The Self-Normalized Estimator for Counterfactual Learning

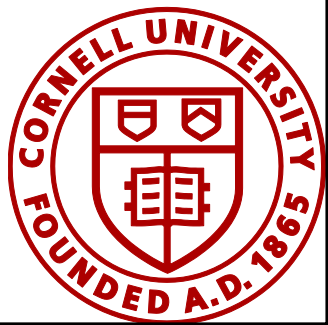
Adith Swaminathan, Thorsten Joachims

Cornell University

210 C #59

This estimator dominates the Horvitz-Thompson estimator when learning from bandit feedback because it avoids *propensity overfitting*

Software: http://www.cs.cornell.edu/~adith/POEM/



Setting: Batch Learning from Bandit Feedback

Use $\langle x_i, y_i, \delta_i \rangle$ logs to find good policy $h(y|x)$

Easier

Supervised learning

Harder

Off-policy RL

Not

Online Contextual Bandit

(~~explore~~-exploit)

Approach: Counterfactual Risk Minimization

Risk estimation via Importance Sampling:

$$\hat{R}(h) = \frac{1}{n}\sum_i \delta_i \frac{1}{\Pr(y_i|x_i)} h(y_i|x_i)$$

Feedback

Propensity

New policy

Learning via ERM:

$$\operatorname*{argmin}_h \hat{R}(h) + \lambda \mathrm{Reg}(h)$$

Change the estimator, stop ERM cheating!

To learn more, drop by **210 C #59**