Latent Variable Models at Scale: Modeling, Regularization, Scalability and Privacy

Pengtao Xie, Machine Learning Department, Carnegie Mellon University

One central task in machine learning (ML) is to extract hidden knowledge and structure from observed data. This task is especially valuable with the prosperity of big data, where the data itself is of limited value unless knowledge can be uncovered therefrom. Latent variable models (LVMs) elegantly fit into this task where the observed variables are used to model data and the latent variables are utilized to characterize the hidden knowledge, and have been widely used in text mining, computer vision, computational biology, recommender system, to name a few. In the realm of big data, the complexity of knowledge and the size of data are way beyond those in small scale data analytics, presenting various challenges for latent variable modeling. In my thesis work, I aim to design, regularize and solve LVMs to effectively and efficiently extract knowledge and structure from big data.

Research Summary

My research targets several key issues in large scale latent variable modeling, including model design, regularization, scalability and privacy, and has proposed

- LVMs to address various modeling issues in big data analytics, such as integrating correlated tasks [12,8], handling multi-modal data [13], incorporating external knowledge [7,6];
- A diversity regularization approach to effectively capture long-tail factors in latent knowledge [1] and reduce the complexity of LVMs without sacrificing their expressiveness [4];
- Distributed machine learning systems that facilitate the efficient learning and inference of LVMs at large scale, with low communication cost and nice convergence guarantees [9,5,3,2];
- Privacy preserving techniques that allow LVMs to run over encrypted data on cloud servers, enjoying the benefits of cloud machine learning without compromising users’ privacy [11,10].

Design LVMs for Complex Modeling

Big data incurs not only the dramatically increased data volume, but also larger complexity of data structure (such as multi-modality [13]), correlation of multiple ML tasks [12,8], rich external knowledge [7,6] and so on, rendering its modeling to be complicated and challenging. I designed LVMs to address these complex modeling issues. To learn distance measures for multi-modal data, I designed a supervised LVM [13], which is built upon a multi-wing harmonium model that embeds heterogeneous feature modalities into a single unified latent space to capture the central characteristics of data. Considering that representation learning and clustering are closely related and can mutually benefit each other, I built LVMs [12,8] to integrate these two tasks into one framework where they are performed in a joint manner to improve the overall performance. In light of external word correlation knowledge can improve the coherence of topic modeling, I designed a Markov Random Field regularized topic model [7] that can effectively leverage the correlations between words to boost topic quality.

Diversity Regularization

The hidden knowledge and structure behind data usually comprises of multiple factors, such as topics
underlying document collection and motifs in biopolymer sequences. In big data, the popularity (or frequency) of these latent factors is usually distributed in a power-law fashion, where a few dominant factors occur very frequently while most factors (in the long-tail region) are of low frequency. These long-tail factors carry crucial information, but unfortunately the standard LVMs are deficient in capturing them. To address this issue, I proposed a diversity regularization approach [1] which encourages the components in LVMs to diversely spread out to improve the coverage of long-tail factors. Another merit of this diversity regularizer is its ability to reduce model complexity without compromising expressiveness [4]. I designed an angle based diversity regularizer, proposed an efficient optimization algorithm and analyzed the theoretical behaviors of this regularization approach.

Large Scale Distributed ML Systems

LVMs are usually parameterized by matrices. When these models are applied to large-scale ML problems starting at millions of feature dimensions and tens of thousands of components, their parameter matrix can grow at an unexpected rate, resulting in high parameter synchronization costs that greatly slow down distributed learning. To address this issue, I proposed a Sufficient Factor Broadcasting (SFB) computation model [5] for efficient distributed learning of a large family of LVMs, which share the following property: the parameter update computed on each data sample is a rank-1 matrix, i.e. the outer product of two “sufficient factors” (SFs). By broadcasting the SFs among worker machines and reconstructing the update matrices locally at each worker, SFB improves communication efficiency — communication costs are linear in the parameter matrix’s dimensions, rather than quadratic — without affecting computational correctness. I presented a theoretical convergence analysis of SFB, and built a system with efficient implementation and an easy-to-use programming interface.

Privacy Preservation

As the problem scale increases, the cost of designing, implementing and maintaining machine learning models and algorithms rapidly grows, exceeding the capability of individuals and small-to-moderate institutions. This catalyzes the development of cloud ML, where ML models are developed and deployed on cloud servers, receive users’ data queries and return analytical results to users. One problem of cloud ML is invasion of users’ privacy since users’ plaintext data is fully exposed to cloud servers. To solve this problem, I leveraged homomorphic encryption to perform ML on encrypted data [11,10], which can strongly protect users’ privacy without sacrificing ML performance. Specifically, I studied neural network (a popular LVM) prediction on encrypted data [10]. I used homomorphic encryption in the following protocol: the data owner encrypts the data and sends the ciphertexts to the cloud to obtain a prediction from a trained model. The model operates on these ciphertexts and sends back the encrypted prediction. In this protocol, not only the data remains private, even the values predicted are available only to the data owner. Using homomorphic encryption and modifications to the activation functions and training algorithms of neural networks, I proposed crypto-nets and proved that they can be constructed and feasible. This method paves the way to build a secure cloud-based neural network prediction service without invading users’ privacy.
References


