Introduction to Machine Learning
Lab 5: Matrix Factorization and Linear Dimensionality Reduction

Hongteng Xu

April 3, 2022

1 Motivation

- Implement PCA and achieve data whitening based on it
- Try to explore the algorithm of robust PCA (RPCA) based on the guidance. Learn how to reformulate problems approximately, and do your second alternating optimization algorithm (Do you remember which is your first alternating optimization algorithm in this course?)
- Learn how to change data statistics by machine learning-guided data manipulation.

2 Tasks

Please read Lecture 6 carefully before doing this lab work.

1. Implement a PCA function, which takes the data matrix and the number of principal components you want as its input.
2. Achieve data whitening based on your PCA method.
3. Implement a robust PCA (RPCA) algorithm via alternating optimization. Given a data matrix $\boldsymbol{X} \in \mathbb{R}^{N\times D}$, the original RPCA problem is

$$\min_{\boldsymbol{X}} \|\boldsymbol{X}_{noise} - \boldsymbol{X}\|_1, \quad s.t.\ \text{rank}(\boldsymbol{X}) \leq L. \tag{1}$$

   Assuming the target $\boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S}$, where $\boldsymbol{L}$ is low-rank and $\boldsymbol{S}$ is sparse, we can solve (1) by optimizing the following two subproblems iteratively. At the k-th iteration:

$$\begin{aligned} &\text{P1:} \quad \boldsymbol{L}^{(k)} = \arg\min_{\boldsymbol{L}} \|\boldsymbol{X}_{noise} - \boldsymbol{L} - \boldsymbol{S}^{(k-1)}\|_F^2, \quad s.t.\ \text{rank}(\boldsymbol{L}) \leq L. \\ &\text{P2:} \quad \boldsymbol{S}^{(k)} = \arg\min_{\boldsymbol{S}} \|\boldsymbol{X}_{noise} - \boldsymbol{L}^{(k)} - \boldsymbol{S}\|_F^2, \quad s.t.\ \|\boldsymbol{S}\|_0 \leq \tau ND, \end{aligned} \tag{2}$$

   where $\|\boldsymbol{S}\|_0$ counts the number of nonzero elements in $\boldsymbol{S}$, and $\tau \in (0, 1)$ controls the ratio of nonzero elements.
4. **Data manipulation:** Given a set of zero-mean data, *i.e.*, $\boldsymbol{X} \in \mathbb{R}^{N\times D}$, whose covariance matrix is $\boldsymbol{\Gamma}_{\boldsymbol{X}} \in \mathbb{R}^{D\times D}$, we have a chance to add two outliers, i.e., $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ into $\boldsymbol{X}$ and formulate a new data matrix $\widehat{\boldsymbol{X}} \in \mathbb{R}^{(N+2)\times D}$. The two outliers satisfy that $\|\boldsymbol{x}_1\|_2 = \|\boldsymbol{x}_2\|_2 = 1$. Try to ensure $\widehat{\boldsymbol{X}}$ to be a zero-mean data matrix and maximize the difference between its covariance matrix $\boldsymbol{\Gamma}_{\widehat{\boldsymbol{X}}}$ and the original $\boldsymbol{\Gamma}_{\boldsymbol{X}}$, i.e., $\max \|\boldsymbol{\Gamma}_{\widehat{\boldsymbol{X}}} - \boldsymbol{\Gamma}_{\boldsymbol{X}}\|_F^2$