the theory of transportability, causal relationships are more robust
  - compare the performance of classifiers with different interventional distributions

Data Mining and Machine Learning Lab

Arizona State University