Nonnegative Matrix Factorization (NNMF)

Note Title

★ What is NNMF?

- a type of low-rank approximation of a given matrix $A \in \mathbb{R}^{m \times n}$ where $a_{ij} \geq 0 \quad \forall i, \forall j$.
- Factors must be nonnegative.
- Certain applications (e.g., text mining, chemometrics, etc.) require nonnegativity in all the factors involved.
- SVD, PCA cannot be used because they involve negative coefficients, negative entries in the factors (i.e., entries of U, V etc.)

★ NNMF Objective

Given a nonnegative matrix $A \in \mathbb{R}^{m \times n}$ and $k < \min(m, n)$, find nonnegative matrices $W \in \mathbb{R}^{m \times k}$, $H \in \mathbb{R}^{k \times n}$ to minimize the objective function

$$J_{NNMF}(W, H) := \frac{1}{2} \| A - WH \|_F^2$$

The product WH is called an (approximate) NNMF of A.

The choice of k is critical in practice, but often $k \ll \min(m, n)$
$\Rightarrow$ a compressed approx. of A.

☆ Numerical Approaches for NNMF

- Minimization of J_{NNMF} is difficult:
  - Many local minima exist in J_{NNMF} in both W & H.
  - Lack of a unique solution
    Consider $D \in \mathbb{R}^{k \times k}$, nonnegative and nonsingular, and suppose D^{-1} is also nonnegative (e.g., D could be $\text{diag}(d_1, \cdots, d_k)$ with $d_j > 0 \quad 1 \le j \le k$.)
    Then if WH is an NNMF of A, so is $WDD^{-1}H$.

- Many algorithms have been proposed. We'll discuss only one of them based on the so-called Alternating Least Squares (ALS).

Algorithm (ALS-NNMF)

rand(m,k) returns m×k random matrix whose entries are uniformly distributed on the unit interval (0,1).

- Initialize W by $W = \text{rand}(m, k)$.
- For $j = 1 :$ maxiter
  - Solve for H in $W^T W H = W^T A$.
  - Set all negative entries of H to 0.
  - Solve for W in $H H^T W^T = H A^T$.
  - Set all negative entries of W to 0.

Compare it with "randn": the standard normal distribution!

Notes: (1) Convergence is not guaranteed yet this algorithm usually works in practice.

(2) $W^TWH = W^TA$, $HH^TW^T = HA^T$ are just a bunch of normal eqn's, e.g., $W^TW h_i = W^T a_i$, $i = 1:n$.

$HH^TW^T = HA^T$ comes from the following:

$$\|A - WH\|_F^2 \to \min.$$

$$\Leftrightarrow \|A^T - H^TW^T\|_F^2 \to \min.$$

$$\Leftrightarrow ((H^T)^T H^T) W^T = (H^T)^T A^T$$

$$\Leftrightarrow HH^TW^T = HA^T.$$

(3) Random initialization like the original algorithm may not be efficient. We can use the following algorithm to initialize the matrix W:

- Compute the first k singular values and the corresponding vectors by
  [U,S,V] = svds(A,k);
- Then do the following:

```
W(:,1) = U(:,1);
for j=2:k
    C = U(:,j)*V(:,j)';
    C = C.*(C>=0);
    [u,s,v] = svds(C,1);
    W(:,j) = u;
end
```

The reasoning behind this initialization is the following:

Good exercise!

If A is nonnegative, then its first singular vectors u_1 & v_1 are also nonnegative. So, it's good to use $W(:,1) = u_1$ and

$H(1,:) = v_1^T$

Unfortunately, u_2, v_2 contains negative entries due to the orthogonality $u_1 \perp u_2$, $v_1 \perp v_2$.

So, construct $C = u_2 v_2^T$, and set all the negative entries of C to 0. Then this C is nonnegative, so can compute the first singular vectors of this C, which are nonnegative and good approximations to u_2, v_2. Then set the first left singular vector as the 2nd column of W. We can repeat this procedure until we fill W.

Example: Problem 2 of HW #1.

Consider the following set of terms (words) and documents (or rather book titles):

	Terms		Documents
T1:	Book (Handbook, BOOK)	D1:	The Princeton Companion to Mathematics
T2:	Equation (Equations)	D2:	NIST Handbook of Mathematical Functions
T3:	Function (Functions)	D3:	Table of Integrals, Series, and Products
T4:	Integral (Integrals)	D4:	Linear Integral Equations
T5:	Linear	D5:	Proofs from THE BOOK
T6:	Mathematics (Mathematical)	D6:	The Book of Numbers
T7:	Number (Numbers)	D7:	Number Theory in Science and Communication
T8:	Series	D8:	Green's Functions and Boundary Value Problems
		D9:	Discourse on Fourier Series
		D10:	Basic Linear Partial Differential Equations
		D11:	Mathematical Physics, An Advanced Course

Term-Document Matrix

$A=$

	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}
T_1	0	1	0	0	1	1	0	0	0	0	0
T_2	0	0	0	1	0	0	0	0	0	1	0
T_3	0	1	0	0	0	0	0	1	0	0	0
T_4	0	0	1	1	0	0	0	0	0	0	0
T_5	0	0	0	1	0	0	0	0	0	1	0
T_6	1	1	0	0	0	0	0	0	0	0	1
T_7	0	0	0	0	0	1	1	0	0	0	0
T_8	0	0	1	0	0	0	0	0	1	0	0

Let's compute the NNMF of A with $k=3$, using MATLAB:

```
>> [W, H] = nnmf(A,3);
```

The resulting matrices are :

$$W = \begin{bmatrix} 1.4366 & 0.0016 & 0 \\ 0 & 1.4181 & 0 \\ 0.9536 & 0 & 0 \\ 0 & 0.6530 & 0.8984 \\ 0 & 1.4181 & 0 \\ 1.2931 & 0 & 0.0023 \\ 0.4829 & 0.0076 & 0 \\ 0 & 0 & 1.3883 \end{bmatrix} \begin{matrix} T1 \\ T2 \\ T3 \\ T4 \\ T5 \\ T6 \\ T7 \\ T8 \end{matrix}$$

w_3

$$\begin{matrix} & D_1 & D_2 & D_3 & D_4 & D_5 & D_6 & D_7 & D_8 & D_9 & D_{10} & D_{11} \end{matrix}$$

$$H = \begin{bmatrix} 0.2681 & 0.7569 & 0 & 0 & 0.2922 & 0.3875 & 0.0954 & 0.1967 & 0 & 0 & 0.2681 \\ 0 & 0 & 0.0370 & 0.7573 & 0.0001 & 0.0042 & 0.0041 & 0 & 0 & 0.6520 & 0 \\ 0.0014 & 0.0004 & 0.8342 & 0.1669 & 0 & 0 & 0 & 0 & 0.5257 & 0 & 0.0014 \end{bmatrix}$$

Let's interpret the results !

w_3 has large entries corresponding to T4 (Integral / Integrals) and T8 (Series).

The responses of the documents to w_3 is the 3rd row of H.

You can see that D_3 and D_9 have high responses, which are reasonable :

D_3 = Table of Integrals, Series, and Products

D_9 = Discourse on Fourier Series

Exercise : Do interpret w_1 and w_2 yourself !!