A Survey on Differentially Private Machine Learning

I. Differentially Private Learning Model Evaluation

One of the core tasks in building a machine learning model is to evaluate its performance. Analysts must be able to evaluate the degree of fit of the hypothetical model to the original data, and assess the model's accuracy on unknown data. In the process of model evaluation, in order to ensure the privacy of users' personal information in the dataset, privacy protection needs to be performed on it, so that the attacker cannot obtain the private information related to the users from the released data. Therefore it is of vital importance to develop differentially private evaluation algorithms for machine learning models.

A. Evaluation Process

There are multiple stages in developing a machine learning model for use in a software application. It follows that there are multiple places where one needs to evaluate the model. Roughly speaking, the first phase, which is referred as prototyping, includes offline evaluation. In the deployed phase, online evaluation is necessary. Online evaluation measures live metrics of the deployed model on live data; offline evaluation measures offline metrics of the prototyped model on historical data (and sometimes on live data as well).

1.Validation and Parameter Tuning:

The available historical dataset is split into two parts: training and validation. The model training process receives training data and produces a model, which is evaluated on validation data.

In [1], they describe a simple and general differentially private method, together with two specific instantiations, for reusing a holdout set for validating results while provably avoiding overfitting to the holdout set. The analyst can perform any analysis on the training dataset, but can only access the holdout set via an algorithm that allows the analyst to validate her hypotheses against the holdout set. Crucially, their algorithm prevents overfitting to the holdout set even when the analyst's hypotheses are chosen adaptively on the basis of the previous responses of their algorithm.

In [2], they develop an end-to-end differentially private method for solving regression problems with convex penalty functions and selecting the penalty parameters by cross-validation. They show how a differentially private procedure for penalized logistic regression with elastic-net regularization can be applied to the analysis of GWAS data and evaluate our method's performance.

[3] introduces a generic differentially private validation procedure for differentially private machine learning algorithms that apply when a certain stability condition holds on the training algorithm and the validation performance metric. The training data size and the privacy budget used for training in their procedure is independent of the number of parameter values searched over. They apply their generic procedure to two fundamental tasks in statistics and machine-learning – training a regularized linear classifier and building a histogram density estimator that result in end-to-end differentially private solutions for these problems.

[4] develops new algorithmic techniques for learning and a refined analysis of privacy costs within the framework of differential privacy. They implementation and experiments demonstrate that they can train deep neural networks with non-convex objectives, under a modest privacy budget, and at a manageable cost in software complexity, training efficiency, and model quality.

[5] introduces methods for releasing the best hyper-parameters and classifier accuracy privately. Leveraging the strong theoretical guarantees of differential privacy and known Bayesian optimization convergence bounds, it proves that under a Gaussian process assumption these private quantities are often near-optimal. Even if this assumption is not satisfied, it can use different smoothness guarantees to protect privacy.

2.Hypothesis Testing:

The online phase has its own testing procedure. The most commonly used form of online testing is A/B testing, which is based on statistical hypothesis testing.

[6][7][8][9][10][11][12][13][14][15]

B. Performance Metrics

Different machine learning tasks have different performance metrics. There are different metrics for the tasks of classification, regression, ranking, clustering, topic modeling, etc. Some metrics, such as precision/recall, are useful for multiple tasks. Even with a model produced by a differentially private algorithm, directly reporting its performance on a database has the potential for disclosure. Thus, differentially private computation of evaluation metrics for machine learning is an important research area.

Any metric based on a single confusion matrix can be made private by applying the standard methods, such as Laplace noise, for differentially private counts or marginals [16]. Thus, differentially private accuracy, recall, specificity, precision, etc. can be obtained.In this subsection, we focus on more complex metrics that are both more useful for machine learning evaluation as well as more challenging to implement privately.

[4]

In [17], they find effective differentially private mechanisms for computing AUC, area under the receiver-operating characteristic(ROC) curve, ROC curve and average precision. They bound the maximum change in AUC between neighboring datasets to find the local sensitivity. However, local sensitivity itself is not suitable for creating differentially private algorithms since adding different amounts of noise for adjacent databases can leak information. Instead, they use $\beta - smoothsensitivity$ which ensures the scale of noise for adjacent databases is within a factor of e^{β}. Similarly, average precision is computed by smooth sensitivity too. To extend the work on AUC to ROC, they use a symmetric binormal ROC curve, i.e., a 1-parameter ROC curve estimator.

In [18], they develop differentially private diagnostics tools for regression. Specifically, they create differentially private versions of residual plots for linear regression and of receiver operating characteristic (ROC) curves as well as binned residual plot for logistic regression. The residual plot and binned residual plot help determine whether or not the data satisfy the assumptions underlying the regression model, and the ROC curve is used to assess the predictive power of the logistic regression model. These diagnostics improve the usefulness of algorithms for computing differentially private regression output, which alone does not allow analysts to assess the quality of the posited model. To generate differentially private ROC curves,they add noise directly to each of the points that make up the curve. However, this method requires the privacy budget to be split among all points in the curve.

C. Other Methods

In addition to the commonly used performance metrics mentioned below, researchers proposes some new thoughts on comparing various machine learning algorithms and models and how to choose the best from them under differential private protection, such as [19] [20]. In this subsection, we give a brief introduction to these evaluation methods.

[19] proposes Pythia, an end-to-end differentially private mechanism for achieving near-optimal error rates using a suite of available privacy algorithms. Pythia is a meta-algorithm, which safely performs automated Algorithm Selection and executes the selected algorithm to return a differentially private result. Using Pythia, data curators do not have to understand available algorithms, or analyze subtle properties of their input data, but can nevertheless enjoy reduced error rates that may be possible for their inputs.

Pythia works in three steps. First it privately extracts a set of feature values from the given input. Then, using a Feature-based Algorithm Selector Pythia chooses a differentially private algorithm A^* from a collection of available algorithms. Lastly, it runs A^* on the given input. An important aspect of this approach is that Pythia does not require intimate knowledge of the algorithms from which it chooses, treating each like a black-box. This makes Pythia extensible, easily accommodating new advances from the research community as they appear.

[20] studies model selection in multivariate linear regression under the constraint of differential privacy. It shows that model selection procedures based on penalized least squares or likelihood can be made differentially private by a combination of regularization and randomization, and proposes two algorithms to do so. It also shows that its private procedures are consistent under essentially the same conditions as the corresponding non-private procedures. Besides, it finds that under differential privacy, the procedure becomes more sensitive to the tuning parameters.

REFERENCES

[1] C. Dwork, M. Hardt, M. Hardt, O. Reingold, O. Reingold, and A. Roth, "Generalization in adaptive data analysis and holdout reuse," in *International Conference on Neural Information Processing Systems*, 2015, pp. 2350–2358. I-A

[2] F. Yu, M. Rybar, C. Uhler, and S. E. Fienberg, *Differentially-Private Logistic Regression for Detecting Multiple-SNP Association in GWAS Databases*. Springer International Publishing, 2014. I-A

[3] K. Chaudhuri and S. Vinterbo, "A stability-based validation procedure for differentially private machine learning," *Advances in Neural Information Processing Systems*, pp. 2652–2660, 2013. I-A

[4] M. Abadi, A. Chu, I. Goodfellow, H. B. Mcmahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," pp. 308–318, 2016. I-A, I-B

[5] M. J. Kusner, J. R. Gardner, R. Garnett, and K. Q. Weinberger, "Differentially private bayesian optimization," in *International Conference on International Conference on Machine Learning*, 2015, pp. 918–927. I-A

[6] A. F. Barrientos, J. P. Reiter, A. Machanavajjhala, and Y. Chen, "Differentially private significance tests for regression coefficients," 2017. I-A

[7] D. Vu and A. Slavkovic, "Differential privacy for clinical trial data: Preliminary evaluations," in *IEEE International Conference on Data Mining Workshops*, 2009, pp. 138–143. I-A

[8] R. Hill, M. Hansen, E. Janssen, S. A. Sanders, J. R. Heiman, and L. Xiong, "A quantitative approach for evaluating the utility of a differentially private behavioral science dataset," in *IEEE International Conference on Healthcare Informatics*, 2014, pp. 276–284. I-A

[9] M. Gaboardi, H. W. Lim, R. Rogers, and S. Vadhan, "Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing," pp. 2111–2120, 2016. I-A

[10] Y. Wang, J. Lee, and D. Kifer, "Differentially private hypothesis testing, revisited," *Statistics*, vol. 22, no. 5, pp. 821 – 825, 2015. I-A

[11] O. Sheffet, "Differentially private least squares: Estimation, confidence and rejecting the null hypothesis," 2015. I-A

[12] C. Dwork, W. Su, and L. Zhang, "Private false discovery rate control," *Computer Science*, 2015. I-A

[13] V. Karwa, A. B. Slavkovi?, and P. Krivitsky, "Differentially private exponential random graphs," in *Privacy in Statistical Databases*, 2014. I-A

[14] C. Uhlerop, A. Slavkovi?, and S. E. Fienberg, "Privacy-preserving data sharing for genome-wide association studies." *Journal of Privacy & Confidentiality*, vol. 5, no. 1, p. 137, 2013. I-A

[15] S. E. Fienberg, A. Slavkovic, and C. Uhler, "Privacy preserving gwas data sharing," in *IEEE International Conference on Data Mining Workshops*, 2011, pp. 628–635. I-A

[16] C. Dwork, "Differential privacy," in *International Colloquium on Automata, Languages, and Programming*, 2006, pp. 1–12. I-B

[17] K. Boyd, E. Lantz, and D. Page, "Differential privacy for classifier evaluation," in *ACM Workshop on Artificial Intelligence and Security*, 2015, pp. 15–23. I-B

[18] Y. Chen, A. F. Barrientos, A. Machanavajjhala, and J. P. Reiter, "Is my model any good: differentially private regression diagnostics," *Knowledge & Information Systems*, vol. 54, no. 1, pp. 1–32, 2017. I-B

[19] I. Kotsogiannis, A. Machanavajjhala, M. Hay, and G. Miklau, "Pythia: Data dependent differentially private algorithm selection," in *ACM International Conference on Management of Data*, 2017, pp. 1323–1337. I-C

[20] S. A. Lei J, Charest A S, "Differentially private model selection with penalized and constrained likelihood," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 181, no. 3, pp. 609–633, 2018. I-C