D5: Databases and Information Systems
Data Mining and Matrices, SS 2017
Homework #4: NMF
Tutorial: **21 June 2017** at 10:15

*You can discuss these problems with other students, but everybody must do and present their own answers. You can use computers etc. to perform the algebraic operations, but you must show the intermediate steps (and "computer said so" is never a valid answer). You are of course free to use material from the Internet, but again, you must present the intermediate steps and you must also be able to explain why the steps are valid and why you chose them. You can mark an answer even if it is not complete or correct, as long as you have made significant progress towards solving it. Note, however, that the TA does the final decision on whether your solution is complete (or correct) enough for a mark.* ***Starting from this problem sheet, we apply more strict evaluation on what constitutes as sufficiently solved problem. You should only mark a problem if you think you have essentially solved it. This is done to give you a better impression on how the exam questions will be graded.***

**Problem 1** (Regularized ALS). In the lecture, we saw regularized ALS algorithms for NMF using both the Frobenius and $L_1$ regularizers. The general form of regularized NMF for $\boldsymbol{A} \in \mathbb{R}_+^{I\times J}$, $\boldsymbol{W} \in \mathbb{R}_+^{I\times K}$, and $\boldsymbol{H} \in \mathbb{R}_+^{K\times J}$ is

$$\begin{aligned} &\text{minimize } \|\boldsymbol{A} - \boldsymbol{W}\boldsymbol{H}\|_F + \alpha R_{\boldsymbol{W}}(\boldsymbol{W}) + \beta R_{\boldsymbol{H}}(\boldsymbol{H}) \\ &\text{subject to } w_{ik} \geq 0, \quad h_{kj} \geq 0, \quad \text{for all } i, k, j\ . \end{aligned} \tag{1.1}$$

The reqularized ALS update rule for $\boldsymbol{W}$ in (1.1) is

$$\boldsymbol{W} \leftarrow [(\boldsymbol{A}\boldsymbol{H}^T - \alpha\boldsymbol{\Phi}_{\boldsymbol{W}})(\boldsymbol{H}\boldsymbol{H}^T)^{-1}]_+\ , \tag{1.2}$$

where

$$\boldsymbol{\Phi}_{\boldsymbol{W}} = \left(\frac{\partial R_{\boldsymbol{W}}(\boldsymbol{W})}{\partial w_{ik}}\right)_{ik} \in \mathbb{R}^{I\times K}$$

is the matrix of partial derivatives of $R_{\boldsymbol{W}}$. Derive the update rules for $\boldsymbol{W}$ presented in the lecture for

a) $R_{\boldsymbol{W}}(\boldsymbol{W}) = \|\boldsymbol{W}\|_F^2$ (*Hint:* The general update rule (1.2) is obtained from the stationary point $\boldsymbol{W} = (\boldsymbol{A}\boldsymbol{H}^T - \alpha\boldsymbol{\Phi}_{\boldsymbol{W}})(\boldsymbol{H}\boldsymbol{H}^T)^{-1}$ (i.e. the point where the gradient is zero). Use this equation to derive the update rule.)

b) $R_{\boldsymbol{W}}(\boldsymbol{W}) = \sum_{i,k} w_{ik}$ (assuming the columns of $\boldsymbol{W}$ are normalized)

**Problem 2** (Hoyer's sparsity function). Recall that Hoyer (2004) defines the function sparsity : $\mathbb{R}^n \to \mathbb{R}$ as

$$\text{sparsity}(\boldsymbol{x}) = \frac{\sqrt{n} - \|\boldsymbol{x}\|_1 / \|\boldsymbol{x}\|_2}{\sqrt{n} - 1}\ , \tag{2.1}$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ are the vector $\ell_1$ and $\ell_2$ norms, respectively. Show that

a) $\text{sparsity}(\boldsymbol{x}) = 1$ if and only if $\boldsymbol{x}$ has exactly one non-zero element; and

b) $\text{sparsity}(\boldsymbol{x}) = 0$ if and only if $|x_i| = |x_j|$ for all $i, j \in \{1, 2, \ldots, n\}$.

**Problem 3** (Multiplicative rules as gradient descent). Lee and Seung's multiplicative updates for NMF would update $\boldsymbol{H}$ as

$$\boldsymbol{H}_{ij} \leftarrow \boldsymbol{H}_{ij}\frac{(\boldsymbol{W}^T\boldsymbol{A})_{ij}}{(\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H})_{ij}}\ . \tag{3.1}$$

Show that this can be considered as a gradient descent approach with gradient updates

$$\boldsymbol{H}_{ij} \leftarrow \boldsymbol{H}_{ij} + \varepsilon_{ij}\Big((\boldsymbol{W}^T\boldsymbol{A})_{ij} - (\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H})_{ij}\Big)\ , \tag{3.2}$$

where the step size $\varepsilon_{ij}$ is set separately for every element.

*Hint:* Find $\varepsilon_{ij}$ such that you can transform (3.2) to (3.1).

**Problem 4** (ALS as Newton's method). Recall that the Newton's method for optimizing a function $f : \mathbb{R}^n \to \mathbb{R}$ involves the following iterative update rule:

$$\boldsymbol{x}_{n+1} \leftarrow \boldsymbol{x} - [\boldsymbol{H}(f(\boldsymbol{x}_n))]^{-1} \nabla f(\boldsymbol{x}_n) \ . \tag{4.1}$$

The alternating least squares optimization for NMF, on the other hand, updates matrix $\boldsymbol{H}$ in decomposition $\boldsymbol{A} \approx \boldsymbol{W}\boldsymbol{H}$ as

$$\boldsymbol{H} \leftarrow [\boldsymbol{W}^+ \boldsymbol{A}]_+ \ . \tag{4.2}$$

Let us assume that $\boldsymbol{W}^T\boldsymbol{W}$ is invertible. Compute first Newton's update rule for matrix $\boldsymbol{H}$ in NMF (truncating negative values to 0), and then show that this update rule is equivalent to the ALS update rule.

**Problem 5** (NMF as $k$-means). The $k$-means algorithm tries to optimize the function

$$\sum_{j=1}^{k} \sum_{i \in C_j} \left\| \boldsymbol{a}_i - \boldsymbol{\mu}_j \right\|_2^2 \ , \tag{5.1}$$

where $\boldsymbol{a}_i \in \mathbb{R}^d, i = 1, \ldots, n$ are the input (row) vectors, $C_j \subset \{1, 2, \ldots, n\}, j = 1, \ldots, k, C_i \cap C_j = \emptyset$ if $i \neq j$, and $\cup_j C_j = \{1, 2, \ldots, n\}$ define the $k$ clusters of $\boldsymbol{a}_i$, and $\boldsymbol{\mu}_j \in \mathbb{R}^d, j = 1, \ldots, k$, are the centroids for the clusters. Given a clustering, the centroid $\boldsymbol{\mu}_j$ is computed as the element-wise average, $\boldsymbol{\mu}_j = \frac{1}{|C_j|} \sum_{i \in C_j} \boldsymbol{a}_i$ (summation and division are element-wise).

Show that if all $\boldsymbol{a}_i$ are non-negative, we can write (5.1) as a special type of semi-orthogonal NMF

$$\|\boldsymbol{A} - \boldsymbol{G}\boldsymbol{M}\|_F^2 \ , \quad \boldsymbol{G}^T\boldsymbol{G} = \boldsymbol{I} \ . \tag{5.2}$$

That is, show how to transform (5.1) into (5.2) and verify that all matrices stay non-negative and that $\boldsymbol{G}$ is column-orthogonal.

**Problem 6** (NMF and pLSA). In the lectures the pLSA was presented as NMF optimizing the generalized KL divergence. In this problem we aim at proving why GKL is used instead of the Frobenius norm.

Recall that in pLSA, the joint probability of a document $d_i$ and term $t_j$ using $K$ topics $(z_k)_{k=1}^K$ is defined as

$$\Pr[d_i, t_j] = \sum_{k=1}^{K} \Pr[z_k] \Pr[d_i \mid z_k] \Pr[t_j \mid z_k] \ . \tag{6.1}$$

Using the NMF formulation with document-term matrix $\boldsymbol{A}$ that is normalized to sum to unity and NMF factor matrices $\boldsymbol{W}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{H}$, where columns of $\boldsymbol{W}$, diagonal of $\boldsymbol{\Sigma}$, and rows of $\boldsymbol{H}$ sum to unity, we can write (6.1) as

$$\Pr[d_i, w_j] = \sum_{k=1}^{K} \sigma_{kk} w_{ik} h_{kj} = (\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{H})_{ij} \ . \tag{6.2}$$

Now, the likelihood of observing $\boldsymbol{A}$ when drawing the data from the distribution (6.2) is proportional to

$$L = L(\boldsymbol{A} \mid \boldsymbol{W}, \boldsymbol{\Sigma}, \boldsymbol{H}) = \prod_i \prod_j \Pr[d_i, w_j]^{\boldsymbol{A}_{ij}} \ . \tag{6.3}$$

Show that NMF with GKL divergence as the error metric is maximizing the likelihood $L$ by showing that maximizing the log-likelihood $\log L(\boldsymbol{A} \mid \boldsymbol{W}, \boldsymbol{\Sigma}, \boldsymbol{H})$ is equivalent to minimizing the (generalized) KL divergence

$$D_{GKL}(\boldsymbol{A} \| \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{H}) = \sum_i \sum_j \left( \boldsymbol{A}_{ij} \ln \frac{\boldsymbol{A}_{ij}}{(\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{H})_{ij}} - \boldsymbol{A}_{ij} + (\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{H})_{ij} \right) \ . \tag{6.4}$$