Differentially Private, Distributed Decision Trees

Kai Wen Wang, Nina Balcan, Travis Dick

November 1, 2018

Webpage: `https://kaiwenw.github.io/JuniorProject.html`

1 Introduction

In the age of information where every interaction with technology and the internet can track our data, privacy of personal data is becoming more relevant than ever before. Previous attempts at individual privacy, such as k-anonymity, do not introduce randomness and have well-known exploits. Differential privacy is a theoretic guarantee for an individual's data privacy when the data is used in training machine learning models. In the real world, differential privacy has recently been adopted by industry leaders, such as Apple [3], in a push for individual privacy.

Another increasingly significant area of research is distributed learning: data can be interspersed across many data-centers belonging to different organizations. Not only is it impractical to collate the data, it is often impossible due to incentives and contracts of the organizations to keep data private. As such, real world problem may not only necessitate distributed learning but also bring in new concerns of privacy, such as the privacy of a whole data center, when building a machine learning model. Therefore, a principled research about the combination of privacy and distributed learning is very relevant and important to many of today's applications of machine learning. I will explore this under the mentorship of Professor Nina Balcan and her Ph.D. student Travis Dick

Our choice to study privacy and distributed learning in the context of decision trees is not arbitrary. Decision trees are a well-understood class of machine learning classifiers that are interesting theoretically and commonly applied in practice. For example, Microsoft Kinect uses random forests, a variant of decision trees, for detecting a persons pose. Preliminary research has been done on differentially private decision trees [4] but only empirical results were shown. There is also a lack of foundational research on distributed decision trees from the perspective of communication complexity. Our work seeks to establish theoretical guarantees on communication complexity and provide empirical results of differentially private, distributed decision trees (DPDDT).

2 Project Goals and Milestones

The challenge of the work lies in finding a delicate balance between accuracy, privacy budget (a metric of differential privacy), and communication complexity, since these metrics are inherently in conflict of each other. For example, a common way of ensuring differential privacy is by adding Laplacian noise to the model's output, which intrinsically decreases accuracy. Limiting communication for the model may also have negative or positive effects accuracy and privacy. The tradeoff of privacy and accuracy in a distributed setting will be a major goal in the theoretic portion of the project. We aim to prove tight lower and upper bounds on the effects of that

each factor may have on another. By the end of the spring semester, we hope to have explored enough theoretical guarantees on accuracy, privacy budget, and communication complexity to have some meaningful conclusions. In addition to theory, we also aim to simulate a working implementation of Differentially Private, Distributed Decision Trees with non-trivial, real world datasets to empirically backup and quantitatively evaluate our theoretical results. It would be especially interesting to extend our algorithms for training regular decision trees to more complex decision tree based models like random forests.

If things are slower than expected, then we hope to have theoretical results and only empirical results on contrived and simple datasets. In this case, we still hope to have combined privacy and distributed learning in a meaningful way. If things go faster than expected, we hope to organize the experimental results, frame our research question in an appealing story, and produce a well-polished paper for a notable conference, such as the Conference on Neural Information Processing Systems (NIPS) which typically accepts papers until mid-May. A detailed milestone timeline of the project is listed in Table 1.

For literature in differential privacy, there is a comprehensive textbook on the subject, called The Algorithmic Foundations of Differential Privacy by Dwork and Roth [2]. So far, I've read through the first two chapters. Decision trees are quite well-studied and I have learned about them in 10-701 (Intro to ML at CMU). Since decision trees are very well-studied, there are many resources available online and in machine learning textbooks such as Machine Learning by Tom Mitchell [5]. For distributed learning, I've skimmed a detailed paper on the subject [1] and watched some lecture videos by Prof. Balcan. I would need to find more resources about distributed learning, especially the theoretical aspects of it. In terms of the implementation aspects of distributed learning, I will have a relevant background since I am taking Distributed Systems (15-440) currently. Software or hardware resources likely would not be a concern for this project. The only potential resources I may need would be some AWS credits to simulate a large distributed network of decision trees. As a student, AWS credits should be feasible to obtain.

Date	Milestone
Dec 15, 2018	Have solid theoretical understanding of key decision tree algorithms (ID3, C4.5, Random Forests). Go through half of the DP book and have a clear understanding of the significance, strengths and weaknesses of DP. Begin theoretical analysis of DP decision trees.
Feb 1, 2019	Done some theoretical analysis on DP with decision trees, such as designing a new DP learning model and proving lower and upper bounds on accuracy and privacy budget. Done reviewing literature for distributed learning.
Feb 15, 2019	Further theoretical results with DP decision trees. Have some basic experimental tests of last milestone. Begin theoretical analysis on distributed decision trees.
Mar 1, 2019	Done some theoretical analysis on distributed learning, such as designing a new distributed learning algorithm for decision trees, as well as lower and upper bounds on accuracy with communication complexity. This could be incompatible with our original DP algorithm, in which case begin thinking of new DPDDT algorithm.
Mar 22, 2019	Merged some theoretical results from DP and distributed learning. Begin thinking of new notions of privacy that arise from a distributed point of view. Done implementing a distributed decision tree.
Apr 5, 2019	Completed implementing DPDDT and tested on real datasets. Ideally have an idea of how to obtain some novel theoretical results that arise from the combination of privacy and distributed learning.
Apr 19, 2019	Finished theoretical results of DPDDT and begin performing final tests on DPDDT. If possible, try more sophisticated tests such as applied in computer vision.
May 3, 2019	Finish up any remaining proofs or testing. Should also have a manuscript that was continuously maintained throughout the semester. If possible, begin polishing the manuscript for submission to NIPS.

Table 1: Specific milestones for the end of fall semester and biweekly updates in the spring.

References

[1] Maria Florina Balcan, Avrim Blum, Shai Fine, and Yishay Mansour. Distributed learning, communication complexity and privacy. In *Conference on Learning Theory*, pages 26–1, 2012.

[2] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[3] Andy Greenberg. Apples differential privacyis about collecting your data–but not your data. *Wired (June 13, 2016)*, 2016.

[4] Geetha Jagannathan, Krishnan Pillaipakkamnatt, and Rebecca N Wright. A practical differentially private random decision tree classifier. In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, pages 114–121. IEEE, 2009.

[5] Tom M Mitchell et al. Machine learning. wcb, 1997.