ICMLA Challenge 2018 - Parts-based decomposition of noisy data

Organizers: Shashanka Ubaru, Kristofer E. Bouchard, and Arif Wani.

Overview

With the ever growing collection of large volumes of scientific data, development of interpretable machine learning tools to analyze such data is becoming more important. However, robust, interpretable machine learning tools are lacking, threatening extraction of scientific insight and discovery. Dimensionality reduction and low rank approximations/decompositions are popular tools used in many applications to analyze high dimensional data. However, popular dimensionality reduction methods, often yield uninterpretable results, particularly for noisy data.

In this challenge, we focus on parts based feature extraction from noisy data using unsupervised learning. We desire to decompose given noisy data into a small set of interpretable (parts based) features. Such a decomposition will not require any training examples, making it a very important tool for exploratory data analysis, particularly in scientific data applications.

Challenge

The proposed challenge includes two main parts:

- *Feature extraction:* to extract parts based features from noisy versions of the data, and
- *Selection and estimation:* to compute the weights that recover the (underlying noiseless) data from these extract features.

In particular, we consider noisy versions of two datasets, namely, swimmer and 2-digit MNIST datasets. The swimmer data is constructed using sparse combinations of 16 parts (arms and legs positions). The 2-digit MNIST dataset is constructed by combining 2 out of 20 individual digit images (10 digits $0-9$ in units and tens place). The first part of the challenge is to extract the 16 and 20 bases/parts for the swimmer and the 2-digit MNIST datasets, respectively, from noisy versions of the data (with additive noise of unknown distribution). The second part of the challenge is to compute the exact 4 and 2 nonzero (sparse) weights (selection), respectively for recovering the original data from these extracted features.

Open-ended Applications: The authors can present *novel applications* for algorithms which yield parts-based decompositions of noisy data. For example, article [Ubaru et al., 2017] presents two applications, in Neuroscience and Analytic Chemistry, respectively for such algorithms. This part of the challenge includes description of the novel application and demonstration of the performance of the proposed algorithm in the application.

Performance evaluation

The performance of the proposed algorithm should be evaluated for the following four aspects:

- *Quality of the features extracted*: evaluated as correlation (or mean squared error) between the recovered bases and the actual bases.
- *Quality of selection and estimation:* evaluation based on exact sparsity, and error in recovery of (noiseless) data.
- *High-quality data reconstruction:* The reconstructed images are ideally denoised version of the data. The reconstruction quality is evaluated as the error in reconstruction the noiseless data.
- *Robustness to noise:* Algorithm should be to robust to various magnitudes (SNR) and types of noise (Poisson, Gaussian, etc.).

Datasets and codes

The swimmer and 2-digit MNIST datasets are available in Matlab (.MAT) format in the following address: `http://www-users.cs.umn.edu/~ubaru001/codes/ICMLA_18_Challenge.zip`. The actual parts (bases) from which the data are created are included for evaluation. A third dataset of random pixel images is also included for validation. The performance of the proposed algorithm can be compared against the performance of the $UoI\text{-}NMF_{cluster}$ algorithm proposed in [Ubaru et al., 2017]. The Matlab code for $UoI\text{-}NMF_{cluster}$ algorithm is also provided.

Submission

A short paper (4 pages) describing the proposed algorithms and results on the provided datasets should be submitted through the main conference submission website. These papers will be reviewed mainly based on:

- Originality and technical soundness of the employed algorithm.
- Performance of the algorithm with respect to the four aspects above.
- Proposal of new applications (if any).

Publication

Accepted papers will be scheduled for presentations at the conference and published in the ICMLA 2018 conference proceedings.

Important Dates

- Paper Submission Deadline: **September 14, 2018**.
- Notification of acceptance: **October 7, 2018**.
- Camera-ready papers & Pre-registration: **October 17, 2018**.
- The ICMLA Conference: **December 17-20, 2018**.

Contact: If you have any questions, contact the organizers:
Shashanka Ubaru: ubaru001@umn.edu or
Kristofer Bouchard: kristofer.bouchard@gmail.com.

References

[Ubaru et al., 2017] Ubaru, S., Wu, K., and Bouchard, K. E. (2017). $UoI\text{-}NMF_{cluster}$: A Robust Nonnegative Matrix Factorization Algorithm for Improved Parts-Based Decomposition and Reconstruction of Noisy Data. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 241–248.