Abstract

This thesis addresses two central tasks prevalent in many modern data processing, storage, and transmission pipelines: Clustering and compression. Specifically, in the first and the second part of this thesis, we study the problems of subspace clustering and random process clustering, respectively. While clustering problems are arguably among the most archetypal problems in unsupervised learning, compression methods are traditionally hand designed. In the third and fourth part of this thesis, we leverage machine learning techniques for compression, a trend that only emerged recently. In more detail, we propose a deep generative model-based framework for lossy data compression on one hand, and we study compression of neural network models for inference on resource-constrained hardware on the other hand.

Subspace clustering, the focus of the first part of this thesis, refers to the problem of clustering unlabeled high-dimensional data points into a union of unknown low-dimensional subspaces. Out of the plethora of subspace clustering algorithms, the sparse subspace clustering (SSC) algorithm (Elhamifar and Vidal, 2013) has attracted significant attention thanks to excellent clustering performance in practical applications. SSC performs spectral clustering based on an adjacency matrix obtained by sparsely representing each data point in terms of all the other data points via the Lasso. When the number of data points is large or the dimension of the ambient space is high, the computational complexity of SSC quickly becomes prohibitive. In this case, replacing the Lasso by the greedy orthogonal matching pursuit (OMP) algorithm results in significantly lower computational

complexity, while often yielding comparable performance (Dyer et al., 2013). The main contribution of the first part of the thesis is an analytical performance characterization of the resulting SSC-OMP algorithm for noisy data. Moreover, we introduce and analyze the SSC-matching pursuit (SSC-MP) algorithm, which employs MP in lieu of OMP. Both SSC-OMP and SSC-MP are proven to succeed even when the subspaces underlying the data intersect and when the data points are contaminated by severe noise. Our experiments show that SSC-MP compares very favorably to other sparsity-based subspace clustering algorithms, both in terms of clustering performance and running time. In addition, we find that, in contrast to SSC-OMP, the performance of SSC-MP is very robust with respect to the choice of parameters in the stopping criteria.

The second part of this thesis deals with the problem of clustering noisy finite-length observations of stationary ergodic random processes according to their generative models without prior knowledge of the model statistics and the number of generative models. Two algorithms, both using the L^1-distance between estimated power spectral densities (PSDs) as a measure of dissimilarity, are analyzed. The first one, termed nearest neighbor process clustering (NNPC), relies on partitioning the nearest neighbor graph of the observations via spectral clustering. The second algorithm consists of a single k-means iteration with farthest point initialization and was considered before in the literature, albeit with a different dissimilarity measure and with asymptotic performance results only. We prove that both algorithms succeed with high probability in the presence of noise and missing entries, and even when the generative process PSDs overlap significantly, all provided that the observation length is sufficiently large. Our results quantify the tradeoff between the overlap of the generative process PSDs, the observation length, the fraction of missing entries, and the noise variance. Furthermore, we provide extensive numerical results for synthetic and real data and find that NNPC outperforms state-of-the-art algorithms in human motion sequence clustering.

In the third part of this thesis, we propose and study the problem of distribution-preserving lossy compression. Motivated by recent

advances in extreme image compression which allow to maintain artifact-free reconstructions even at very low bitrates, we propose to optimize the rate-distortion tradeoff under the constraint that the reconstructed samples follow the distribution of the training data. Such a compression system recovers both ends of the spectrum: On one hand, at zero bitrate it learns a generative model of the data, and at high enough bitrates it achieves perfect reconstruction. Furthermore, for intermediate bitrates it smoothly interpolates between learning a generative model of the training data and perfectly reconstructing the training samples. We study several methods to approximately solve the proposed optimization problem, including a novel combination of Wasserstein generative adversarial networks and Wasserstein autoencoders, and present an extensive theoretical and empirical characterization of the proposed compression systems.

The fourth and last part of this thesis targets hardware-friendly compression of neural network models in the sense that the compressed models require only few multiplications at inference time. Specifically, we perform end-to-end learning of low-cost approximations of (generalized) matrix multiplications in deep neural network (DNN) layers by casting matrix multiplications as 2-layer sum-product networks (SPNs) (arithmetic circuits) and learning their (ternary) edge weights from data. The SPNs disentangle multiplication and addition operations and enable us to impose a budget on the number of multiplication operations. Combining our method with knowledge distillation techniques and applying it to image classification and language modeling DNNs, we obtain a first-of-a-kind reduction in number of multiplications (over 99.5%) while maintaining the predictive performance of the full-precision models. Finally, we demonstrate that the proposed framework is able to rediscover Strassen's matrix multiplication algorithm, learning to multiply 2×2 matrices using only 7 multiplications instead of 8.