Exercises
Computational Intelligence Lab
SS 2020

Machine Learning Institute
Dept. of Computer Science, ETH Zürich
Prof. Dr. Thomas Hofmann
Web http://da.inf.ethz.ch/cil

Series 5 Solutions (Non-Negative Matrix Factorization)

Solution 1 (Convex Relaxation for Exact Matrix Recovery):

Let us consider the singular vector decomposition of matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$,

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top, \tag{1}$$

where matrices $\mathbf{U} \in \mathbb{R}^{m \times m}$ and $\mathbf{V} \in \mathbb{R}^{n \times n}$ are orthogonal, and $\mathbf{D} \in \mathbb{R}^{m \times n}$ is a diagonal rectangular matrix with non-negative real numbers on its diagonal, which, for instance, for the case $m < n$ can be represented as follows:

$$\mathbf{D} = \begin{bmatrix}\mathbf{\Sigma} & \mathbf{0}\end{bmatrix}, \quad \mathbf{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_m) = \begin{bmatrix}\sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_m\end{bmatrix}, \quad \sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_m \geq 0. \tag{2}$$

1. Since matrices $\mathbf{U}$ and $\mathbf{V}$ are orthogonal, and hence full rank matrices, the rank of the matrix $\mathbf{A}$ is equal to the number of its positive singular values

$$\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{D}) = \#\{\sigma_i > 0\}. \tag{3}$$

On the other hand, the Euclidean operator norm[1] of $\mathbf{A}$ is equal to its largest singular value σ_1,

$$\|\mathbf{A}\|_2 = \sigma_{\max}(\mathbf{A}) = \sigma_1. \tag{4}$$

Therefore, if $\|\mathbf{A}\|_2 \leq 1$ and hence $\forall i \ \sigma_i \leq 1$, one can derive the following inequality,

$$\text{rank}(\mathbf{A}) = \#\{\sigma_i > 0\} = \sum_{i:\ \sigma_i > 0} 1 \geq \sum_{i:\ \sigma_i > 0} \sigma_i = \sum_i \sigma_i = \|\mathbf{A}\|_*. \tag{5}$$

2. A function $f: \ X \to \mathbb{R}$ is convex if $\forall x, y \in X$

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \quad \forall \lambda \in [0,1]. \tag{6}$$

Let $\mathbf{U}_\lambda \mathbf{D}_\lambda \mathbf{V}_\lambda^\top$ be the SVD decomposition of $\lambda\mathbf{A} + (1-\lambda)\mathbf{B}$. Then, we have

$$\begin{aligned}\|\lambda\mathbf{A} + (1-\lambda)\mathbf{B}\|_* &= \text{trace}\left(\mathbf{D}_\lambda\right) &(7)\\ &= \text{trace}\left(\mathbf{U}_\lambda^\top(\mathbf{U}_\lambda\mathbf{D}_\lambda\mathbf{V}_\lambda^\top)\mathbf{V}_\lambda\right) &(8)\\ &= \text{trace}\left(\mathbf{U}_\lambda^\top(\lambda\mathbf{A} + (1-\lambda)\mathbf{B})\mathbf{V}_\lambda\right) &(9)\\ &= \lambda\,\text{trace}\left(\mathbf{U}_\lambda^\top\mathbf{A}\mathbf{V}_\lambda\right) + (1-\lambda)\,\text{trace}\left(\mathbf{U}_\lambda^\top\mathbf{B}\mathbf{V}_\lambda\right). &(10)\end{aligned}$$

Our proof is done once we bound both terms: trace $\left(\mathbf{U}_\lambda^\top\mathbf{A}\mathbf{V}_\lambda\right) \leq \|\mathbf{A}\|_*$ and trace $\left(\mathbf{U}_\lambda^\top\mathbf{B}\mathbf{V}_\lambda\right) \leq \|\mathbf{B}\|_*$. Let

[1] https://en.wikipedia.org/wiki/Operator_norm

$\mathbf{U}_A\mathbf{D}_A\mathbf{V}_A^\top$ be the SVD decomposition of $\mathbf{A}$. Then, we get

$$\begin{aligned}
\text{trace}\left(\mathbf{U}_\lambda^\top \mathbf{A}\mathbf{V}_\lambda\right) &= \sum_{i=1}^{\min(m,n)} \left[\mathbf{U}_\lambda^\top \mathbf{A}\mathbf{V}_\lambda\right]_i^i && (11)\\
&= \sum_{i=1}^{\min(m,n)} \left[\mathbf{U}_\lambda^\top \mathbf{U}_A\mathbf{D}_A\mathbf{V}_A^\top\mathbf{V}_\lambda\right]_i^i && (12)\\
&= \sum_{i=1}^{\min(m,n)}\sum_{j=1}^{\min(m,n)} \left[\mathbf{U}_\lambda^\top \mathbf{U}_A\right]_j^i \sigma_j(\mathbf{A}) \left[\mathbf{V}_A^\top\mathbf{V}_\lambda\right]_i^j && (13)\\
&= \sum_{j=1}^{\min(m,n)} \sigma_j(\mathbf{A}) \sum_{i=1}^{\min(m,n)} \left[\mathbf{U}_\lambda^\top \mathbf{U}_A\right]_j^i \left[\mathbf{V}_A^\top\mathbf{V}_\lambda\right]_i^j && (14)\\
&\leq \sum_{j=1}^{\min(m,n)} \sigma_j(\mathbf{A}) \left\|\left[\mathbf{U}_\lambda^\top \mathbf{U}_A\right]_j\right\|_2 \left\|\left[\mathbf{V}_A^\top\mathbf{V}_\lambda\right]^j\right\|_2 && (15)\\
&= \sum_{j=1}^{\min(m,n)} \sigma_j(\mathbf{A}) && (16)\\
&= \|\mathbf{A}\|_*, && (17)
\end{aligned}$$

where the superscript i above a matrix denotes its i-th row and the subscript i below a matrix denotes its i-th column. Similarly, one can bound $\text{trace}\left(\mathbf{U}_\lambda^\top \mathbf{B}\mathbf{V}_\lambda\right) \leq \|\mathbf{B}\|_*$, and therefore,

$$\|\lambda\mathbf{A} + (1-\lambda)\mathbf{B}\|_* \leq \lambda\|\mathbf{A}\|_* + (1-\lambda)\|\mathbf{B}\|_*, \tag{18}$$

which concludes the proof.

3. We are going to rewrite the problem[2]

$$\min_{\mathbf{B}} \|\mathbf{B}\|_*, \quad \text{subject to } \|\mathbf{A}-\mathbf{B}\|_{\mathbf{G}} = 0, \tag{19}$$

as a problem of semidefinite programming (SDP) in the following form,

$$\begin{aligned}
&\min_{\mathbf{B},\mathbf{W_1},\mathbf{W_2}} \frac{1}{2}\operatorname{Tr}(\mathbf{W_1}) + \frac{1}{2}\operatorname{Tr}(\mathbf{W_2}) && (20)\\
&\text{subject to } \underbrace{\begin{bmatrix}\mathbf{W_1} & \mathbf{B}\\ \mathbf{B}^\top & \mathbf{W_2}\end{bmatrix} \succeq 0}_{\text{cone constraints}} \text{ and } \underbrace{\|\mathbf{A}-\mathbf{B}\|_{\mathbf{G}} = 0}_{\text{affine constraints}}.
\end{aligned}$$

In what follows, we assume $m = n$ for simplicity. We are going to prove the equivalence of (19) and (20) with the help of the Schur complement lemma:

$$\begin{bmatrix}\mathbf{W}_1 & \mathbf{B}\\ \mathbf{B}^\top & \mathbf{W}_2\end{bmatrix} \succeq 0 \iff \begin{cases}\mathbf{W}_1 \succeq 0\\ \mathbf{W}_2 - \mathbf{B}^\top\mathbf{W}_1^+\mathbf{B} \succeq 0\\ (\mathbf{I} - \mathbf{W}_1\mathbf{W}_1^+)\mathbf{B} = 0\end{cases}, \tag{21}$$

where $\mathbf{A}^+$ denotes the pseudoinverse of a matrix $\mathbf{A}$, which is a generalization of the inverse matrix defined for any rectangular matrix.[3] The pseudoinverse of a matrix is tightly connected to its SVD decomposition. If $\mathbf{U}\mathbf{D}\mathbf{V}^\top$ is the SVD decomposition of matrix $\mathbf{A}$, then the pseudoinverse is equal to $\mathbf{A}^+ = \mathbf{V}\mathbf{D}^+\mathbf{U}^\top$.
Using the Schur complement lemma, the SDP problem (20) can be reformulated as follows:

$$\begin{aligned}
&\min_{\mathbf{B},\mathbf{W_1},\mathbf{W_2}} \frac{1}{2}\operatorname{Tr}(\mathbf{W_1}) + \frac{1}{2}\operatorname{Tr}(\mathbf{W_2}) && (22)\\
&\text{subject to } \|\mathbf{A}-\mathbf{B}\|_{\mathbf{G}} = 0,\\
&\qquad \mathbf{W}_1 \succeq 0,\\
&\qquad \mathbf{W}_2 - \mathbf{B}^\top\mathbf{W}_1^+\mathbf{B} \succeq 0,\\
&\qquad (\mathbf{I} - \mathbf{W}_1\mathbf{W}_1^+)\mathbf{B} = 0.
\end{aligned}$$

[2] This one is a bonus question, similar questions will not be asked in the exam.
[3] https://en.wikipedia.org/wiki/Moore%E2%80%93Penrose_inverse

Since matrix $\mathbf{W}_1$ is symmetric positive semidefinite ($\mathbf{W}_1 \succeq 0$), its SVD decomposition can be parametrized by an orthogonal matrix $\mathbf{U}$ and a diagonal positive semidefinite matrix $\mathbf{D} \succeq 0$:

$$\mathbf{W}_1 = \mathbf{U}\mathbf{D}\mathbf{U}^\top \succeq 0. \tag{23}$$

Therefore, the pseudoinverse of $\mathbf{W}_1$ is equal to

$$\mathbf{W}_1^+ = \mathbf{U}\mathbf{D}^+\mathbf{U}^\top \succeq 0. \tag{24}$$

Note that replacing the constraint $\mathbf{W}_2 - \mathbf{B}^\top\mathbf{W}_1^+\mathbf{B} \succeq 0$ with the equation $\mathbf{W}_2 = \mathbf{B}^\top\mathbf{W}_1^+\mathbf{B}$ does not affect the solution. To see this, recall that the trace of a symmetric matrix equals the sum of its eigenvalues, so $\mathbf{W}_2$ with $\mathbf{W}_2 \succ \mathbf{B}^\top\mathbf{W}_1^+\mathbf{B}$ cannot be a solution because its eigenvalues can be further decreased and, thus, make the objective (22) smaller. With this, we can prove that the problem (19) is equivalent to

$$\begin{aligned}
\min_{\mathbf{B},\mathbf{U},\mathbf{D}} \quad & \frac{1}{2}\operatorname{Tr}(\mathbf{U}\mathbf{D}\mathbf{U}^\top) + \frac{1}{2}\operatorname{Tr}(\mathbf{B}^\top(\mathbf{U}\mathbf{D}^+\mathbf{U}^\top)\mathbf{B}) =: \mathcal{L} \\
\text{subject to } & \|\mathbf{A}-\mathbf{B}\|_{\mathbf{G}} = 0, \\
& \mathbf{D} = \operatorname{diag}(d_1,\dots,d_n) \succeq 0, \\
& \mathbf{U} \text{ orthogonal}, \\
& (\mathbf{I} - \mathbf{U}\mathbf{D}\mathbf{D}^+\mathbf{U}^\top)\mathbf{B} = 0.
\end{aligned} \tag{25}$$

Expanding (25), we get

$$\begin{aligned}
\mathcal{L} &= \frac{1}{2}\operatorname{Tr}(\mathbf{U}\mathbf{D}\mathbf{U}^\top) + \frac{1}{2}\operatorname{Tr}(\mathbf{B}^\top(\mathbf{U}\mathbf{D}^+\mathbf{U}^\top)\mathbf{B}) \\
&= \frac{1}{2}\operatorname{Tr}(\mathbf{U}^\top\mathbf{U}\mathbf{D}) + \frac{1}{2}\operatorname{Tr}((\mathbf{U}\mathbf{D}^+\mathbf{U}^\top)\mathbf{B}\mathbf{B}^\top) \\
&= \frac{1}{2}\operatorname{Tr}(\mathbf{D}) + \frac{1}{2}\operatorname{Tr}(\mathbf{D}^+(\mathbf{U}^\top\mathbf{B}\mathbf{B}^\top\mathbf{U})) \\
&= \frac{1}{2}\sum_{i:\, d_i>0} d_i + \frac{1}{2}\sum_{i:\, d_i>0}\frac{1}{d_i}[\mathbf{U}^\top\mathbf{B}\mathbf{B}^\top\mathbf{U}]_i^i.
\end{aligned}$$

Keeping $\mathbf{U}$ and $\mathbf{B}$ constant and optimizing for $\mathbf{D}$, we obtain the stationarity condition

$$\mathbf{D} = \operatorname{diag}(d_1,\dots,d_n), \quad \text{with } d_i = \sqrt{[\mathbf{U}^\top\mathbf{B}\mathbf{B}^\top\mathbf{U}]_i^i}. \tag{26}$$

Feeding it back into (25), we have that

$$\mathcal{L} = \sum_{i=1}^n \sqrt{[\mathbf{U}^\top\mathbf{B}]^i[\mathbf{B}^\top\mathbf{U}]_i} = \sum_{i=1}^n \sqrt{[\mathbf{B}^\top\mathbf{U}]_i^\top[\mathbf{B}^\top\mathbf{U}]_i} = \sum_{i=1}^n \left\|[\mathbf{B}^\top\mathbf{U}]_i\right\|_2 = \sum_{i=1}^n \|\mathbf{B}^\top\mathbf{u}_i\|_2 \geq \|\mathbf{B}\|_*. \tag{27}$$

Moreover, equality is achieved in (27) if we choose $\mathbf{B}$ such that its SVD decomposition is $\mathbf{U}\mathbf{D}_B\mathbf{V}^\top$. With this choice, the entries of the diagonal matrix $\mathbf{D}$ given at stationary points by (26) satisfy

$$d_i = \sqrt{[\mathbf{U}^\top\mathbf{B}\mathbf{B}^\top\mathbf{U}]_i^i} = \sqrt{[\mathbf{U}^\top\mathbf{U}\mathbf{D}_B\mathbf{V}^\top\mathbf{V}\mathbf{D}_B\mathbf{U}^\top\mathbf{U}]_i^i} = \sqrt{[\mathbf{D}_B^2]_i^i},$$

and therefore, $\mathbf{D} = \mathbf{D}_B$, which satisfies the constraint $(\mathbf{I} - \mathbf{U}\mathbf{D}\mathbf{D}^+\mathbf{U}^\top)\mathbf{B} = 0$ in (25). To recap, with $\mathbf{B}$ restricted to the set $\{\mathbf{U}\mathbf{D}\mathbf{V}^\top : \mathbf{V}\mathbf{V}^\top = \mathbf{I}_n\}$, the minima are preserved, the last 3 constraints in (25) are satisfied, and the objective becomes the nuclear norm of $\mathbf{B}$, as shown in (27). Thus, the problems (19, 20, 22, 25) are equivalent.

Solution 3 (pLSA and LDA theory, 2):

i. Consider two topics for one document and one word, then (see lecture slides and exercise description)

$$-\ell(\mathbf{x}) = -\log(u_1v_1 + u_2v_2), \quad \mathbf{x} = (u_1, v_1, u_2, v_2).$$

The above function is not convex. Pick

$$\mathbf{x} = (1,1,0,0), \quad \mathbf{y} = (0,0,1,1)$$

$$-\ell(\mathbf{x}/2 + \mathbf{y}/2) = -\log(1/2) > 0 = (-\ell(\mathbf{x}) - \ell(\mathbf{y}))/2,$$

which violates convexity. Note: this does not mean the problem is necessarily hard! One can solve it with Projected Gradient Descent, and find a local minimizer. However, this will be slow.

ii. Let $Q_{zij} \in \{0,1\}$ be 1 if word j of document i is associated with topic z, otherwise $Q_{zij} = 0$. The log-likelihood, conditioned on this information, is

$$-\log(\ell(\mathbf{U},\mathbf{V})) = -\sum_{ij} x_{ij} \log\left(\sum_z Q_{zij} u_{zi} v_{zj}\right).$$

Note that, in the sum with respect to z, only one term is non-zero (and is equal to one). Hence, we can rewrite this as

$$\begin{aligned}
-\log(\ell(\mathbf{U},\mathbf{V})) &= -\sum_{ij} x_{ij} \sum_z Q_{zij} \log(u_{zi} v_{zj}) \\
&= -\sum_{ij} x_{ij} \sum_z Q_{zij} \left(\log(u_{zi}) + \log(v_{zj})\right) \\
&= -\sum_{ij} x_{ij} \sum_z Q_{zij} \left(\log(u_{zi}) + \log(v_{zj}) - \log(Q_{zij})\right).
\end{aligned}$$

This using the convention $0\log(0) = 0$. Note that this corresponds exactly to the lower bounding function seen in the lecture for *variational parameters* q_{zij} such that $\sum_z q_{zij} = 1$:

$$\ell(\mathbf{U},\mathbf{V}) \geq \ell_q(\mathbf{U},\mathbf{V}) = \sum_{ij} x_{ij} \sum_z q_{zij} \left(\log(u_{zi}) + \log(v_{zj}) - \log(q_{zij})\right).$$

We will proceed in this *more general case* and optimize ℓ_q, so that we actually get a proof for the M-step formulas of pLSA. The above objective is convex in each u_{zi} and v_{zj}. Hence, the closed-form solution can be obtained by setting the gradient of the Lagrangian function to zero.

$$\mathcal{L}_{\mathbf{U},\mathbf{V}}(\alpha,\beta) = -\ell_q(\mathbf{U},\mathbf{V}) + \sum_i \alpha_i \left(\sum_z u_{zi} - 1\right) + \sum_z \beta_z \left(\sum_j v_{zj} - 1\right).$$

We proceed with the gradient:

$$\frac{\partial \mathcal{L}}{\partial u_{zi}} = 0 \Leftrightarrow -\sum_j x_{ij} q_{zij} \frac{1}{u_{zi}} + \alpha_i = 0 \Leftrightarrow u_{zi} = \frac{\sum_j x_{ij} q_{zij}}{\alpha_i}.$$

Finally, setting $\partial\mathcal{L}/\partial\alpha_i$ to zero zero yields

$$\sum_z u_{zi} = 1 \Leftrightarrow \frac{\sum_z \sum_j x_{ij} q_{zij}}{\alpha_i} = 1 \Leftrightarrow \alpha_i = \sum_j x_{ij}.$$

Replacing α_i in the formulation of u_{zi} concludes the derivation of u_{zi}. Similarly, we can derive the optimum for the v_{zj}s.

Solution 4 (Implementing pLSA for Discovering Topics in a Corpus):

You can find the code at

https://github.com/dalab/lecture_cil_public/tree/master/exercises/ex5

1. *Why does the maximizer of the lower bound ℓ_q increase at each iteration?*
Solution: Recall that the EM steps are (see slides)

$$\text{E-step:} \qquad q_{zij} = \frac{u_{zi} v_{zj}}{\sum_k u_{ki} v_{kj}}$$
$$\text{M-step:} \qquad (\mathbf{U}, \mathbf{V}) = \arg\max_{\mathbf{U},\mathbf{V}} \ell_q(\mathbf{U}, \mathbf{V}), \quad \text{subject to } \sum_z u_{zi} = 1, \sum_j v_{zj} = 1,$$

where ℓ_q was defined in the solution for the last exercise. Clearly the lower-bound maximizer does not decrease in M-step. So, we just need to show that the ℓ_q does not decrease in the E-step.
We claim that the E-step is derived from the following maximization problem

$$\max_q \ell_q(\mathbf{U}, \mathbf{V}), \quad \text{subject to } \sum_z q_{zij} = 1.$$

To prove this, we need to construct the Lagrangian function and set its gradient to zero:

$$\mathcal{L}_q(\alpha) = -\ell_q(\mathbf{U}, \mathbf{V}) + \sum_{ij} \alpha_{ij} \left(\sum_z q_{zij} - 1 \right).$$

We first derive the optimality condition on q_{zij}:

$$\frac{\partial L}{\partial q_{zij}} = x_{ij} \left(-\log(u_{zi}) - \log(v_{zj}) + \log(q_{zij}) + 1 + \alpha_{ij} \right) = 0 \Leftrightarrow q_{zij} = C x_{ij} u_{zi} v_{zj}$$

Then the optimality condition on α_{ij} implies that $C^{-1} = x_{ij} \sum_z u_{zi} v_{zj}$.

2. *Why does the log-likelihood increase on each iteration?*
Solution: From the last point, the EM algorithm can be written as

$$\text{E step:} \qquad q^{n+1} = \arg\max_q \ell_q(\mathbf{U}^n, \mathbf{V}^n)$$
$$\text{M step:} \qquad (\mathbf{U}^{n+1}, \mathbf{V}^{n+1}) = \arg\max_{\mathbf{U},\mathbf{V}} \ell_{q^{n+1}}(\mathbf{U}, \mathbf{V}),$$

where we skipped constraints. One can readily check (exercise) that $\ell_{q^{n+1}}(\mathbf{U}^n, \mathbf{V}^n) = \log \ell(\mathbf{U}^n, \mathbf{V}^n)$. Hence

$$\log \ell(\mathbf{U}^n, \mathbf{V}^n) = \ell_{q^{n+1}}(\mathbf{U}^n, \mathbf{V}^n) \leq \ell_{q^{n+1}}(\mathbf{U}^{n+1}, \mathbf{V}^{n+1}) \leq \ell_{q^{n+2}}(\mathbf{U}^{n+1}, \mathbf{V}^{n+1}) = \log \ell(\mathbf{U}^{n+1}, \mathbf{V}^{n+1}).$$

Indeed, EM is an alternating maximisation algorithm.