Karlsruhe Institute of Technology

Bachelorarbeit

Automatische Auswahl von maschinellen Lernverfahren für kausale Inferenz

von

Atanas Dimitrov

Pervasive Computing Systems / TECO
Institut für Telematik
Fakultät für Informatik

Abgabadatum: 02.09.2019

Verantwortlicher Betreuer: Prof. Dr. Michael Beigl
Betreuerin: Ployplearn Ravivanpong

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Bachelorarbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe, die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht und weiterhin die Richtlinien des KIT zur Sicherung guter wissenschaftlicher Praxis beachtet habe.

Karlsruhe, 02.09.2019

Abstract

Diese Bachelorarbeit beschäftigt sich mit dem Vergleich von Methoden für kausale Inferenz und mit der automatischen Auswahl von dem besten von denen abhängig von dem vorhandenen Datensatz. Dazu benutzen und erweitern wir Synth-Validation - ein Verfahren, mit dem von den echten Daten synthetische Daten mit einem gewünschtem durchschnittlichen Behandlungseffekt erstellt und dann ausgewertet werden. Dabei haben die Datensätze, auf denen wir unsere Experimente durchführen, unterschiedliche Natur - echten Rohdaten mit größeren oder kleineren Zahl von Kovariaten, von Rohdaten synthetisch generierten Daten und zufällig generierten Daten. Die kausale Inferenz Verfahren, von denen Synth-Validation auswählt, benutzen ausschließlich Algorithmen aus dem maschinellem Lernen. Es wird die Fähigkeit von Synth-Validation gemessen, den Verfahren zu wählen, der die nächste Schätzung von dem durchschnittlichen Behandlungseffekt hat. Das wird unter unterschiedlichen Konstellationen unterstellt - nach der Art der Daten, nach der Anzahl der Elementen in der Stichprobe usw.

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einführung

1.1 Kausalität und kausale Inferenz

1.2 Motivation

1.3 Ziele und Methodik

2 Methoden für kausale Inferenz

2.1 Lineare Verfahren

2.1.1 Covariate Matching

2.1.2 Propensity Score Matching

2.1.3 Inverse Probability Weighting

2.2 Maschinelle Lernverfahren

2.2.1 Gradient Boosting

2.2.2 Lasso

2.2.3 Kausale Wälder

2.2.4 Targeted Maximum Likelihood Estimation

3 Synth-Validation

3.1 Generierung von synthetischen Daten

3.1.1 Auswahl von synthetischen Effekten

3.1.2 Schätzung von bedingten Erwartungswerten

3.2 Methodenauswahl

4 Implementierung

4.1 Fremde Bibliotheken

4.2 Lesen/Schreiben von Daten

4.3 Ziehen von Stichproben

4.4 Methoden für kausale Inferenz

4.5 Synth-Validation

4.5.1 Datenstrukturen

4.5.2 Schätzung

4.5.3 Constrained Boosting

4.5.4 Methodenauswahl

4.6 Benchmark von Synth-Validation

4.7 Erstellung von Abbildungen

4.8 Experimenten

4.9 Anderer Code

5 Ergebnisse und Evaluation

5.1 Methodik und Daten

5.2 TODO

6 Schlussfolgerung

6.1 Zusammenfassung

6.2 Diskussion

Appendices

References

[ATW+19] Susan Athey, Julie Tibshirani, Stefan Wager, et al. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019. No citations.

[SJT+17] Alejandro Schuler, Ken Jung, Robert Tibshirani, Trevor Hastie, and Nigam Shah. Synth-validation: Selecting the best causal inference method for a given dataset. *arXiv preprint arXiv:1711.00083*, 2017. No citations.

[WA18] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. No citations.