CSCE658 S2024: Possible Class Readings and Projects

1 Coresets and Dimensionality Reduction for Data Science

Coresets and dimensionality reduction are two ways to decrease the effective size of a dataset for a specific task. A coreset is a subset of weighted points of the dataset, with dimensionality reduction decreases the number of features of the dataset. The goal is to improve the runtime of data science algorithms due to the smaller effective size of the dataset, without sacrificing too much accuracy with respect to the original dataset. With the improved runtimes, the model can then be trained on larger datasets, which can then actually improve overall accuracy. Potential papers of interest:

(1) [BLG+19] studies the application of coresets to neural networks, so that after the weights of each neuron are trained, a number of edges and nodes are subsequently pruned, leading to faster evaluation times of future inputs

(2) [MOB+20] studies the application of coresets to neural networks, where the neurons are pruned before the neural network is trained, leading to faster training and evaluation times of future inputs

(3) [TZM+23] studies coresets for radial basis function neural networks, which can approximate any continuous function

(4) [LBL+20] applies sampling-based approaches to improve the performance of convolutional neural networks (CNNs)

2 Social Aspects of Algorithmic Design

Often we would like additional functionality from our algorithms.

(1) [DYZH21] surveys fairness in deep learning work from the computational perspective

(2) [CKLV17] presents an algorithm for socially fair k-means clustering

(3) [AEKM20] studies the problem of fair correlation clustering

(4) [BC20] shows that the problem of histogram estimation has different behaviors in the central setting of differential privacy and the shuffle model of differential privacy

(5) [CGK+23] studies the problem of releasing the values of all-pairs shortest-paths in the central setting of differential privacy

3 Oracles and Learning-Augmented Algorithms

Learning-augmented algorithms or data-driven algorithm design is the study of incorporating possibly erroneous external advice into algorithmic design. The goal is to perform better than oblivious algorithms if the advice is good but not lose significant algorithmic performance if the advice is bad. Potential papers of interest:

(1) [EFS+22] studies the incorporation of advice for clustering data to achieve *more accurate* approximation algorithms

(2) [LLW22] incorporates advice for binary search trees to achieve *faster* query times

(3) [HIKV19] studies the incorporation of advice for frequency estimation in the streaming model to achieve algorithms with *less* memory usage, given additional distributional assumptions

(4) [Mit18] studies the incorporation of advice for improving the *accuracy* of bloom filters, which are data structures for membership queries, i.e., answering queries on whether certain datasets are contained within a dataset

(5) [CKT19] studies the problem of submodular optimization when an evaluation oracle may not output the exact value but rather an approximate value

References

[AEKM20] Sara Ahmadian, Alessandro Epasto, Ravi Kumar, and Mohammad Mahdian. Fair correlation clustering. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS*, pages 4195–4205, 2020.

[BC20] Victor Balcer and Albert Cheu. Separating local & shuffled differential privacy via histograms. In *1st Conference on Information-Theoretic Cryptography, ITC*, pages 1:1–1:14, 2020.

[BLG+19] Cenk Baykal, Lucas Liebenwein, Igor Gilitschenski, Dan Feldman, and Daniela Rus. Data-dependent coresets for compressing neural networks with applications to generalization bounds. In *7th International Conference on Learning Representations, ICLR*, 2019.

[CGK+23] Justin Y. Chen, Badih Ghazi, Ravi Kumar, Pasin Manurangsi, Shyam Narayanan, Jelani Nelson, and Yinzhan Xu. Differentially private all-pairs shortest path distances: Improved algorithms and lower bounds. In *Proceedings of the 2023 ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 5040–5067, 2023.

[CKLV17] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, pages 5029–5037, 2017.

[CKT19] Victoria G. Crawford, Alan Kuhnle, and My T. Thai. Submodular cost submodular cover with an approximate oracle. In *Proceedings of the 36th International Conference on Machine Learning, ICML*, pages 1426–1435, 2019.

[DYZH21] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. Fairness in deep learning: A computational perspective. *IEEE Intell. Syst.*, 36(4):25–34, 2021.

[EFS+22] Jon C. Ergun, Zhili Feng, Sandeep Silwal, David P. Woodruff, and Samson Zhou. Learning-augmented k-means clustering. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[HIKV19] Chen-Yu Hsu, Piotr Indyk, Dina Katabi, and Ali Vakilian. Learning-based frequency estimation algorithms. In *7th International Conference on Learning Representations, ICLR*, 2019.

[LBL+20] Lucas Liebenwein, Cenk Baykal, Harry Lang, Dan Feldman, and Daniela Rus. Provable filter pruning for efficient neural networks. In *8th International Conference on Learning Representations, ICLR*, 2020.

[LLW22] Honghao Lin, Tian Luo, and David P. Woodruff. Learning augmented binary search trees. In *International Conference on Machine Learning, ICML*, volume 162, pages 13431–13440, 2022.

[Mit18] Michael Mitzenmacher. A model for learned bloom filters and optimizing by sandwiching. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2018.

[MOB+20] Ben Mussay, Margarita Osadchy, Vladimir Braverman, Samson Zhou, and Dan Feldman. Data-independent neural pruning via coresets. In *8th International Conference on Learning Representations, ICLR*, 2020.

[TZM+23] Murad Tukan, Samson Zhou, Alaa Maalouf, Daniela Rus, Vladimir Braverman, and Dan Feldman. Provable data subset selection for efficient neural networks training. In *International Conference on Machine Learning, ICML*, pages 34533–34555, 2023.