Abstract

Machine learning (ML) applications have become ubiquitous during the last years, such as hyperspectral unmixing, recommender systems (for example in services like Spotify and Netflix), computer vision, text mining and audio source separation, to name a few. Many models and algorithms have been proposed, including linear dimensionality reduction (LDR), the perceptron, a precursor to modern neural networks, expectation-maximization, and the more recent algorithms for neural networks. Furthermore, there are two main categories of ML algorithms: supervised vs. unsupervised. Supervised algorithms have the labels of the training data available, while unsupervised algorithms train on data without labels.

Our focus in this thesis is on a subclass of ML models called matrix factorization models. These are unsupervised algorithms whose goal is to decompose a given data matrix into two smaller matrices, referred to as factors. Both factors are typically significantly smaller than the initial data matrix. There are multiple applications for matrix factorizations. Compression is one of them, since the factors are smaller than the original matrix. In text mining, matrix factorization retrieves topics from a collection of documents. In recommender systems, it creates groupings (clusters) of data points with common characteristics. As an example, when the input matrix represents users and products (for example songs or movies), these clusters represent groups with similar tastes. In addition, depending on the application, constraints can be imposed to the factors. An example of a constrained model is nonnegative matrix factorization (NMF) where the elements of the factors need to be nonnegative.

In this thesis, we focus on three models: (1) semi-binary Matrix Factorization (semi-bMF), where one factor has no constraints and the other can only have elements that are either 0 or 1, (2) Boolean matrix factorization (BMF) where both factors must only have elements that are either 0 or 1 and (3) Boolean matrix tri-factorization (BMTF), where we are factorizing into three factors. Furthermore, the matrix product used in the Boolean factorizations is different than the standard matrix product operation. This makes the computation harder but allows for better approximations and additional interpretability properties. For both models, we present new algorithms that are competitive with the state of the art and show that they can provide meaningful results in several applications, including topic modeling, image analysis, and clustering.