DS-GA 3001: TTML paper pointers and project ideas

David S. Rosenberg

ATE / CATE estimation

- The X-learner paper we discussed mentions that the poor coverage characteristics of the bootstrap confidence intervals may be due to the bias of the CATE estimator – is this avoidable? What leads to more or less bias? Do simpler base hypothesis spaces help or hurt? Can we fix things if we discretize the input space? In the limit of discretization, we're back to ATE estimation. [KSBY19]
- The X-learner paper uses honest random forests and BART as base learners. How is performance affected if we use base learners that are more common in machine learning, such as [regular] random forests, gradient boosting, neural networks, etc?
- In the X-learner paper, in the transphobia experiment, it seems there's big difference in the outcome of the X-RF and the T-RF, despite having treatment and control groups of the same size – what's going on here? What drives the difference? Can we reproduce this effect in a simulation to gain more insight?
- This paper does a lot of empirical comparisons on CATE estimators, but does not include X-learner, among many other things they exclude (citing computational constraints). Could be interested to expand their work.
- Explores issues with ATE estimation in high dimensions `https://arxiv.org/abs/1604.07125`.

Covariate shift

- A paper by Reddi et al. applies control variate methods to the covariate shift problem [RPS15].
- The paper What is the Effect of Importance Weighting in Deep Learning? (Byrd and Lipton) makes some claims about using importance weighting in the deep learning setting [BL18].

Online bandits and contextual bandits

- The paper that sparked interest in Thompson sampling in the ML community is Chapelle and Li's "An Empirical Evaluation of Thompson Sampling" [CL11]. They try some variants of Thompson sampling and compare to a UCB method. May be interested to expand their methods and their datasets. How does policy gradient work for various classes of ML models, for various approaches to baselines / control variates?
- This paper (and accompanying github code) is bake-off of a wide range of contextual bandit algorithms.
- Neural contextual bandits with UCB-based exploration by Zhou, Li, and Gu might be of interest.

Offline bandits / counterfactual evaluation

- Experiments in [DLL11] show direct methods performing rather poorly. [JS16, SJ15] seem to take it as given that these methods don't work well – in fact, they don't even compare to those methods. Is this an oversight? More recent work tells essentially the opposite story [BWRB20]. One possibility is that direct methods don't work well unless you have a sufficiently expressive hypothesis space (e.g. something like a deep neural network rather than a logistic regression) – of course for these large capacity models you need a reasonably large training set. So what's going on here? Is there a consistent performance pattern we can learn here?
- One potential issue with direct methods for policy learning is that we build our reward estimator using the logging policy, but as we optimize through policy space, we're applying the reward estimator to a different distribution of actions. Can we use importance weighted learning to make a better reward estimator? But then we'd have to have a different reward estimator everytime we move through policy space. That's essentially the idea of this paper: "Batch learning from bandit feedback through bias corrected reward imputation" [WBBJ19].
- In the missing data homework, we found that fitting the propensity score function data yielded better performance than the actual propensity score function. Can we get a similar gain from approximating the logging policy by learning from the data?
- Bayesian Counterfactual Risk Minimization – also tries out learning the logging policy.
- Effective Evaluation Using Logged Bandit Feedback from Multiple Loggers [ABSJ17]

- - Confident Off-Policy Evaluation and Selection through Self-Normalized Importance Weighting [KVGS20] is a recent paper that presents sophisticated lower confidence bounds on off-policy value estimates. Experimenting with this directly is of interest. Can one optimize these lower bounds, in the spirit of POEM and NORM-POEM from [SJ15], to get an new and improved policy learning algorithm?
- CAB: Continuous Adaptive Blending for Policy Evaluation and Learning (claims to be SOTA).[SWSJ19]
- On the Design of Estimators for Bandit Off-Policy Evaluation [VBDJ19]
- Doubly robust off-policy evaluation with shrinkage [SDKD20]
- Bandit Overfitting in Offline Policy Learning [BWRB20]
- (BanditNet) Deep Learning with Logged Bandit Feedback (ICLR 2018)
- AutoML for Contextual Bandits [DKC+19] is a fairly straightforward paper that argues that with sufficiently sophisticated reward prediction, the direct method we don't need methods like importance weighting. Is this true? Can you find datasets where this approach breaks? Even in the large-data setting?
- Advanced control variate methods for off-policy evaluation in this paper by Vlassis et al.
- More Robust Doubly Robust Off-policy Evaluation (PMLR 2018) [FCG18]
- The CoinDICE paper is about estimating confidence intervals for policy values. The paper is for general RL, but Appendix C gives specialization to contextual and multiarmed bandit.
- A whole slew of other papers referenced by Joachims' seminar on counterfactual machine learning.
- (Continuous action spaces) Policy Evaluation and Optimization with Continuous Treatments (AIStats 2018)
- Extremely large action spaces: [LDJ20]

Calibrated probability prediction

- Binary Classifier Calibration: Non-Parametric Approach [NCH14]
  - A nice paper blending theory and practice. Among other things, gives a theorem bounding the reduction in AUC performance from using a histogram binning calibrator. On the practical side, there's a discussion on what one should use for the calibration dataset – the training set, a holdout set, or something else.

- Obtaining Well Calibrated Probabilities Using Bayesian Binning and a lot more by Mahdi Pakdaman Naeini.
- Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration (NeurIPS 2019) [KPNK+19]
- Verified uncertainty calibration (NeurIPS 2019) [KLM19] gives a calibration method that has the sample efficiency of a parametric model but the calibration verifiability of histogram binning. Also, perhaps the bigger contribution, is that they show the "debiased estimator" of calibration from the meteorological literature is much more sample efficient.
- Spline approach [Luc18]
- Approach to calibrating deep neural networks: [MKS+20]

Data sets

- Here's a nice dashboard of Covid data – perhaps you can think of something interesting to do with it: https://covid.jerschow.com/I I don't have any specific ideas for projects relevant to this class (except perhaps feature importance that we'll get to later?). But the creator of the dashboard would be interested in discussing if you like.

References

[ABSJ17] Aman Agarwal, Soumya Basu, Tobias Schnabel, and Thorsten Joachims. Effective evaluation using logged bandit feedback from multiple loggers. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page nil, 8 2017.

[BL18] Jonathon Byrd and Zachary C. Lipton. What is the effect of importance weighting in deep learning? *CoRR*, 2018.

[BWRB20] David Brandfonbrener, William F. Whitney, Rajesh Ranganath, and Joan Bruna. Bandit overfitting in offline policy learning. *CoRR*, 2020.

[CL11] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, pages 2249–2257, Red Hook, NY, USA, 2011. Curran Associates Inc.

[DKC+19] Praneet Dutta, Man Kit, Cheuk, Jonathan S Kim, and Massimo Mascaro. Automl for contextual bandits. *CoRR*, 2019.

[DLL11] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 1097–1104, Madison, WI, USA, 2011. Omnipress.

[FCG18] Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. *CoRR*, 2018.

[JS16] Thorsten Joachims and Adith Swaminathan. Counterfactual evaluation and learning for search, recommendation and ad placement. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval - SIGIR '16*, SIGIR '16, page nil, - 2016.

[KLM19] A. Kumar, P. Liang, and T. Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[KPNK+19] Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 12316–12326. Curran Associates, Inc., 2019.

[KSBY19] Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

[KVGS20] Ilja Kuzborskij, Claire Vernade, András György, and Csaba Szepesvári. Confident off-policy evaluation and selection through self-normalized importance weighting. *CoRR*, 2020.

[LDJ20] Romain Lopez, Inderjit Dhillon, and Michael I. Jordan. Learning from extreme bandit feedback. *CoRR*, 2020.

[Luc18] Brian Lucena. Spline-based probability calibration. *CoRR*, 2018.

[MKS+20] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip H. S. Torr, and Puneet K. Dokania. Calibrating deep neural networks using focal loss. *CoRR*, 2020.

[NCH14] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Binary classifier calibration: Non-parametric approach. *CoRR*, 2014.

[RPS15] Sashank Reddi, Barnabas Poczos, and Alex Smola. Doubly robust covariate shift correction. In *AAAI Conference on Artificial Intelligence*, 2015.

[SDKD20] Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudik. Doubly robust off-policy evaluation with shrinkage. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9167–9176. PMLR, 13–18 Jul 2020.

[SJ15] Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3231–3239. Curran Associates, Inc., 2015.

[SWSJ19] Yi Su, Lequn Wang, Michele Santacatterina, and Thorsten Joachims. CAB: Continuous adaptive blending for policy evaluation and learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6005–6014. PMLR, 09–15 Jun 2019.

[VBDJ19] Nikos Vlassis, Aurelien Bibaut, Maria Dimakopoulou, and Tony Jebara. On the design of estimators for bandit off-policy evaluation. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6468–6476. PMLR, 09–15 Jun 2019.

[WBBJ19] Lequn Wang, Yiwei Bai, A. Bhalla, and T. Joachims. Batch learning from bandit feedback through bias corrected reward imputation. In *ICML Workshop on Real-World Sequential Decision Making*, 2019.