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 4 Solutions (Matrix Approximation & Reconstruction)

Solution 1 (Constrained Optimization with Lagrange Multipliers):

1. Let $\lambda_1 = \max f|_{S^{n-1}}$ and $\mathbf{v}_1 \in S^{n-1}$ a point maximizing f, i.e., $\lambda_1 = f(\mathbf{v}_1)$. Prove that $\boldsymbol{A}\mathbf{v}_1 = \lambda_1 \mathbf{v}_1$.

   The theorem of Lagrange multipliers says that it exists a value $\lambda \in \mathbb{R}$, such that

   $$\nabla(f - \lambda g)(\mathbf{v}_1) = 0 \tag{1}$$

   where $g(\mathbf{x}) = \|\mathbf{x}\|_2^2 - 1$. It follows immediately that $\nabla g(\mathbf{x}) = 2\mathbf{x}$. In order to compute $\nabla f(\mathbf{x})$, we write:

   $$f(\mathbf{x}) = \langle \mathbf{x}, \boldsymbol{A}\mathbf{x} \rangle = \sum_{i,j=1}^{n} \boldsymbol{A}_{ij}\mathbf{x}_i\mathbf{x}_j$$

   and we observe that

   $$\frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_j} = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}$$

   It follows that:

   $$\begin{aligned} \frac{\partial f}{\partial \mathbf{x}_k}(x) &= \frac{\partial}{\partial \mathbf{x}_k} \sum_{i,j=1}^{n} \boldsymbol{A}_{ij}\mathbf{x}_i\mathbf{x}_j \\ &= \sum_{i,j=1}^{n} \boldsymbol{A}_{ij}\delta_{ik}\mathbf{x}_j + \sum_{i,j=1}^{n} \boldsymbol{A}_{ij}\delta_{jk}\mathbf{x}_i \\ &= \sum_{j=1}^{n} \boldsymbol{A}_{kj}\mathbf{x}_j + \sum_{i=1}^{n} \boldsymbol{A}_{ik}\mathbf{x}_i \\ &= \sum_{j=1}^{n} \boldsymbol{A}_{kj}\mathbf{x}_j + \sum_{i=1}^{n} \boldsymbol{A}_{ki}\mathbf{x}_i \\ &= 2\sum_{j=1}^{n} \boldsymbol{A}_{kj}\mathbf{x}_j \\ &= 2(\boldsymbol{A}\mathbf{x})_k \end{aligned}$$

   and therefore $\nabla f(\mathbf{x}) = 2\boldsymbol{A}\mathbf{x}$. By substituting the results above into eq. (1), we get $2\boldsymbol{A}\mathbf{v}_1 = 2\lambda\mathbf{v}_1$, that is $\boldsymbol{A}\mathbf{v}_1 = \lambda\mathbf{v}_1$. Now, by multiplying this expression by $\mathbf{v}_1$ to the left and by recalling that $||\mathbf{v}_1|| = 1$, we get

   $$\langle \mathbf{v}_1, \boldsymbol{A}\mathbf{v}_1 \rangle = \lambda \langle \mathbf{v}_1, \mathbf{v}_1 \rangle = \lambda,$$

   that is, $f(\mathbf{v}_1)(= \lambda_1) = \lambda$. Therefore $\boldsymbol{A}\mathbf{v}_1 = \lambda_1\mathbf{v}_1$.

2. Now, maximize f on the set $S^{n-2} = \left\{\mathbf{x} \in S^{n-1} : \langle \mathbf{x}, \mathbf{v}_1 \rangle = 0\right\}$. More specifically, with $g(\mathbf{x})$ as before and $h(\mathbf{x}) := \langle \mathbf{x}, \mathbf{v}_1 \rangle$, consider $g(\mathbf{x}) = 0$ and $h(\mathbf{x}) = 0$ as the new constraints. Assuming that $\lambda_2 = \max f|_{S^{n-2}}$ and $\mathbf{v}_2 \in S^{n-2}$ is a point maximizing f, prove that $\boldsymbol{A}\mathbf{v}_2 = \lambda_2\mathbf{v}_2$.

   The theorem of Lagrange multipliers ensures that there exists $\lambda, \mu \in \mathbb{R}$, such that:

   $$\nabla(f - \lambda g - \mu h)(\mathbf{v}_2) = 0 \tag{2}$$

   The gradient of h is $\nabla h(\mathbf{x}) = \mathbf{v}_1$ for all $\mathbf{x}$. Substituting this last expression along with $\nabla f(\mathbf{x}) = 2\boldsymbol{A}\mathbf{x}$ and $\nabla g(\mathbf{x}) = 2\mathbf{x}$ into eq. 2 yields

   $$2\boldsymbol{A}\mathbf{v}_2 = 2\lambda\mathbf{v}_2 + \mu\mathbf{v}_1 \tag{3}$$

By multiplying this equation to the left by $\mathbf{v}_1$, recalling that $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle = 0$ and that $\boldsymbol{A}$ is symmetric, we get

$$\begin{aligned} 2\langle \mathbf{v}_1, A\mathbf{v}_2 \rangle &= 2\lambda \langle \mathbf{v}_1, \mathbf{v}_2 \rangle + \mu \langle \mathbf{v}_1, \mathbf{v}_1 \rangle \\ 2\langle \boldsymbol{A}\mathbf{v}_1, \mathbf{v}_2 \rangle &= 0 + \mu \\ 2\lambda_1 \langle \mathbf{v}_1, \mathbf{v}_2 \rangle &= \mu \\ 0 &= \mu. \end{aligned}$$

Therefore, plugging back into eq. (3), we have $\boldsymbol{A}\mathbf{v}_2 = \lambda \mathbf{v}_2$. By multiplying this equation by $\mathbf{v}_2$ to the left we obtain

$$\langle \mathbf{v}_2, \boldsymbol{A}\mathbf{v}_2 \rangle = \lambda \langle \mathbf{v}_2, \mathbf{v}_2 \rangle = \lambda,$$

that is, $f(\mathbf{v}_2)\,(=\lambda_2) = \lambda$. Therefore $\boldsymbol{A}\mathbf{v}_2 = \lambda_2 \mathbf{v}_2$.

3. Applying the same rationale as above, prove that $\boldsymbol{A}\mathbf{v}_3 = \lambda_3 \mathbf{v}_3$, where $\lambda_3 = \max f|_{S^{n-3}} = f(\mathbf{v}_3)$ and $S^{n-3} = \{\mathbf{x} \in S^{n-1} : \langle \mathbf{x}, \mathbf{v}_1 \rangle = 0, \langle \mathbf{x}, \mathbf{v}_2 \rangle = 0\}$.

   Let $k(\mathbf{x}) = \langle \mathbf{x}, \mathbf{v}_2 \rangle$ be the new constraint such that $k(\mathbf{x}) = 0$. There exist λ, μ, ν such that

   $$\nabla(f - \lambda g - \mu h - \nu k)(\mathbf{v}_3) = 0$$

   or

   $$2\boldsymbol{A}\mathbf{v}_3 = 2\lambda \mathbf{v}_3 + \mu \mathbf{v}_1 + \nu \mathbf{v}_2.$$

   Multiplying this equation by $\mathbf{v}_1$ yields $\mu = 0$; the multiplication by $\mathbf{v}_2$ provides $\nu = 0$. Therefore, $\boldsymbol{A}\mathbf{v}_3 = \lambda \mathbf{v}_3$. If we multiply this last equation by $\mathbf{v}_3$, we get $f(\mathbf{v}_3)\,(=\lambda_3) = \lambda$, from which it follows $\boldsymbol{A}\mathbf{v}_3 = \lambda_3 \mathbf{v}_3$. Obviously, we have that $\lambda_1 \geq \lambda_2 \geq \lambda_3$ and $\mathbf{v}_1 \perp \mathbf{v}_2 \perp \mathbf{v}_3$ by construction.

4. By iterating the above procedure, conclude that $\{\mathbf{v_k}\}_{k=1}^n$ forms an orthonormal basis of $\mathbb{R}^n$, with $\boldsymbol{A}\mathbf{v}_k = \lambda_k \mathbf{v}_k$, $\lambda_k = \max f|_{S^{n-k}} = f(\mathbf{v}_k)$, and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$.

   We optimize over the set

   $$S^{n-k} = \{\mathbf{x} \in S^{n-1} : \langle \mathbf{x}, \mathbf{v}_1 \rangle = \langle \mathbf{x}, \mathbf{v}_2 \rangle = \cdots = \langle \mathbf{x}, \mathbf{v}_{k-1} \rangle = 0\}$$

   such that, given $\lambda_k = \max f|_{S^{n-k}} = f(\mathbf{v}_k)$, we have $\boldsymbol{A}\mathbf{v}_k = \lambda_k \mathbf{v}_k$. The last three sets will be S^2 (two-dimensional sphere, for $k = n-2$), S^1 (circumference, for $k = n-1$) and S^0 (for $k = n$) which consists of two points symmetric with respect to the origin, i.e. $S^0 = \{\mathbf{v}_n, -\mathbf{v}_n\}$, S^0 being the space orthogonal to the vector $\mathbf{v}_{n-1}$ in S^1. Clearly $f|_{S^0}$ is constant since in general $f(\mathbf{x}) = f(-\mathbf{x})$. In order to prove that the last vector $\mathbf{v}_n$ (which is orthogonal to $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}$) is an eigenvector of $\boldsymbol{A}$, we just need to observe that

   $$\langle \mathbf{v}_j, \boldsymbol{A}\mathbf{v}_n \rangle = \langle \boldsymbol{A}\mathbf{v}_j, \mathbf{v}_n \rangle = \lambda_j \langle \mathbf{v}_j, \mathbf{v}_n \rangle = 0, \quad \forall j = 1, 2, \ldots, n-1.$$

   In words, the vector $\boldsymbol{A}\mathbf{v}_n$ is orthogonal to the vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}$ and it is therefore proportional to $\mathbf{v}_n$, i.e. $\boldsymbol{A}\mathbf{v}_n = \lambda \mathbf{v}_n$. It follows immediately that $\lambda = \lambda_n = f(\mathbf{v}_n)$. By construction, we have that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, and $\lambda_1 = f(\mathbf{v}_1)$ which is the maximum eigenvalue of $\boldsymbol{A}$, while $\lambda_n = f(\mathbf{v}_n)$ is the minimum eigenvalue of $\boldsymbol{A}$. This method proves that all the eigenvalues of $\boldsymbol{A}$ are real numbers since $\lambda_k = f(\mathbf{v}_k) \in \mathbb{R}\ \forall k = 1, \ldots, n$. The procedure terminates at S^0 because we have constructed n vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ orthogonal to each other and with unitary norm, that is, an orthonormal basis of $\mathbb{R}^n$ consisting of eigenvectors of $\boldsymbol{A}$.

5. Recap in a few words how the Lagrange multiplier method is used as part of PCA.

   PCA applies the above result to the variance-covariance matrix $\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top$.

Solution 2 (Alternating Least Squares for Collaborative Filtering):

1. Is the objective function $L(\mathbf{U}, \mathbf{V})$ convex? If not, prove it.

   The objective is not convex. To prove that, it is sufficient to provide a counter example for $m = n = 1$. This counter example can be generalized to other dimensions by setting all the entries in $\mathbf{U}$ and $\mathbf{V}$ to zero except for those with indexes $(1,1)$:

   $$\mathbf{U} = \begin{bmatrix} u & 0 & \ldots \\ 0 & 0 & \ldots \\ \vdots & \vdots & \ddots \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} v & 0 & \ldots \\ 0 & 0 & \ldots \\ \vdots & \vdots & \ddots \end{bmatrix}.$$

In these cases, the objective reduces to

$$L(u, v) = (a - uv)^2 + \lambda u^2 + \lambda v^2.$$

We are going to use the following theorem: a twice differentiable function is convex on a convex set if and only if its Hessian is positive semi-definite on the interior of that convex set.[1] One can easily verify that the objective $L(u, v)$ is twice differentiable and its Hessian is

$$\nabla^2 L(u, v) = 2 \begin{bmatrix} v^2 + \lambda & 2uv - a \\ 2uv - a & u^2 + \lambda \end{bmatrix}. \tag{4}$$

By setting $u = v = \sqrt{2\lambda + 2|a|}$, we can find that

$$\begin{aligned} \det\left(\nabla^2 L(u, v)\right) &= 4(v^2 + \lambda)(u^2 + \lambda) - 4(2uv - a)^2 \\ &= 4\Big[(3\lambda + 2|a|)^2 - (4\lambda + \underbrace{4|a| - a}_{>2|a|})^2\Big] < 0. \end{aligned}$$

Thus, the Hessian (4) is not positive semi-definite everywhere in $\mathbb{R}^2$ by Sylvester's criterion,[2] and hence $L(u, v)$ is not convex in $\mathbb{R}^2$.

2. Is the objective $L(\mathbf{U}, \mathbf{V})$ convex with respect to $\mathbf{U}$?

   Yes. Notice that the Hessian of $L(\mathbf{U}, \mathbf{V})$ with respect to $\mathbf{u}_i$ is

$$\nabla^2_{\mathbf{u}_i} L(\mathbf{U}, \mathbf{V}) = 2 \sum_{j:(i,j)\in\mathcal{I}} \mathbf{v}_j \mathbf{v}_j^\top + \lambda \mathbf{I}_k,$$

   a positive definite matrix for any $\lambda > 0$. The Hessian of $L(\mathbf{U}, \mathbf{V})$ with respect to $\mathbf{U}$ will be a block diagonal matrix consisting of $\{\nabla^2_{\mathbf{u}_i} L(\mathbf{U}, \mathbf{V})\}_{i=1}^m$ – note that the cross derivatives $\nabla_{\mathbf{u}_i} \nabla_{\mathbf{v}_j}$ vanish. Finally, since the spectrum of block diagonal matrices is the union of the constituent matrices,[3] the Hessian $\nabla^2_{\mathbf{U}} L(\mathbf{U}, \mathbf{V})$ is positive definite. Hence, $L(\mathbf{U}, \mathbf{V})$ is convex with respect to $\mathbf{U}$.

3. Derive the update rule for $\mathbf{u}_i$. Note that the update rule for $\mathbf{v}_j$ is symmetric to that for $\mathbf{u}_i$.

$$\nabla_{\mathbf{u}_i} L(\mathbf{U}, \mathbf{V}) = -2 \sum_{j:(i,j)\in\mathcal{I}} (a_{ij} - \mathbf{u}_i^\top \mathbf{v}_j)\mathbf{v}_j + 2\lambda \mathbf{u}_i$$

   Setting it to zero, we obtain

$$\begin{aligned} \sum_{j:(i,j)\in\mathcal{I}} a_{ij} \mathbf{v}_j &= \sum_{j:(i,j)\in\mathcal{I}} (\mathbf{u}_i^\top \mathbf{v}_j)\mathbf{v}_j + \lambda \mathbf{u}_i \\ &= \sum_{j:(i,j)\in\mathcal{I}} \mathbf{v}_j(\mathbf{u}_i^\top \mathbf{v}_j) + \lambda \mathbf{u}_i \\ &= \sum_{j:(i,j)\in\mathcal{I}} \mathbf{v}_j(\mathbf{v}_j^\top \mathbf{u}_i) + \lambda \mathbf{u}_i \\ &= \Big(\sum_{j:(i,j)\in\mathcal{I}} \mathbf{v}_j \mathbf{v}_j^\top\Big)\mathbf{u}_i + \lambda \mathbf{u}_i \\ &= \Big(\sum_{j:(i,j)\in\mathcal{I}} \mathbf{v}_j \mathbf{v}_j^\top + \lambda \mathbf{I}_k\Big)\mathbf{u}_i. \end{aligned}$$

   Noticing that the matrix is invertible,[4] the update rule is

$$\mathbf{u}_i = \Big(\sum_{j:(i,j)\in\mathcal{I}} \mathbf{v}_j \mathbf{v}_j^\top + \lambda \mathbf{I}_k\Big)^{-1} \sum_{j:(i,j)\in\mathcal{I}} a_{ij} \mathbf{v}_j. \tag{5}$$

[1] https://en.wikipedia.org/wiki/Convex_function#Functions_of_several_variables

[2] https://en.wikipedia.org/wiki/Sylvester%27s_criterion

[3] https://math.stackexchange.com/q/1307998/261538

[4] Check that it is positive definite for $\lambda > 0$. While doing so, notice that the eigenvalues are lower bounded by λ.

4. Suppose the computational complexity of inverting a $k \times k$ matrix is $\mathcal{O}(k^3)$, let n_i be the number of items rated by user i. Find the computational complexity of the update step (5). Use big O notation.

   The complexity is $\mathcal{O}(n_i k^2 + k^3)$. The first term comes from computing the sum of n_i matrices with shape $k \times k$ in (5) and the second term comes from inverting the resulting matrix.

5. For a recommender system, $\mathbf{u}_i$ and $\mathbf{v}_j$ can be interpreted as the low-dimensional representations of the user i and the item j correspondingly. Interpret the update steps of the ALS algorithm in terms of obtaining low-dimensional representations for a recommender system.

   The updates can be interpreted as follows: given low-dimensional representations of the items (resp. users), compute independently the best representation of each user (resp. item). Moreover, recall that (5) is the solution of a ridge regression problem, so further intuition can be gained based on that. For instance, assuming (incorrectly) that $\lambda = 0$, the vector $\mathbf{u}_i \in \mathbb{R}^k$ given by (5) can be seen as the coordinates of the projection of the ratings vector $\mathbf{a}_i = [a_{ij_1},\ a_{ij_2},\ \ldots,\ a_{ij_{n_i}}] \in \mathbb{R}^{n_i}$ on the k-dimensional sub-space spanned by the vectors $\{\mathbf{v}_{j_1},\ \mathbf{v}_{j_2},\ \ldots,\ \mathbf{v}_{j_{n_i}}\}$.[5]

Solution 3 (SGD for Collaborative Filtering):

Consider the given objective function as a sum

$$f(\mathbf{U}, \mathbf{Z}) = \frac{1}{|\Omega|} \sum_{(d,n) \in \Omega} \underbrace{\frac{1}{2} \left[\mathbf{X}_{dn} - (\mathbf{U}\mathbf{Z}^T)_{dn}\right]^2}_{f_{d,n}}$$

where $\mathbf{U} \in \mathbb{R}^{D \times K}$ and $\mathbf{Z} \in \mathbb{R}^{N \times K}$.

- **Stochastic Gradient:** For one fixed element (d, n) of the sum, we derive the gradient entry (d', k) of $\mathbf{U}$, that is, $\frac{\partial}{\partial u_{d',k}} f_{d,n}(\mathbf{U}, \mathbf{Z})$, and analogously for the $\mathbf{Z}$ part.

$$\frac{\partial}{\partial u_{d',k}} f_{d,n}(\mathbf{U}, \mathbf{Z}) = \begin{cases} -\left[\mathbf{X}_{dn} - (\mathbf{U}\mathbf{Z}^T)_{dn}\right] z_{n,k} & \text{if } d' = d \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial}{\partial z_{n',k}} f_{d,n}(\mathbf{U}, \mathbf{Z}) = \begin{cases} -\left[\mathbf{X}_{dn} - (\mathbf{U}\mathbf{Z}^T)_{dn}\right] u_{d,k} & \text{if } n' = n \\ 0 & \text{otherwise} \end{cases}$$

- **Full Gradient:** We have access to all elements $(d, n) \in \Omega$, so we can calculate the partial derivatives of the full gradient for all $(d, n) \in \Omega$. For one specific $(d, n) \in \Omega$, the partial derivatives are the same as that in the stochastic gradient above.

[5] See https://en.wikipedia.org/wiki/Ordinary_least_squares#Projection for more on this interpretation.