東京大学 大学院ゼミナールNo.3

「A Random Matrix Approach and Improvement of Machine Learning for Very Large Data」

2019年12月11日(水) 13:00~ 於 農学部1号館1階100番教室

生物測定学研究室　修士課程1年　Dang Tung Thanh

Abstract

Statistical Machine Learning has undergone a phase transition from a pure academic endeavor to being one of the drivers of modern commerce and science. Currently, data repositories for such applications currently exceed exabytes and are rapidly increasing in size. Beyond their sheer magnitude, these datasets and associated applications' considerations pose significant challenges for theory and practical software development of machine learning approaches. Recent results such as those on tera-scale learning and on very large neural network suggest that scale is a very important ingredient in quality modeling.

N. Halko et al (2011)[1] followed and developed the randomization algorithms that offer a set of powerful tools for performing low-rank matrix approximation. These techniques exploited modern computational architectures more fully than classical methods and opened the possibility of dealing with truly massive data sets. This paper presented a modular framework for constructing randomized algorithms that compute partial matrix decompositions. These methods used random sampling to identify a subspace that captures most of the action of a matrix. The input matrix was then compressed—either explicitly or implicitly—to this subspace, and the reduced matrix was manipulated deterministically to obtain the desired low-rank factorization.

Nathan Halko et al (2011)[2] applied the popularized randomized methods for principal component analysis (PCA) which is among the most popular tools in machine learning, statistics, and data analysis more generally in the case of very large data set. When data sets were too large to be stored in the random-access memory (RAM) of a typical computer system, these randomized algorithms showed a lot of important advances. For example, they required only a couple of iterations to produce nearly optimal accuracy, with overwhelmingly high probability. Moreover, they achieved some improvements in standard computations such as Out-of-core computations and computational costs.

Kevin J. Galinsky et al (2016)[3] introduced a method that infers selection with very large sample size using principal components (PCs) by identifying variants whose differentiation along top PCs is significantly greater than the null distribution of genetic drift. This research tried to employ recent advances in random matrix theory to accurately approximate top PCs while reducing time and memory cost from quadratic to linear in the number of individuals, a computational improvement of many orders of magnitude. This method was applied on 54,734 individuals of European descent from the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort; it required only 57 min of compute time and 2.6 GB of RAM for this analysis, orders of magnitude better than any other publicly available software. Moreover, using the PC-based test for natural selection, this paper replicated previously known selected loci and identify three new genome-wide significant signals of selection, including selection in Europeans at ADH1B.

To continue the impressive successes of the combination between machine learning and Random Matrix Theory (RMT), Cosme Louart (2018)[4] proposed an original random matrix-based approach to understand the end-to-end regression performance of single-layer random artificial neural networks, sometimes referred to as extreme learning machines, when both dimensions of dataset are large and scale proportionally with the number n of neurons in the network. This approach had several interesting features both for theoretical and practical considerations. It was first one of the few known attempts to move the random matrix realm away from matrices with independent or linearly dependent entries. In terms of practical applications, the findings of this paper shed light on the already incompletely understood extreme learning machines which have proved extremely efficient in handling machine learning problems involving large to huge datasets at a computationally affordable cost.

Xiaoyi Mai et al. (2018)[5] provided a quantitative performance study of the generalized graph-based semi-supervised algorithm for large dimensional Gaussian mixture data and radial kernels, technically following the random matrix approaches. The main contribution of this paper is that due to the large data assumption, most of the intuition leading up to the aforementioned algorithms collapse as both dimensions of data set go to infinite at a similar rate, and few algorithms remain consistent in this regime. Thus, corrective measures and a new data-driven parametrization scheme were proposed along with a theoretical analysis of the asymptotic performances of the resulting approach. As a result, significant performance gains are observed on practical data classification using the proposed parametrization.

Zhenyu Liao et al. (2019)[6] proposed the random matrix theory to improve the large dimensional performance analysis of kernel least squares support vector machines (LS-SVMs). The main findings were that, in the big data regime and under suitable conditions on the input statistics, a nontrivial asymptotic classification error rate (i.e., neither 0 nor 1) can be obtained and the decision function of LS-SVM converges to a Gaussian random variable whose mean and variance depend on the statistics of the two different classes as well as on the behavior of the kernel function. Most importantly, the analysis of the paper provided a deeper understanding of the mechanism into play in SVM-type methods and in particular of the impact on the choice of the kernel function as well as some of their theoretical limits in separating high-dimensional Gaussian vectors.

Benjamin Erichson et al. (2019)[7] followed randomized matrix algorithms to present the R package that provided the randomized singular value decomposition functions to accelerate the computation of principal component analysis (PCA) and robust principal component analysis (RPCA). More generally, the concept of randomness allows also one to efficiently compute modern matrix decompositions such as the interpolative decomposition (ID) and CUR decomposition.

References

1. Halko, Nathan, Per-Gunnar Martinsson, and Joel A. Tropp. "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions." SIAM review 53.2 (2011): 217-288.
2. Nathan, et al. "An algorithm for the principal component analysis of large data sets." SIAM Journal on Scientific computing 33.5 (2011): 2580-2594.
3. Galinsky, Kevin J., et al. "Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia." The American Journal of Human Genetics 98.3 (2016): 456-472.
4. Louart, Cosme, Zhenyu Liao, and Romain Couillet. "A random matrix approach to neural networks." The Annals of Applied Probability 28.2 (2018): 1190-1248.

5. Mai, Xiaoyi, and Romain Couillet. "A random matrix analysis and improvement of semi-supervised learning for large dimensional data." The Journal of Machine Learning Research 19.1 (2018): 3074-3100.
6. Liao, Zhenyu, and Romain Couillet. "A large dimensional analysis of least squares support vector machines." IEEE Transactions on Signal Processing 67.4 (2019): 1065-1074.
7. Erichson, N. Benjamin, et al. "Randomized Matrix Decompositions Using R." Journal of Statistical Software 89.1 (2019): 1-48.