CS 596: Theoretical Machine Learning

Lecturer: Pranjal Awasthi
Scribe: Orsan Aytekin

Lecture 15
Nov.10, 2016

In many applications we are not just interested in an arbitrary factorization of the data matrix, but one with special properties. One such instance is Non-negative Matrix Factorization (NMF). Here we wan to factorize an nxm matrix M with the property that it has no negative elements.

$$\begin{aligned} M &= U\Sigma V^T \\ &= \begin{bmatrix} \bar{u_1} & \bar{u_2} & \dots & \bar{u_r} \end{bmatrix} \begin{bmatrix} \bar{\sigma_1} & & \\ & \ddots & \\ & & \bar{\sigma_r} \end{bmatrix} \begin{bmatrix} \bar{v_1}^T \\ \vdots \\ \bar{v_r}^T \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1\bar{u_1} & \sigma_2\bar{u_2} & \dots & \sigma_r\bar{u_r} \end{bmatrix} \begin{bmatrix} \bar{v_1}^T \\ \vdots \\ \bar{v_r}^T \end{bmatrix} \end{aligned}$$

Here, U is $n \times r$ orthogonal matrix, Σ is $r \times r$ diagonal rectangular matrix, V is $m \times r$ matrix. We want both U and V to be entry wise non-negative. We will often write this in the form $M = UV^T$, where the Σ matrix is combined with U and V is an orthonormal matrix.

Here are some definitions: $r(M)$ = rank of M, $r_+(M)$ = non-negative rank of M. This is the minimum r for which non-negative factorization exists. We have the following properties:

- $min(n, m) \geq r_+(M) \geq r(M)$.
- As opposed to SVD, non-negative factorization is NP-hard.
- runtime of the best known algorithm is $(nm)^{O(r_+^2)}$

From now on, we will use r to denote the non-negative rank of the matrix. Below we will see a polynomial time algorithm to solve non worst case instances of NMF. In order to formalize such instances we need the following definition

Definition 1 (Separability)**.** *M is separable if $\forall j \in [r]$ $\exists i \in [m]$ such that*

$$V_{ij} = \begin{cases} \neq 0, & \text{if } i = j \\ 0, & i \neq j \end{cases}$$

In other words, separability says that for every column of V there exists a "special" row associated with it. This is true in many practical settings. For example in the Netflix example we have M = [user-by-movie] = [user-by-genre] x [genre-by-movies]. Separability says that for every genre, there exists a movie that belongs to only that genre and no one else.

Theorem 1. *If M is separable, then one can find $M = UV^T$ such that $U \geq 0$, $V \geq 0$ in polynomial time ($poly(n, m, r)$).*

We will first make certain simplifications that are without loss of generality.

Lemma 1. *Let M be a separable matrix. Scale all the columns of M to have ℓ_1 norm 1. Rename M to be the new matrix. Then there exists a factorization of $M = UV^T$ where the column ℓ_1 norms of U and V are 1.*

Proof. To scale the columns of M multiply M by a diagonal matrix D. We get $MD = UV^TD = UW^T$. Now scale the columns of U by using another diagonal matrix D_1 such that $M = UD_1D_1^{-1}W^T = U'V'^T$. Now it is easy to see that since all matrices are non-negative, ℓ_1 norm 1 on M and U' also implies that the columns of V' are scaled to ℓ_1 norm 1. □

Lemma 2. *Let M be a separable matrix that has a non negative factorization with $M = UV^T$ and ℓ_1 column norms of M, U, V be 1. Then there exists a similar factorization where no column of U can be written as a convex combination of the other columns.*

Proof. Let U_i be a column that can be written as $U_i = \sum_{j \neq i} \alpha_j U_j$ and $\sum_j \alpha_j = 1$. Delete column i from U and row i from V^T. Update remaining entries of V as follows: $\forall k \in [m], l \neq i$, $V_{k,j} \leftarrow V_{k,j} + \alpha_j V_{k,i}$. It is easy to verify that the remaining new factors will also have column ℓ_1 norm 1 and matrix V will be separable. □

Proof of Theorem 1. Once we have a matrix that satisfies the above two properties then it is easy to see that every column of U will appear in M as a column. Furthermore, no such column can be written as a convex combination of other columns. Hence, one can identify all columns of M that cannot be written as convex combinations of other columns to get the matrix U. Once we have U, V can be recovered by solving a linear system.

1 Additional Readings

- The original NMF paper. `http://www.columbia.edu/~jwp2128/Teaching/W4721/papers/nmf_nature.pdf`.
- The paper of Donoho and Stodden and that introduced separability. `https://web.stanford.edu/~vcs/papers/NMFCDP.pdf`.
- The paper of Arora et al. that provides the algorithm for separable instances. `https://arxiv.org/abs/1111.0952`.

□