Discussion: Adaptive Text Embeddings for Causal Inference (Veitch, Sridhar & Blei, 2020)

Julian Ashwin

University of Oxford

NLP Reading Group, November 26, 2020

Summary

- **Motivation:** to test causal hypotheses, we often have to adjust for confounding features.
- **Problem:** these confounding features might appear as unstructured data such as text.
- **Approach:** Develop *causally sufficient embeddings* that provide a supervised low-dimensional representation of documents.
- **Answer:** Show improved causal estimation in synthetic datasets and two real-world examples.

Intuition

Does adding a theorem to a paper affect its chance of acceptance?

- Inclusion of theorem is straightforward to measure and observable
- But any causal link could be confounded by the subject of the paper:
  - Some subjects are more likely to be treated (i.e. have a theorem)
  - Outcome also varies by subject (i.e. some more likely to be accepted than others)
- We want to use the text to adjust for the subject and estimate the causal effect.
- If treatment is binary, all we need is the propensity score (prob of treatment given text) and the expected outcome given text.

Strategy

- An observation consists of outcome Y_i, treatment T_i and text W_i.
- $Z_i = f(W_i)$ is the part of the text that might confound the causal effect.
- The causal effect is then

$$\psi = \mathbb{E}[\mathbb{E}[Y|Z, T = 1]] - \mathbb{E}[Y|Z, T = 0]|T = 1]$$

- The conditional expectation for Y is $Q(t, z) = \mathbb{E}[Y|t, z]$
- The propensity score is $g(z) = P(T = 1|z)$.
- So we need estimators $\hat{g}(z_i)$ and $\hat{Q}(t, z)$, which we then use for, e.g.

$$\hat{\psi} = \frac{1}{n}\sum_i \left[\hat{Q}(1, z_i) - \hat{Q}(0, z_i)\right]\hat{g}(z_i) / \left(\frac{1}{n}\sum_i t_i\right)$$

Two Models

1. Causal BERT
2. Causal Amortized Topic Model

Causal Bert

Three outputs:

1. Document-level embeddings
    - Unsupervised embedding, I think just the standard BERT:

$$\lambda_i = f((\xi_{w_{i1}}, \xi_{w_{il}}), \gamma^U)$$

2. Map from embeddings to treatment probability
    - Logit linear layer $\lambda_i \rightarrow \tilde{g}(\lambda_i; \gamma^g)$
3. Map from embeddings to expected outcomes
    - 2-hidden layer neural net for each value of t:

$$\lambda_i \rightarrow \tilde{Q}(0, \lambda_i; \gamma^{Q_0})$$

$$\lambda_i \rightarrow \tilde{Q}(1, \lambda_i; \gamma^{Q_1})$$

Estimate these jointly, so objective includes prediction of outcome and treatment as well as unsupervised embedding.

Causal Amortized Topic Model

1. Topic model estimated using feedforward "encoder" neural network
   - Produces topic proportions for each document, θ_i
2. Logit linear mapping for propensity score: $\theta_i \rightarrow \tilde{g}(\theta_i; \gamma^g)$
3. Linea mapping from topics to outcome: $\theta_i \rightarrow \tilde{Q}(\theta_i; \gamma^Q)$

Also estimated jointy, so the loss function is includes prediction of outcome and treatment as well as unsupervised embedding.

Key Assumption

Key assumption for causal identification is that adjusting for z is sufficient to capture all relevant information from w

Additional assumption: there are no confounding variables that are external to the text.

- do referees recognise papers by more prestigious authors?
- do well-known Reddit users get more positive feedback?

Semi-synthetic data

- Can't observe true causal effect, so use semi-synthetic dataset:
  - Simulate an outcome that dependson both the treatment and a confounder.
  - Confounders used: title buzziness and subreddit $\tilde{z}$
  - Simulate outcome from observed treatment and the propensity score given the observed confounder, e.g.

$$Y_i = t_i + b_1(\pi(\tilde{z} - 0.5)) + \epsilon_i$$

- Both language modelling and the supervison elements are important in recovering the ground truth causal effect.

Results

(a) Language Modeling Helps

Dataset:	Reddit (NDE)	PeerRead (ATT)
Ground truth	1.00	0.06
Unadjusted	1.24	0.14
NN $\hat{\psi}^Q$	1.17	0.10
NN $\hat{\psi}^{\text{plugin}}$	1.17	0.10
BERT (sup. only) $\hat{\psi}^Q$	0.93	0.19
BERT (sup. only) $\hat{\psi}^{\text{plugin}}$	1.17	0.18
C-ATM $\hat{\psi}^Q$	1.16	0.10
C-ATM $\hat{\psi}^{\text{plugin}}$	1.13	0.10
C-BERT $\hat{\psi}^Q$	1.07	0.07
C-BERT $\hat{\psi}^{\text{plugin}}$	1.15	0.09

(b) Supervision Helps

Dataset:	Reddit (NDE)	PeerRead (ATT)
Ground truth	1.00	0.06
Unadjusted	1.24	0.14
BOW $\hat{\psi}^Q$	1.17	0.13
BOW $\hat{\psi}^{\text{plugin}}$	1.18	0.14
BERT $\hat{\psi}^Q$	-15.0	-0.25
BERT $\hat{\psi}^{\text{plugin}}$	-14.1	-0.28
LDA $\hat{\psi}^Q$	1.20	0.07
LDA $\hat{\psi}^{\text{plugin}}$	1.20	0.09
ATM $\hat{\psi}^Q$	1.17	0.08
ATM $\hat{\psi}^{\text{plugin}}$	1.17	0.08

Comments

- What if there are non-text confounding factors?
- What if the treatment is non-binary? Makes the separate neural network for each value of t impractical...
- Can we adapt this to identifying the causal effect of the text, rather than just using it as a control?
- Just gives point estimates, how can we run a hypothesis test?