Feature Selection From Differentially Private Correlations

AUTHORS

Ryan Swope, Amol Khanna, Philip Doldo, Edward Raff

Saptarshi Roy

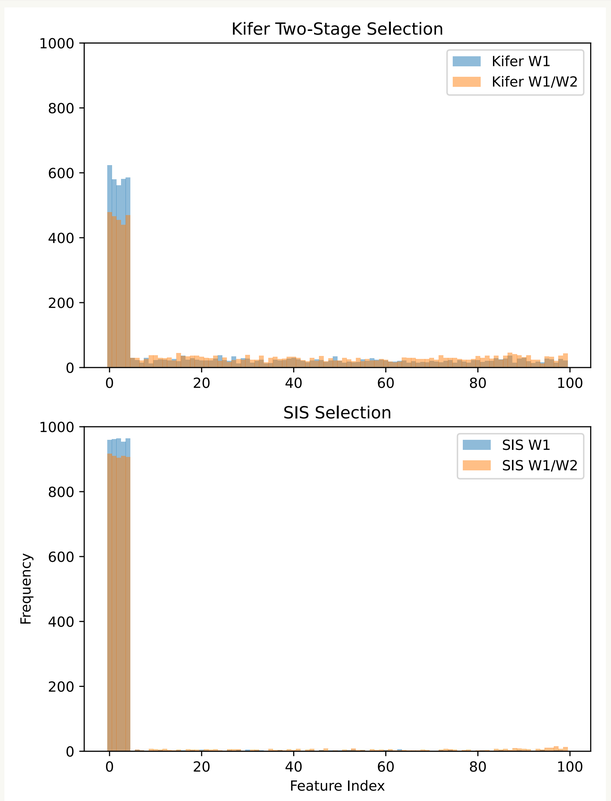
AFFILIATIONS

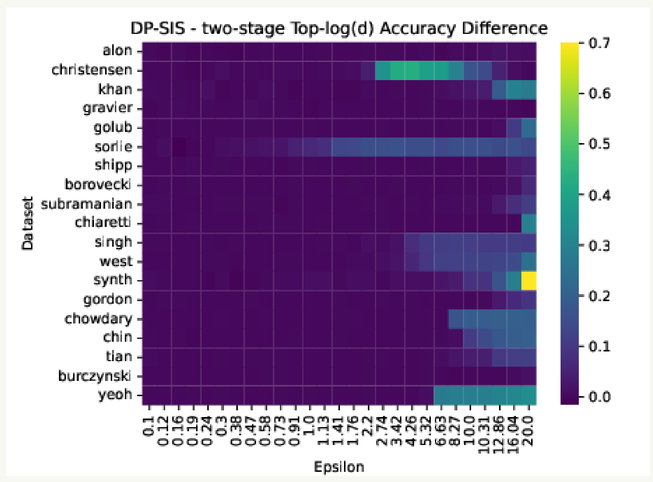
Booz Allen Hamilton

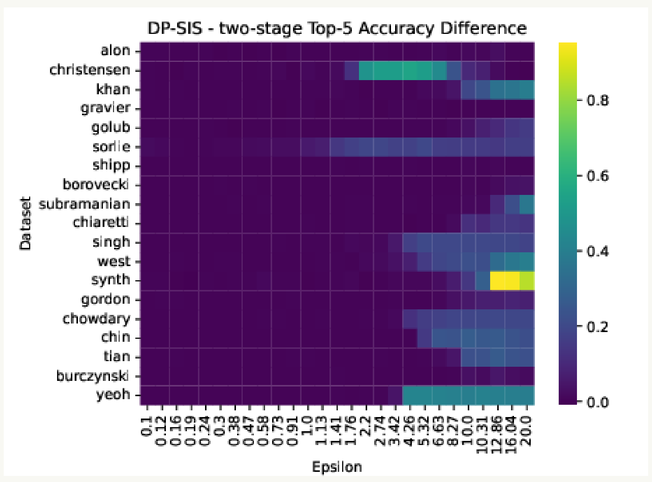
University of Michigan

ABSTRACT

Data scientists often seek to identify the most important features in high-dimensional datasets. This can be done through $L1$-regularized regression, but this can become inefficient for very high-dimensional datasets. Additionally, high-dimensional regression can leak information about individual datapoints in a dataset. In this paper, we empirically evaluate the established baseline method for feature selection with differential privacy, the two-stage selection technique, and show that it is not stable under sparsity. This makes it perform poorly on real-world datasets, so we consider a different approach to private feature selection. We employ a correlationsbased order statistic to choose important features from a dataset and privatize them to ensure that the results do not leak information about individual datapoints. We find that our method significantly outperforms the established baseline for private feature selection on many datasets.

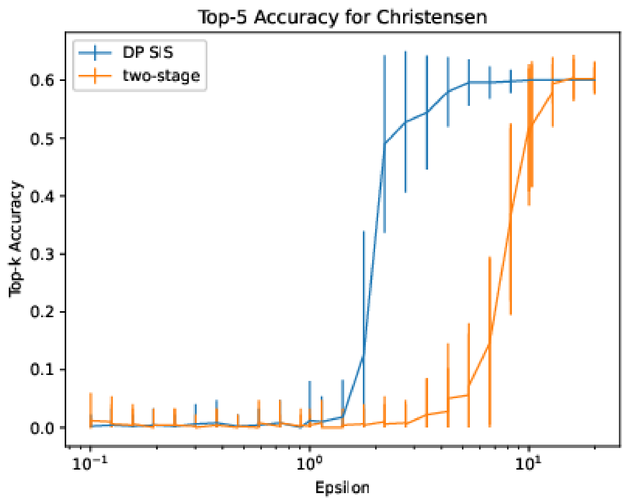


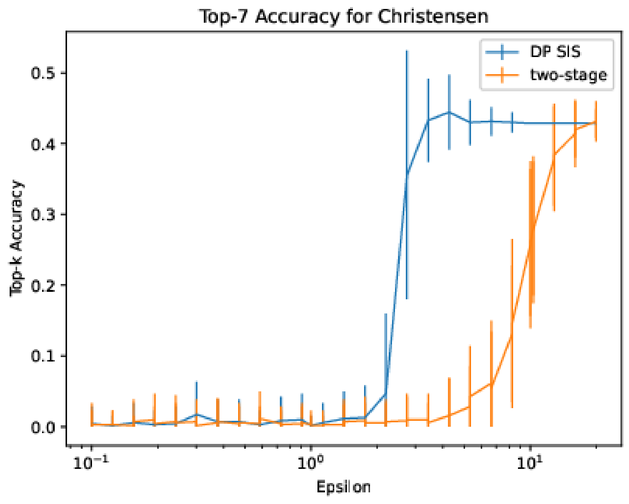


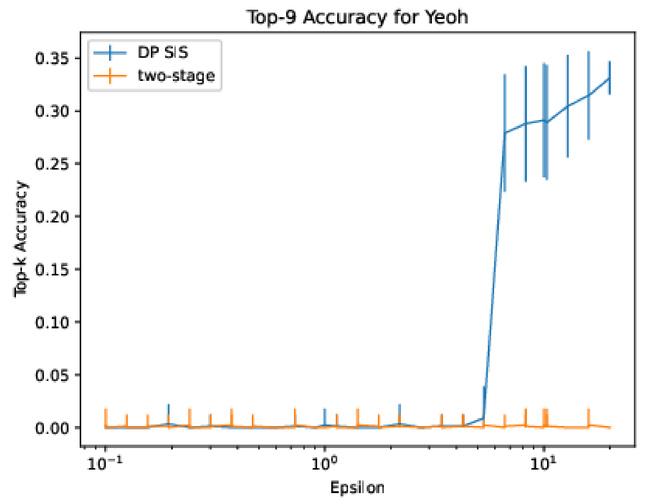


EXPERIMENTS

We show that the support selection of the Kifer two-stage approach is numerically unstable (above, left), and that for a number of high dimenstionality datasets, DP-SIS with canonical Lipschitz achieves similar or higher top-k accuracy than two-stage. The heatmaps above show this behavior at a high level, while the figures below show a few examples over a practical range of ϵ.







CONCLUSION

This paper seeks to improve upon the state-of-the-art in computationally feasible differentially private feature selection for highdimensional linear regression. We identify that the current computationally feasible method, the two-stage approach, requires building $\sqrt{N}$ estimators on disjoint partitions of a dataset to identify the most commonly selected features in these estimators. While this is computationally feasible, it is still difficult since it requires building many high-dimensional estimators. Additionally, it requires high-dimensional sparse regression estimators to be algorithmically stable, which they are not. SIS is an alternative method for high-dimensional feature selection. It selects the k features which are most correlated with the target vector. By making SIS private with a differentially private top-k selector, we can develop a differentially private feature selector based on SIS. We choose to use DP-SIS based on its superior performance. SIS outperforms the two-stage method on most datasets in reasonable ranges of ϵ. Additionally, DP-SIS can be used with any feature selection metric, making it flexible to improved metrics for feature selection developed in the future. Finally, we end with a final comment. This paper explores employing a simple metric for feature selection in differential privacy, and pits it against a more complicated mechanism. The simpler metric works better despite it not being able to identify cases in which subsets of features are individually weakly correlated with the target but jointly strongly correlated with the target. This is because when using differential privacy, there is a constant tugof-war between employing more expressive methods with more noise and simpler methods with less noise. In the case presented in this paper, feature selection had higher accuracy when a simpler method was selected which was more stable and required less noise.