Learning in Wasserstein Space

Jianbo Ye
College of Information Sciences and Technology
The Pennsylvania State University
University Park, PA 16801, USA
jxy198@ist.psu.edu

ABSTRACT

Learning from empirical probability measures is an emerging problem that has potential benefiting multiple domains. My research focuses on developing scalable and effective learning algorithms that handle large-scale data in form of measures. In particular, the Wasserstein space provides a powerful geometry that houses the compositions of data, which can be of great interest to domain experts that analyze unstructured data.

1. INTRODUCTION

Traditional machine learning algorithms often assume data is represented by a set of feature vectors, where methods such as clustering, classifications, and some generative models are applied to efficiently gain non-trivial understandings from data. However, when each instance/object one deal with is of high compositional complexity, it is not a free food that each can be converted into a single vector. In practice, these objects are naturally stored as sets, trees, graphs, and sequences. Large schools of work by computer scientists and domain experts develop discriminative and informative features for object descriptions and representations, such that off-the-shelf machine learning toolbox can be easily applied. A fundamental question raised here is "can we study those highly complicated objects, e.g. exploring patterns, without converting them into feature vectors in the first place", or "can machine learning directly deal with such non-vectorizable data?"

The answer seems obvious, as there is a family of learning models that do not explicitly depend on vector representations. For example, affinity-based manifold learning captures the geometry of data by exploring an affinity graph constructed from instance pairs. However, it is known that manifold approaches scale poorly with the growing size of data, and spectral algorithms they adopt can often induce large approximation errors. The scalability issue is also a technical hurdle for other affinity-based methods, such as kernel machines and some clustering methods, especially when a highly expensive affinity or distance function is utilized.

One actively pursued topic of my research is about unsupervised learning from empirical probability measures with an emphasis on its computational scalability. Instead of treating each object or instance as a vector, the new data-driven framework that I am looking into is a more flexible nonparametric framework that assumes a cloud of weighted points, called empirical measure, represents each instance. Such non-parametric representation eliminates the restriction of a fixed quantization codebook across instances, and readily allows data sparsity. Given what has been done for vector data, we have seen a growing trend to retrieve patterns from a set of empirical probabilities rather than just points. Representative is the family of Wasserstein geometrical methods (e.g., [1,4,5]) and kernel methods for probabilities (e.g. [3]).

2. Discrete Distribution Clustering

To promote the emerging framework aforementioned, my recent work develops scalable clustering algorithms that can efficiently group large-scale discrete distributions under Wasserstein distances, collectively referred to D2-Clustering [1,2], which can be viewed as the "K-means" counterpart for probability measures. As a metric, Wasserstein distance (or earth mover's distance) defines a powerful geometry to compare distributions by calculating the optimal mass transportation from one to the other. This power comes, however, with high computational price: solving optimal transport does not admit closed form solution and its cost scales poorly with the size of support points, aka the index of instance complexity. This cost has brought even higher computational barrier for implementing sophisticated learning algorithms in spirit of the Wasserstein geometry, when a large set of discrete measures are analyzed. Classical methods such as averaging, clustering, independent or principal component analysis, and classification do not have immediate feasible extensions for distributions. It is especially the case when we are to further scale up any of the three: dimensionality of space, support size of distribution, or number of distributions. Benefit from the recent advance of numerical methods (e.g. entropic regularization and Bregman ADMM) as well as powerful parallel computing infrastructures, optimal transport, originally heavily explored in imaging science and graphics, has attracted new surges for its machine learning potential. With an expertise in numerical techniques and continuing supports from my advisors Prof. James Z. Wang and Prof. Jia Li, my recent work, AD2-Clustering [1], reduces the computation of D2-Clustering by introducing new numerical solutions. AD2-Clustering for the first time practically handle the clustering of many large distribution-valued datasets, such as protein sequences, density images, and documents, making them solvable within a reasonable estimate of computation time using today's hardware. Promising results on those datasets have been reported. Particularly, another recently submitted work of mine achieves state-of-the-art performance of document clustering upon six representative datasets with varying lengths and difficulty levels, compared to nine methods covering a wide spectrum of methodologies used in relevant literature [6]. This work validates the significance of D2-clustering, complements our understanding of document analysis, and creates a window to further examine the use of word embeddings.

3. Proposed Work

In data science and machine learning realm, it is often a challenge to develop data-driven solutions that meaningfully exploit unstructured and complicated instances. AD2-Clustering is thus a straightforward yet important tool to explore those data. In the new workflow of AD2-Clustering, one will focus on composing task sensitive features for basic elements that make up a complicated instance. The properties of their high level compositions are then automatically inferred within the Wasserstein framework. This highlights its practicality and

flexibility on many real-world tasks compared to the traditional ways directly handcrafting the high-level features and clustering them afterwards. In summary, I perceive those machine learning methods in spirit of learning from probability measures can be of wide interest for generating innovative and effective data-driven solutions in both research and industry; yet, their computational feasibility needs much further investigations, provided what has been done till today. There are many related problems that remain unsolved including but not limited to robust methods with new perspectives, streaming methods for stochastic learning, mixture models for distributions, integrations with existing classifiers, and efficient approximations for domain specific data. Put aside the common nonparametric setting of distribution-valued data, some investigations might also be put into the perspective that examines parametric and semi-parametric probabilistic measures. To name a few of them, I am also interested in Gaussian mixture models and hidden Markov models, and how to exploit their Wasserstein inspired Riemannian structures in related applications.

4. REFERENCES

1. Ye, J., Wu P., Wang, J. Z., and Li, J. 2015. Accelerated discrete distribution clustering under Wasserstein distance. *Submitted to IEEE Trans. Signal Processing (TSP).* URL=http://arxiv.org/abs/1510.00012.

2. Li, J. and Wang, J. Z. 2008. Real-time computerized annotation of pictures. *IEEE Trans. Pattern Anal. & Machine Intelligence (TPAMI)*. 30, no. 6 (2008): 985-1002. DOI= http://dx.doi.org/10.1109/TPAMI.2007.70847.

3. Muandet, K., Fukumizu, K., Dinuzzo, F., and Scholkopf, B. 2012. Learning from distributions via support measure machines. In *Advances in Neural Information Processing Systems (NIPS)*. (2012): 10-18. URL=http://papers.nips.cc/paper/4825-learning-from-distributions-via-support-measure-machines.

4. Cuturi, M. and Doucet, A. 2014. Fast computation of Wasserstein barycenters. *Proc. Int. Conf. Mach. Learn. (ICML)*. (2014): 685-693. URL=http://www.jmlr.org/proceedings/papers/v32/cuturi14.html

5. Seguy, V. and Cuturi, M. 2015. An algorithmic approach to compute principal geodesics in the Wasserstein space. In *Advances in Neural Information Processing Systems (NIPS)*. (2015): to appear. URL= http://arxiv.org/abs/1506.07944

6. Ye, J., Li Y., Wu Z., Wang J.Z., Li J. 2015. Quantitative measuring gains acquired from word embeddings — a non-parametric unsupervised perspective. *Submitted to TACL.* (2016).