A Practical Data Repository for Causal Learning with Big Data

Bench'19

Lu Cheng (Arizona State University)
Raha Moraffah (Arizona State University)
Ruocheng Guo (Arizona State University)
K.S. Candan (Arizona State University)
Adrienne Raglin (US Army Research Laboratory)
Huan Liu (Arizona State University)

Agenda

Introduction

Causal Effect Estimation

Causal Machine Learning

Causal Discovery

Introduction

A simple definition of causality:

A variable T causes Y iff changing T leads to a change in Y, while *keeping everything else constant*.

Introduction

Machine Learning is doing well. Why should we care about causality?

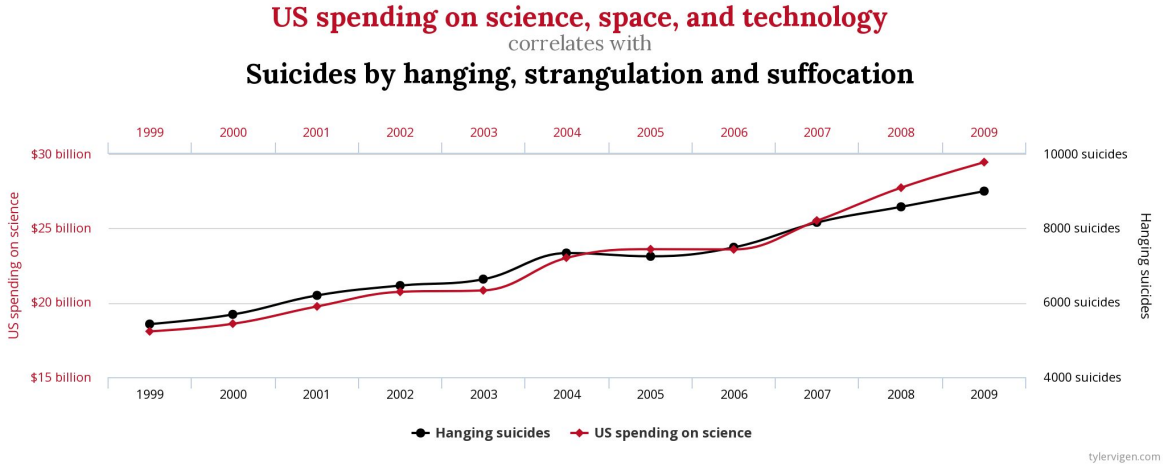
Will the predictions always be robust?

Does prediction always guide decision making?

Introduction

Will the predictions always be robust?

Correlation can be *spurious*



Credit: http://www.tylervigen.com/spurious-correlations

Introduction

Does prediction always guide decision making?

	Algorithm A	Algorithm B
CTR for Low-income users	10/400 (2.5%)	4/200 (2%)
CTR for High-income users	40/600 (6.6%)	50/800 (6.2%)
CTR for all users	**50/1000 (5%)**	**54/1000 (5.4%)**

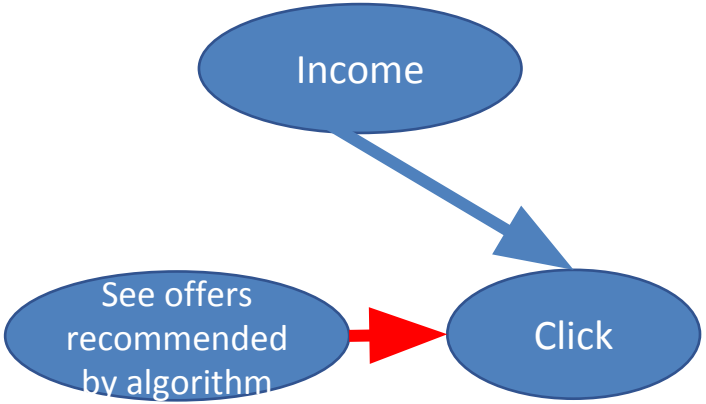
- Observation 1: CTR is higher for algorithm A in both low and high-income group.
- Observation 2: CTR is higher for algorithm B in the whole population.
- Which algorithm is better?

Introduction

Does prediction always guide decision making?

- Which algorithm is better? The underlying causal graph tells the answer.

Higher CTR for algorithm B due to **algorithm itself**



- The conditional probability Pr(click|algorithm) reflects the true causal effect (algorithm->click).
- Decision: Algorithm B is better.

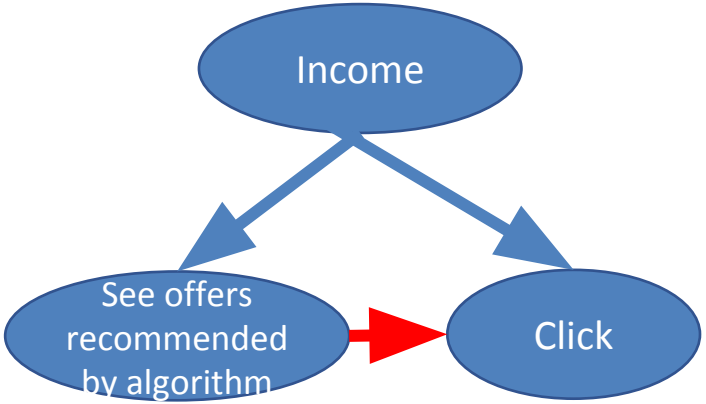
	Algorithm A	Algorithm B
CTR for all users	**50/1000 (5%)**	**54/1000 (5.4%)**

Introduction

Does prediction always guide decision making?

- Which algorithm is better? The underlying causal graph tells the answer.

Higher CTR for algorithm B due to **confounding bias**



- The conditional probability Pr(click|algorithm) does not reflect the true causal effect (algorithm->click).
- We need to block the influence from the confounder (income). We do this by subgrouping.
- Decision: Algorithm A is better.

	Algorithm A	Algorithm B
CTR (Low-income)	10/400 (2.5%)	4/200 (2%)
CTR (High-income)	40/600 (6.6%)	50/800 (6.2%)

Agenda

Introduction

Causal Effect Estimation

Causal Machine Learning

Causal Discovery

Causal Effect Estimation

Sometimes, with prior knowledge, we know there may exist a cause-effect pair (recommendation -> CTR), but we aim to estimate how significant the effect is.

Definition:

*The **causal effect** is the magnitude by which the outcome variable Y is changed resulting from a unit change in the cause (treatment) variable T.*

Motivating examples:

- Economists want to understand how effective is a job training program (T) on job seekers' employment rate/income (Y).

Causal Effect Estimation

Some definitions

Potential Outcome. Given an instance i and the treatment t, the potential outcome of i under treatment t, denoted by y_i^t, is the value that y would have taken if the treatment of instance i had been set to t.

Individual treatment effect:

$$\tau_i = \mathbb{E}[y_i^t] - \mathbb{E}[y_i^c], \tag{1}$$

Where c and t signify the *control* and a treatment.

We can also calculate the average treatment effect

$$\bar{\tau} = \mathbb{E}_i[\tau_i].$$

Causal Effect Estimation

Typical data: observational data {(xi,ti,yi)}

Challenges:

- Counterfactuals: Only one of the potential outcomes can be observed, so we need to estimate the other outcomes (i.e., **counterfactual outcomes**).
- Confounding bias: outcome is influenced by variables other than the treatment, we need to figure out which are these variables and control their influence without knowing the underlying causal relations
  - Some of these variables are a part of xi or highly correlated with xi.
  - However, some of them may not be measured.

Causal Effect Estimation

The widely used datasets:

Jobs. Job training -> Employment. The first part is from the randomized experiment by LaLonde (297 treated and 425 control). The second part is the a larger comparison group (2,490 control). The features describe each job seeker.

Infant Health Development Program. Home visits -> children's cognitive test scores. This is a dataset with true features but simulated treatments and outcomes. This dataset comprises 747 instances. Features describe the children and their mothers.

Causal Effect Estimation

The widely used datasets:

Twins. Born weight -> mortality in the first year of life. Researchers focused on the twins with weights less than 2kg to get a more balanced dataset in terms of the outcome. This results in a dataset consisting of 11,984 such twins. Each twin-pair is represented by features relating to the parents, the pregnancy status and birth status.

Causal Effect Estimation

Limitations of existing datasets:

- Size and dimension are often limited: from economics, education and healthcare experiments.
- A/B tests data from tech companies: hard to be open-source.
- Only deal with relatively simple treatment variables
    - For example, in search engine, the treatment (a ranked list of items) can take too many values, for which, dataset for treatment effect estimation can be extremely therefore ineffective for solving the problem.

Agenda

Introduction
Causal Effect Estimation
Causal Machine Learning
Causal Discovery

Causal Inference for Recommendation

- Problem: given a user and a set of products, we need to recommend a ranked list of items to her.
- Challenge: selection bias in the supervision signals. Users would only click or rate the items they like.

Datasets

- Test sets have to be randomized.
- The input data of learning a recommendation policy consists of products each user decided to look at and those each user liked/clicked. The treatment is the recommended products and the outcome is whether this user clicks this product.
- Standard datasets for recommender systems are not applicable in the evaluation of the deconfounded recommender systems due to the lack of outcomes for counterfactuals.

Randomized Control Trial Dataset

- *Yahoo-R3*. Music ratings collected from Yahoo! Music services. This dataset contains ratings for 1,000 songs collected from 15,400 users with two different sources. One of the sources consists of ratings for randomly selected songs collected using an online survey conducted by Yahoo! Research. The other source consists of ratings supplied by users during normal interaction with Yahoo! Music services.

Semi-synthetic Datasets

- Simulations are based on datasets for recommendation system, such as *MovieLens10M, Netflix, ArXiv*
- The key is to ensure the different data distributions between training/validation and testing
- One common approach is to create two training/validation/test splits from the standard datasets – **regular** and **randomized**
- To construct randomized test sets, we first sample a test set with roughly 20% of the total exposures (entries with ratings/clicks) such that each item has uniform probability. Training and validation sets are generated by randomly selecting remaining data with 70/10 proportions.

Simulated Datasets

- *Coat Shopping Dataset*. This is a synthetic dataset that simulates customers shopping for a coat in an online store. The training data was generated by giving Amazon Mechanical Turkers from a simple web-shop interface.

Limitations

- RCT for a recommender system is not practical in real-world applications.
    - For example, a recommender system that randomly recommends songs to its users can largely degrade user experience.
- The mismatch of synthetic data and the observed data from true environment is often unavoidable in practice
- Treatments and the corresponding potential outcomes are not well defined

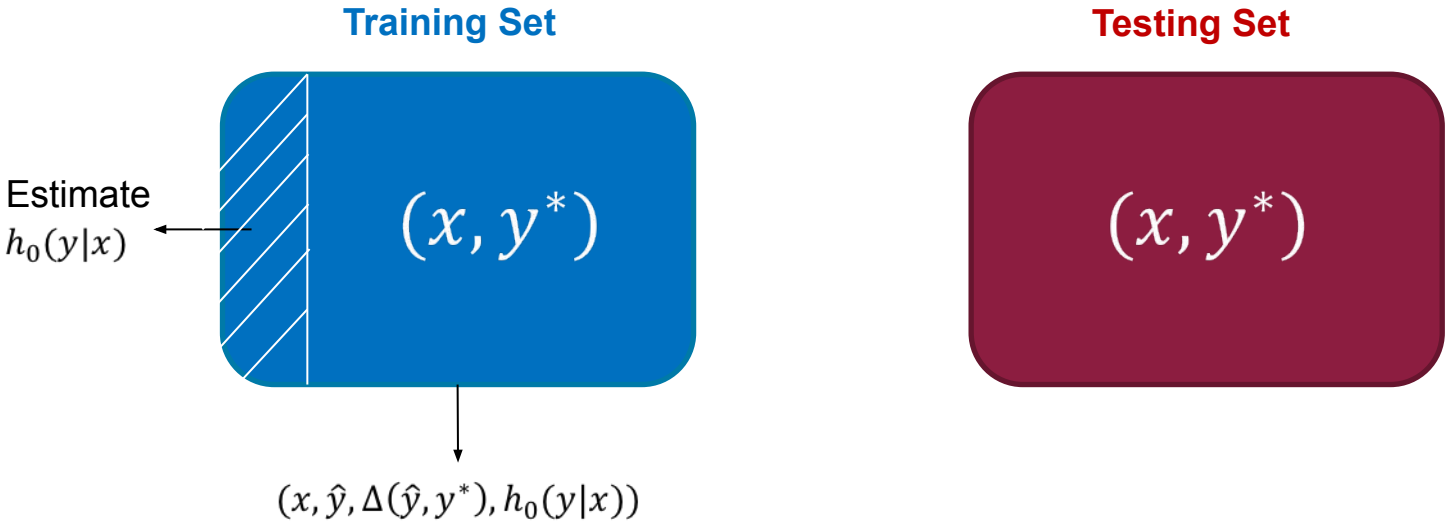
Off-policy Evaluation

- Problem Description
  - Given an **existing policy** h_0 that selects actions based on item features and observes rewards, e.g., search engines, **input feature** x, output prediction of **selected action** $y \sim h_0(y|x)$ and the **feedback** δ, the goal is to evaluate a **new policy** h.
  - Example: In a recommender system, x is the attribute of the items, y is the set of items recommended by the system, and δ denotes the user feed back, e.g., whether a user clicks on the item or not.
- Challenges
  - Evaluation of the new policy is challenging due to the selection bias and partial information

Datasets from the Real World

- Datasets from the Real World
  - *Music Streaming Sessions Dataset.* This dataset from Spotify consists of over 160 million listening sessions with user interaction information. It has metadata for approximately 3.7 million unique tracks referred to in the logs, making it the largest collection of such track data currently available to the public. A subset of this dataset is crawled and labelled by a uniformly random shuffle to satisfy RCT.
- Semi-simulated Datasets.
  - *Bandit Data Generation.* The key is to convert the training partition of a full-information multi-class classification dataset into a partial information bandit dataset for training off-policy learning methods while the test dataset remains intact to evaluate the new policy.

Bandit Data Generation



Limitations

- It might not be clear how it can be used in other applications of off-line evaluations
    - E.g., medical study. It is unethical to assign multiple treatments that might have detrimental interactions between drugs to patients
- The predefined hypothesis h_0 can largely affect the performance of the new policy
- The mismatch of synthetic data and the observed data from true environment is often unavoidable in practice, resulting in policies that do not generalize to the real environment

Agenda

Introduction
Causal Effect Estimation
Causal Machine Learning
Causal Discovery

Causal Discovery

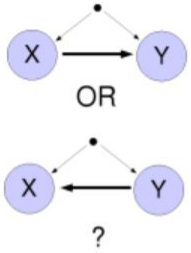
- Problem statement:
  - Causal discovery addresses the problem of learning the underlying causal mechanisms and the causal relations amongst variables in the data.

  - Research in this area can be divided into two major categories:
    1) **Learning causal direction (causal or anti causal relations)** between two variables. Specifically, given the observations $\{(x_i, y_i)\}_{i=1}^{n}$ of random variables, the goal is to infer the causal direction, i.e. whether $x \rightarrow y$ or $y \rightarrow x$.
    2) **Learning the underlying Causal Bayesian Network (CBN)** which shows the causal relationships among all the variables observed in the data.

Datasets: Causal Direction

- Tübingen Cause-Effect Pairs (TCEP): This dataset consists of real-world cause-effect samples collected across various subject areas.The ground-truths are true causal direction provided by human experts.
- Pittsburgh Bridges: A dataset with 108 bridges and 4 cause-effect pairs as following:

  1) Bridge Erection (Crafts, Emerging, Mature, Modern)→Span (Long, Medium, Short)
  2) Material (Steel, Iron, Wood)→Span (Long, Medium, Short)
  3) Material (Steel, Iron, Wood)→Lanes (1, 2, 4, 6)
  4) Purpose (Walk, Aqueduct, RR, Highway)→type (Wood, Suspen, Simple-T, Arch, Cantilev,
  CONT-T).



Datasets: Causal Bayesian Network Discovery

- Lung Cancer Simple Set (LUCAS)
  - A synthetic dataset which was made publicly available through the causality workbench
  - Ground-truth consists of 12 binary variables and their causal relations: 1) Smoking, 2) Yellow Fingers, 3) Anxiety, 4) Peer Pressure, 5) Genetics, 6) Attention Disorder, 7) Born on Even Day, 8) Car Accident, 9) Fatigue, 10) Allergy, 11) Coughing and 12) Lung Cancer

Evaluation methods

- False discovery rate (FDR)
  - (Reversed edges + False Positive)/ Positive
- True positive rate (TPR)
  - True Positive / True
- Structural Hamming distance (SHD)
  - Describes the number of changes that have to be made to a network for it to turn into the one that it is being compared with.

Advantages, Disadvantages and Limitations

- Advantages:
  - There exists a number of real-world datasets for the task of learning the causal direction between two variables that can be used in future research.
  - These datasets are collected for real world scenarios and are annotated by the experts in corresponding fields, which make these desirable and useful for research in this field.
- Disadvantages:
  - No large-scale data is available for the task of finding the underlying Causal Bayesian Network

Advantages, Disadvantages and Limitations

- Limitations:
  - Existing data for causal discovery is small and therefore it is not a good practice to use the machine learning models that require huge amount of data
  - Collecting data for causal discovery task is a tremendously difficult task due to the lack of human experts and resources

Summary

You can find corresponding data and algorithm repositories on github.

https://github.com/rguo12/awesome-causality-algorithms

https://github.com/rguo12/awesome-causality-data

We talked about three popular causal inference/learning problems, their evaluation metrics and the existing datasets for evaluating models for solving these problems.

Creating real-world large scale and high dimensional datasets for studying causal inference/learning in academia is still an open problem.

A comprehensive introduction to causal inference/learning:

Guo, Ruocheng, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. "A Survey of Learning Causality with Data: Problems and Methods." *arXiv preprint arXiv:1809.09337* (2018).