Title

JT Cho
joncho@
seas.upenn.edu

Karinna Loo
kloo@
seas.upenn.edu

Veronica Wharton
whartonv@
seas.upenn.edu

1 Introduction

For our CIS 625 final project, our team — JT Cho, Karinna Loo, and Veronica Wharton — took a closer look at the topic of fairness in machine learning. The paper that piqued our interest was *Rawlsian Fairness for Machine learning* (Joseph et al., 2016), which describes two online algorithms in the linear contextual bandit framework that both learn at a rate comparable to (but necessarily worse than) the best algorithms absent of a fairness constraint and also satisfy a specified fairness constraint. The authors present theoretical and empirical results. Our team sought to reimplement the algorithms presented by Joseph et al. (2016) and then expand upon their empirical analyses. We were also interested in exploring further fairness analyses using real-world data.

2 Project overview

Our project consisted of the following steps:

1. We read the paper *Rawlsian Fairness for Machine Learning* (Joseph et al., 2016).
2. We implemented the TopInterval, IntervalChaining, and RidgeFair algorithms from the paper in Python.
3. We ran our implementations on a Yahoo! dataset containing a fraction of the user click log for news articles displayed in the Featured Tab of the Today Module on the Yahoo! Front Page during the first ten days in May 2009 (Yahoo!, 2009), to see how well they performed on real data.
4. To empirically evaluate our implementations, we ran experiments similar to those in (Joseph et al., 2016) with randomly-drawn contexts.
5. We compiled our findings into a written report.

3 Algorithm implementations

The code for our implementations can be found here: `https://github.com/jtcho/FairMachineLearning/blob/master/fairml.py`

4 Implementation: TopInterval

5 Implementation: IntervalChaining

6 Implementation: RidgeFair

7 Experimental results

We ran experiments that compared the regret of INTERVALCHAINING (IC) with the regret of TOPINTERVAL (TI). As in Joseph et al. (2016), we present three sets of empirical results:

- Varying T (the number of rounds) - we measured the average regret of INTERVALCHAINING and TOPINTERVAL as a function of increasing T.
- Varying k (the number of arms/groups) - we measured the average regret of INTERVALCHAINING and TOPINTERVAL as a function of increasing k.
- Varying d (the number of features) - we measured the average regret of INTERVALCHAINING and TOPINTERVAL as a function of increasing d.

For each increasing variable (T, k, or d), we present nine metrics as a function of the variable, each averaged over 500 trials. Contexts are drawn uniformly at random from $[0, 1]^d$ and standard Gaussian noise. Joseph et al. (2016) only present the average regret difference (metric #3).

1. Average regret (TI) - the average regret of TOPINTERVAL across all rounds.
2. Average regret (IC) - the average regret of INTERVALCHAINING across all rounds.

3. Average regret difference (TI vs. IC) - the difference between the average regrets of TOPINTERVAL and INTERVALCHAINING across all rounds.

4. Cumulative regret (TI) - the cumulative regret of TOPINTERVAL across all rounds.

5. Cumulative regret (IC) - the cumulative regret of INTERVALCHAINING across all rounds.

6. Cumulative regret difference (TI vs. IC) - the difference between the cumulative regrets of TOPINTERVAL and INTERVALCHAINING across all rounds.

7. Final regret (TI) - the regret of TOPINTERVAL in the final round.

8. Final regret (IC) - the regret of INTERVALCHAINING in the final round.

9. Final regret difference (TI vs. IC) - the difference between the final regrets of TOPINTERVAL and INTERVALCHAINING.

8 Conclusion

References

[Joseph et al.2016] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2016. Rawlsian fairness for machine learning. *CoRR*, abs/1610.09559.

[Yahoo!2009] Yahoo! 2009. Yahoo! front page today module user click log dataset. `https://webscope.sandbox.yahoo.com/catalog.php?datatype=r`. Accessed: 2017-04-03.