CSE 494 CSE/CBS 598 (Fall 2007): Numerical Linear Algebra for Data Exploration— Nonnegative Matrix Factorization

Instructor: Jieping Ye

1 Nonnegative Matrix Factorization

- An overview of SVD
  - Let A be an $m \times n$ matrix, with $m \geq n$. It can be factorized as

$$A = U \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^T,$$

    where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal, and $\Sigma \in \mathbb{R}^{m\times n}$ is diagonal

$$\Sigma = \text{diag}\left(\sigma_1, \sigma_2, \cdots, \sigma_n\right), \quad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0.$$

- Weak interpretability: The entries in U and V may be negative, even when A contains nonnegative entries only, such as term-document matrix.
- Nonnegative Matrix Factorization
  - Let A be an $m \times n$ matrix, with nonnegative entries. NMF computes two matrices $W \in \mathbb{R}^{m\times k}$ and $H \in \mathbb{R}^{k\times n}$, which solve the following optimization problem:

$$\min_{W\geq 0, H\geq 0} ||A - WH||_F.$$

  - Here, $W \geq 0$ indicates that all entries in W are nonnegative.
- (Lee and Seung, nature 1999) started a flurry of research into the Nonnegative Matrix Factorization. There are hundreds of papers in the area since 1999.
  - NMF is shown to induce parts-based representation.
- Many applications:
  - Image Processing and Computer Graphics
  - Text analysis
  - Bioinformatics

1.1 Alternating Least Squares (ALS) Algorithm

- This minimization problem above is nonlinear considered as an optimization problem for W and H at the same time. However, if one of the unknown matrices were known, W, say, then the problem of computing H, would be a standard, non-negatively constrained, least squares problem with a matrix right hand side. Therefore the most common way of solving it is to use an alternating least squares (ALS) procedure:
  - Guess an initial value $W^{(1)}$.

- for $k = 1, 2, \cdots$ until convergence
  (a) Solve $\min_{H \geq 0} ||A - W^{(k)} H||_F$, giving $H^{(k)}$.
  (b) Solve $\min_{W \geq 0} ||A - W H^{(k)}||_F$, giving $W^{(k+1)}$.

- The solution is not unique: if (W, H) is the solution, then $(WD, D^{-1}H)$ is also the solution for any diagonal matrix D with positive diagonal entries.

- How to solve $\min_{H \geq 0} ||A - W^{(k)} H||_F$?

  - Let a_j and h_j are the j-th columns of A and H, respectively.
  - Writing out the columns one by one, we see that the above matrix least squares problem is equivalent to n independent vector least squares problems:

$$\min_{h_j \geq 0} ||a_j - W^{(k)} h_j||_F, \quad j = 1, 2, \cdots, n.$$

  - The vector least squares problem can be solved by an active-set algorithm. MATLAB function: **lsqnonneg**.
  - The resulting algorithm for vector least squares problem is time-consuming.
  - As a cheaper alternative, one can take the unconstrained least squares solution, and then set all negative elements in H equal to zero.

1.2 A Multiplicative Update Algorithm

- Let

$$J = ||A - WH||_F.$$

- The objective function J can be re-written as:

$$\begin{aligned} J &= \text{trace}\left((A - WH)(A - WH)^T\right) \\ &= \text{trace}\left(AA^T - 2AH^TW^T + WHH^TW^T)\right) \\ &= \text{trace}(AA^T) - 2\text{trace}(AH^TW^T) + \text{trace}(WHH^TW^T). \end{aligned}$$

- Let $W = (w_{ij})$ and $h = (h_{ij})$. All w_{ij} and h_{ij} are constrained to be nonnegative. This leads to a constrained optimization problem.

- Let α_{ij} and β_{ij} be the Lagrange multiplier for constraints $w_{ij} \geq 0$ and $h_{ij} \geq 0$, respectively, and $\alpha = (\alpha_{ij})$, $\beta = (\beta_{ij})$, the Lagrange L is defined as follows:

$$L = J - \text{trace}(\alpha W^T) - \text{trace}(\beta H^T).$$

- The derivatives of L with respect to W and H are:

$$\begin{aligned} \frac{\partial L}{\partial W} &= -AH^T + WHH^T - \alpha, \\ \frac{\partial L}{\partial H} &= -W^TA + W^TWH - \beta. \end{aligned}$$

- Using the Kuhn-Tucker condition $\alpha_{ij} w_{ij} = 0$ and $\beta_{ij} h_{ij} = 0$, we get the following equations for w_{ij} and h_{ij}:

$$\begin{aligned}(AH^T)_{ij}\ w_{ij} - (WHH^T)_{ij}\ w_{ij} &= 0,\\ (W^TA)_{ij}\ h_{ij} - (W^TWH)_{ij}\ h_{ij} &= 0.\end{aligned}$$

- These equations lead to the following updating formulas:

$$\begin{aligned}w_{ij} &\leftarrow w_{ij}\frac{(AH^T)_{ij}}{(WHH^T)_{ij}+\epsilon}\\ h_{ij} &\leftarrow h_{ij}\frac{(W^TA)_{ij}}{(W^TWH)_{ij}+\epsilon}.\end{aligned}$$

- It is proven by Lee and Seung that the objective function J is non-increasing under the above iterative updating rules, and that the convergence of the iteration (to a local minimum) is guaranteed.

- Software package:
  - Cichocki and Zdunek 2006. NMFLAB for Signal Processing, available at http://www.bsp.brain.riken.jp/ICALAB/nmflab.html.

- Survey articles:
  - Algorithms and applications for approximation nonnegative matrix factorization. M. Berry, M. Browne, A. Langville, P. Pauca, and R. J. Plemmons. Computational Statistics and Data Analysis, 2006.
  - Nonnegative Matrix Approximation: Algorithms and Applications S. Sra and I. S. Dhillon UTCS Technical Report TR-06-27, June 2006.

1.3 Initialization Issue

- One problem with several of the algorithms for non-negative matrix factorization is that convergence to a global minimum is not guaranteed. It often happens that convergence is slow and that a sub-optimal approximation is reached.

- An efficient procedure for computing a good initial approximation based on the SVD of A.

1.4 NMF for Clustering

- A non-negative factorization $A = WH$ can be used for clustering: the data vector a_j is assigned to cluster i if h_{ij} is the largest element in column j of H.

- Related articles on document clustering based on NMF:
  - Xu, W., Liu, X., Gong, Y. Document-Clustering based on Non-Negative Matrix Factorization. In: Proceedings of SIGIR 2003.
  - F. Shahnaz and *et al.* Document clustering using nonnegative matrix factorization. Information Processing and Management: an International Journal, 2006.