Proofs, Exercises and Literature - Recommender Systems and Dimensionality Reduction

1 Proofs

1.1 MF is Nonconvex

Theorem 1 (MF is Nonconvex)**.** *The rank-r matrix factorization problem, defined for a matrix* $D \in \mathbb{R}^{n \times d} \neq \mathbf{0}$ *and a rank* $1 \leq r < \min\{n, d\}$ *as*

$$\min_{X,Y} RSS(X, Y) = \|D - YX^\top\|^2 \qquad s.t.\ X \in \mathbb{R}^{d \times r}, Y \in \mathbb{R}^{n \times r}$$

is a nonconvex optimization problem.

Proof. We show that the $RSS(X, Y)$ is not a convex function. Therefore we assume first that the $RSS(X, Y)$ is a convex function and show then that this assumption leads to a contradiction. Assuming that the $RSS(X, Y)$ is a convex function means that the following inequality has to hold for all matrices $X_1, X_2 \in \mathbb{R}^{d \times r}$ and $Y_1, Y_2 \in \mathbb{R}^{n \times r}$ and $\alpha \in [0, 1]$:

$$RSS(\alpha X_1 + (1-\alpha)X_2, \alpha Y_1 + (1-\alpha)Y_2) \leq \alpha RSS(X_1, Y_1) + (1-\alpha)RSS(X_2, Y_2). \quad (1)$$

For any global minimizer (X, Y) of the rank-r MF problem, $(\gamma X, \frac{1}{\gamma}Y)$ is also a global minimizer for $\gamma \neq 0$. However, for $\alpha = 1/2$ we have that the convex combination attains a function value of

$$\begin{aligned} RSS(\alpha X + (1-\alpha)(\gamma X), \alpha Y + (1-\alpha)(\tfrac{1}{\gamma}Y)) &= RSS\left(\tfrac{1}{2}X + \tfrac{1}{2}(\gamma X), \tfrac{1}{2}Y + \tfrac{1}{2}(\tfrac{1}{\gamma}Y)\right) \\ &= RSS\left(\tfrac{1}{2}(1+\gamma)X, \tfrac{1}{2}(1+\tfrac{1}{\gamma})Y\right) \\ &= \|D - \tfrac{1}{4}(1+\gamma)(1+\tfrac{1}{\gamma})YX^\top\|^2. \end{aligned}$$

We observe that the approximation error in the last equation goes to infinity if $\gamma \to \infty$. Hence, there exists multiple $\gamma > 0$ for which the RSS of the convex combination of two global miinimizers is larger than zero. This contradicts the assumption that the $RSS(X, Y)$ is convex. □

1.2 PCA

Given a dataset, represented by the $n \times d$ matrix D of n observations of d features $\mathtt{F}_1, \ldots, \mathtt{F}_d$, we define a new feature:

$$\mathtt{F}_{d+1} = \sum_{k=1}^{d} \alpha_k \mathtt{F}_k.$$

We have n observations of this new feature, given by

$$D_{\cdot d+1} = \sum_{k=1}^{d} \alpha_k D_{\cdot k} = D\boldsymbol{\alpha} \in \mathbb{R}^n \tag{2}$$

we compute the sample mean as the following matrix-vector product

$$\begin{aligned}
\mu_{\mathtt{F}_{d+1}} &= \frac{1}{n}\sum_{i=1}^{n} D_{id+1} && \text{(Definition sample mean)} \\
&= \frac{1}{n}\mathbf{1}^\top D_{\cdot d+1} && (\mathbf{1} \in \{1\}^n \text{ is constant one vector}) \\
&= \frac{1}{n}\mathbf{1}^\top D\boldsymbol{\alpha} && \text{(Eq. (2))} \\
&= \begin{pmatrix}\mu_{\mathtt{F}_1} & \ldots & \mu_{\mathtt{F}_d}\end{pmatrix}\boldsymbol{\alpha} && \text{(Computation of mean)} \\
&= \boldsymbol{\mu}_{\mathtt{F}}^\top \alpha, && (3)
\end{aligned}$$

where the vector $\boldsymbol{\mu}_{\mathtt{F}}$ gathers all the sample mean values for the given d features. We compute now the sample variance as

$$\begin{aligned}
\sigma^2_{\mathtt{F}_{d+1}} &= \frac{1}{n}\sum_{i=1}^{n}(D_{id+1} - \mu_{\mathtt{F}_{d+1}})^2 && \text{(Definition sample variance)} \\
&= \frac{1}{n}\|D_{\cdot d+1} - \mathbf{1}\mu_{\mathtt{F}_{d+1}}\|^2 && \text{(Definition Euclidean norm, } \mathbf{1} \in \{1\}^n) \\
&= \frac{1}{n}\left\|D\boldsymbol{\alpha} - \mathbf{1}\boldsymbol{\mu}_{\mathtt{F}}^\top\alpha\right\|^2 && \text{(Eq. (3))} \\
&= \frac{1}{n}\left\|\left(D - \mathbf{1}\boldsymbol{\mu}_{\mathtt{F}}^\top\right)\boldsymbol{\alpha}\right\|^2
\end{aligned}$$

We are interested in the direction of maximal variance, so we can restrict the length of vector $\boldsymbol{\alpha}$: $\|\boldsymbol{\alpha}\| = 1$
The direction of largest variance $\boldsymbol{\alpha}$ is the solution to the following optimization problem:

$$\begin{aligned}
\max_{\|\boldsymbol{\alpha}\|=1} \sigma^2_{d+1} &= \max_{\|\boldsymbol{\alpha}\|=1} \frac{1}{n}\left\|\left(D - \mathbf{1}\boldsymbol{\mu}_{\mathtt{F}}^\top\right)\boldsymbol{\alpha}\right\|^2 \\
&= \max_{\|\boldsymbol{\alpha}\|=1} \frac{1}{n}\boldsymbol{\alpha}^\top\left(D - \mathbf{1}\boldsymbol{\mu}_{\mathtt{F}}^\top\right)^\top\left(D - \mathbf{1}\boldsymbol{\mu}_{\mathtt{F}}^\top\right)\boldsymbol{\alpha} \\
&= \max_{\|\boldsymbol{\alpha}\|=1} \frac{\boldsymbol{\alpha}^\top C^\top C\boldsymbol{\alpha}}{n},
\end{aligned}$$

where $C = D - \mathbf{1}\boldsymbol{\mu}_{\mathtt{F}}^\top$ is the centered data matrix.

2 Exercises

2.1 SVD

1. Let's have a look at a very simple movie ratings matrix of six users and four movies:

$$D = \begin{pmatrix} 5 & 5 & 1 & 1 \\ 5 & 5 & 1 & 1 \\ 5 & 5 & 1 & 1 \\ 1 & 1 & 5 & 5 \\ 1 & 1 & 5 & 5 \\ 5 & 5 & 5 & 5 \end{pmatrix}$$

   (a) Which patterns of movie preferences can you detect in the matrix D?

   (b) Can you denote a rank-2 factorization of D which reflects the assignment of users to the patterns you found?

   (c) Compute a rank-2 truncated SVD of D. Do the movie patterns denoted by the SVD solution reflect the patterns you identified?

Solution: This matrix has two obvious movie patterns. The first pattern is $(5, 5, 1, 1)$, indicating that movie 1 and 2 is liked (rated with 5/5 stars) and the last two movies are not liked (rated with 1/5 stars). The second pattern is $(1, 1, 5, 5)$, indicating that the first two movies are not liked and the last two are liked. The last user likes all movies, which could be understood like he belongs to both patterns. In fact, we can write the matrix D as the following rank-2 MF:

$$D = \begin{pmatrix} 5 & 5 & 1 & 1 \\ 5 & 5 & 1 & 1 \\ 5 & 5 & 1 & 1 \\ 1 & 1 & 5 & 5 \\ 1 & 1 & 5 & 5 \\ 5 & 5 & 5 & 5 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 5 & 5 & 1 & 1 \\ 1 & 1 & 5 & 5 \end{pmatrix}. \tag{4}$$

This particular factorization will yet bot be found by SVD, since the rows in the second factor matrix are not orthogonal.

The truncated SVD yields:

$$U_{\cdot\{1,2\}} \Sigma_{\{1,2\}\{1,2\}} V^{\top}_{\cdot\{1,2\}} = \begin{pmatrix} 0.4 & -0.4 \\ 0.4 & -0.4 \\ 0.4 & -0.4 \\ 0.3 & 0.5 \\ 0.3 & 0.5 \\ 0.6 & 0.1 \end{pmatrix} \begin{pmatrix} 16.8 & 0 \\ 0 & 8.8 \end{pmatrix} \begin{pmatrix} 0.6 & 0.6 & 0.4 & 0.4 \\ -0.4 & -0.4 & 0.6 & 0.6 \end{pmatrix}$$

We see that the decomposition of SVD is quite different from the one in Eq. (4). The first pattern $V_{\cdot 1}$ indicates more general information of the data, like movie popularity. Since there are more users who like movies 1 and 2 than users who like movies 3 and 4, movies 1 and 2 get a slightly higher score of 0.6 (vs. 0.4) in the first pattern $V_{\cdot 1}$. The second pattern $V_{\cdot 2}$ indicates the duality in taste of movies 1,2 and 3,4 by the sign (-0.4 vs. 0.6). Likewise, we recognize the difference between users in the matrix U. There are three user patterns: $(0.4, -0.4)$, $(0.3, 0.5)$ and $0.6, 0.1$.

2. Consider the movie recommendation matrix from the lecture, whose missing values are imputed with the mean value of $\mu = 3$:

$$D = \begin{pmatrix} 5 & \mu & 1 & 1 \\ \mu & 1 & 5 & \mu \\ 2 & 1 & 5 & 3 \\ 4 & \mu & 4 & 2 \\ 5 & 5 & \mu & 1 \\ \mu & 1 & 5 & 3 \end{pmatrix}$$

In the example of the lecture, we have used a rank of 2. Try using a rank of 1, 3 and 4 and evaluate the obtained recommendations. Which rank would you choose?

Solution: For this example, a rank of 2 is indeed the best fit. The rank-1 truncated SVD is given as follows:

$$U_1 \Sigma_1 V_1^\top = \begin{pmatrix} 2.9 & 2.0 & 3.2 & 1.8 \\ 3.3 & 2.2 & 3.6 & 2.0 \\ 3.3 & 2.2 & 3.6 & 2.0 \\ 3.8 & 2.6 & 4.2 & 2.3 \\ 4.0 & 2.8 & 4.4 & 2.5 \\ 3.6 & 2.5 & 4.0 & 2.2 \end{pmatrix}.$$

This factorization is not able to approximate the known movie preferences, and hence, we can not derive suitable predictions from this model. The rank 2 factorization is known from the lecture and can be used to give recommendations:

$$U_2 \Sigma_2 V_2^\top = \begin{pmatrix} 4.4 & 3.7 & 1.4 & 0.6 \\ 2.1 & 0.9 & 5.0 & 2.9 \\ 2.1 & 0.9 & 5.0 & 2.9 \\ 4.1 & 2.9 & 3.9 & 2.1 \\ 5.5 & 4.4 & 2.7 & 1.3 \\ 2.7 & 1.4 & 5.1 & 3.0 \end{pmatrix}.$$

The rank of 3 and 4 factorizations of the truncated SVD are adapting too well to the data. We have

$$U_3\Sigma_3V_3^\top = \begin{pmatrix} 5.0 & 3.0 & 1.0 & 0.9 \\ 2.0 & 1.0 & 5.0 & 2.9 \\ 2.0 & 1.0 & 5.0 & 2.9 \\ 4.0 & 3.0 & 4.0 & 2.1 \\ 5.0 & 5.0 & 3.0 & 1.0 \\ 3.0 & 1.1 & 4.9 & 3.1 \end{pmatrix}$$

$$U_4\Sigma_4V_4^\top = \begin{pmatrix} 5.0 & 3.0 & 1.0 & 1.0 \\ 2.0 & 1.0 & 5.0 & 3.0 \\ 2.0 & 1.0 & 5.0 & 3.0 \\ 4.0 & 3.0 & 4.0 & 2.0 \\ 5.0 & 5.0 & 3.0 & 1.0 \\ 3.0 & 1.0 & 5.0 & 3.0 \end{pmatrix}.$$

The imputed mean values are almost perfectly approximated and can this factorization can hence not be used to provide recommendations.

3. Show that the minimum approximation error of a rank-r matrix factorization of the data $D \in \mathbb{R}^{n\times d}$ is equal to the sum of the $\min\{n,d\} - r$ smallest singular values of D:

$$\sigma_{r+1}^2 + \ldots + \sigma_{\min\{n,d\}}^2 = \min_{X,Y} \|D - YX^\top\|^2 \quad \text{s.t. } X \in \mathbb{R}^{d\times r}, Y \in \mathbb{R}^{n\times r}.$$

Solution: We have seen in the lecture that the global minimizers X and Y of the rank-r matrix factorization problem satisfy

$$YX^\top = U_{\cdot\mathcal{R}}\Sigma_{\mathcal{R}\mathcal{R}}V_{\cdot\mathcal{R}}^\top,$$

where $D = U\Sigma V^\top$ is the singular value decomposition of D and $\mathcal{R} = \{1,\ldots,r\}$. We define the matrix

$$\Sigma_0 = \left(\begin{array}{c|c} \Sigma_{\mathcal{R}\mathcal{R}} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right) \in \mathbb{R}^{n\times d}.$$

That is, we create the matrix Σ_0 by attaching $n - r$ constant-zero rows and $d - r$ constant-zero columns. Then, we can write

$$YX^\top = U\Sigma_0V^\top.$$

Due to the orthogonal invariance of the Frobenius norm, we have

$$\begin{aligned}\|D - YX^\top\|^2 &= \|U\Sigma V^\top - U\Sigma_0 V^\top\|^2 \\ &= \|U^\top(U\Sigma V^\top - U\Sigma_0 V^\top)V\|^2 \\ &= \|\Sigma - \Sigma_0\|^2 \\ &= \sigma_{r+1}^2 + \ldots + \sigma_{\min\{n,d\}}^2\end{aligned}$$

2.2 PCA

1. Show that the constraint $Z^\top Z = I$ for $Z \in \mathbb{R}^{n\times r}$, $r < n$ which is imposed for the objective of PCA implies that Z has orthogonal columns which all have a Euclidean norm of one.

Solution: The constraint $Z^\top Z = I$ implies that the entry s, t for $1 \leq s, t \leq r$ of $Z^\top Z$ is equal to zero if $s \neq t$ and equal to one otherwise. We have

$$\begin{aligned} Z^\top Z &= I \\ \Leftrightarrow (Z^\top Z)_{s,t} = Z_{\cdot s}^\top Z_{\cdot t} &= \begin{cases} 0 & \text{if } s \neq t \\ 1 & \text{otherwise} \end{cases} \\ Z_{\cdot s}^\top Z_{\cdot t} &= 0 \qquad \text{for } 1 \leq s \neq t \leq r \qquad (5) \\ Z_{\cdot s}^\top Z_{\cdot s} &= \|Z_{\cdot s}\|^2 = 1 \qquad \text{for } 1 \leq s \leq r \qquad (6) \end{aligned}$$

From Eq. (7) follows that the columns of Z are orthogonal and from Eq. (8) follows that the columns of Z have a Euclidean norm of one.

2. We define a new feature $\mathtt{F}_3 = \mathtt{F}_1 + 2\mathtt{F}_2$, given the following data:

$\mathtt{F}_1$	$\mathtt{F}_2$
2	-2
0	3
1	-2
1	1

   (a) Compute the sample variance of the new feature by computing the new feature values and then computing the variance of these values.

   (b) Compute the sample variance of the new feature by means of the formula derived in the lecture: $\sigma_{\mathtt{F}_3}^2 = \frac{1}{n}\|(D - \mathbf{1}\mu_{\mathtt{F}}^\top)\alpha\|^2$

(c) Plot the data points and the vector α which defines the new feature $\mathtt{F}_3$. Does α indicate a direction of high or low sample variance in the data? How can you compute the variance in the direction of α?

(d) Compute the variance and direction of maximum variance in the data.

Solution: (a) We compute the new feature values, the mean μ_3, and the centered feature values $\mathtt{F}_3 - \mu_3$:

$\mathtt{F}_1$	$\mathtt{F}_2$	$\mathtt{F}_3$	$\mathtt{F}_3 - \mu_3$
2	-2	-2	-3
0	3	6	5
1	-2	-3	-4
1	1	3	2
		$\mu_3 = 1$	

The sample variance of the new feature $\mathtt{F}_3$ is then given as the mean value of the samples $(\mathtt{F}_3 - \mu_3)^2$, that is, we compute

$$\sigma^2_{\mathtt{F}_3} = (3^2 + 5^2 + 4^2 + 2^2)/4 = 13.5$$

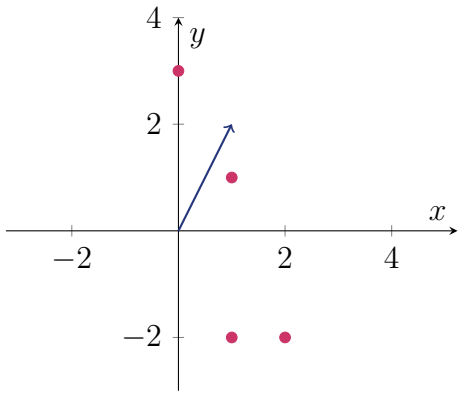
(b) To compute the variance by means of the formula $\sigma^2_{\mathtt{F}_3} = \frac{1}{n}\|(D - \mathbf{1}\mu_{\mathtt{F}}^\top)\alpha\|^2$, we compute first the mean values of the features and then the centered data matrix $D - \mathbf{1}\mu_{\mathtt{F}}$:

	$\mathtt{F}_1$	$\mathtt{F}_2$	$\mathtt{F}_1 - \mu_1$	$\mathtt{F}_2 - \mu_2$	$\mathtt{F}_3 - \mu_3 = \mathtt{F}_1 - \mu_1 + 2(\mathtt{F}_2 - \mu_2)$
	2	-2	1	-2	-3
	0	3	-1	3	5
	1	-2	0	-2	-4
	1	1	0	1	2
μ	1	0	0	0	0

Now, the variance is given as the squared L_2-norm of the samples $\mathtt{F}_3 - \mu_3$:

$$\sigma^2_{\mathtt{F}_3} = (3^2 + 5^2 + 4^2 + 2^2)/4 = 13.5$$

(c) We plot the data points and the vector $\alpha = (1, 2)$.



The variance in the direction of α is not particularly high. Geometrically, the variance would be given by projecting the points on α and then having a look how much the projected samples spread. We can compute the variance by normalizing the vector $\tilde{\alpha} = \alpha/\|\alpha\| = (1,2)/\sqrt{5}$ and then compute the sample variance of the feature $\mathtt{F}_{\tilde{\alpha}} = \tilde{\alpha}_1\mathtt{F}_1 + \tilde{\alpha}_2\mathtt{F}_2$:

$$\sigma^2_{\mathtt{F}_3} = \frac{1}{n}\|(D - \mathbf{1}\mu_{\mathtt{F}}^\top)\tilde{\alpha}\|^2 = (3^2 + 5^2 + 4^2 + 2^2)/5/4 = 2.7$$

(d) We compute the direction of maximum variance in the data by the SVD of the centered data matrix (cf. notebook). The first right singular vector $V_{\cdot 1} = (0.27, -0.96)$ of the centered data matrix denotes the direction of maximum variance (the first principle component) and the squared first singular value divided by the number of samples denotes the variance $\sigma^2_{\mathtt{F}_{V_{\cdot 1}}} = 4.85$ (all values are rounded to two decimal points).

3. Given a data matrix $D \in \mathbb{R}^{n\times d}$, Show that every right singular vector $V_{\cdot k}$ of the centered data matrix $C = D - \mathbf{1}\boldsymbol{\mu}_{\mathtt{F}}^\top$ indicates a new feature $\mathtt{F}_{V_{\cdot k}} = V_{1k}\mathtt{F}_1 + \ldots + V_{dk}\mathtt{F}_d$ whose sample variance is given by the corresponding squared singular value divided by the number of samples σ_k^2/n.

Solution: We have derived in the lecture that the sample variance of a feature defined by $\mathtt{F}_\alpha = \alpha_1\mathtt{F}_1 + \ldots + \alpha_d\mathtt{F}_d$ is given by

$$\sigma^2_{\mathtt{F}_\alpha} = \frac{1}{n}\left\|\left(D - \mathbf{1}\boldsymbol{\mu}_{\mathtt{F}}^\top\right)\boldsymbol{\alpha}\right\|^2 = \frac{1}{n}\left\|C\boldsymbol{\alpha}\right\|^2.$$

If we choose now $\alpha = V_{\cdot k}$, and insert the singular value decomposition of $C = U\Sigma V^\top$,

then we get

$$
\begin{aligned}
\sigma^2_{\mathbf{F}_{V_{\cdot k}}} &= \frac{1}{n}\left\|CV_{\cdot k}\right\|^2 \\
&= \frac{1}{n}\left\|U\Sigma V^\top V_{\cdot k}\right\|^2 \\
&= \frac{1}{n}\left\|\Sigma V^\top V_{\cdot k}\right\|^2 && \text{(orthogonal invariance)} \\
&= \frac{1}{n}\left\|\Sigma \begin{pmatrix} V_{\cdot 1}^\top V_{\cdot k} \\ \vdots \\ V_{\cdot k}^\top V_{\cdot k} \\ \vdots \\ V_{\cdot 1}^\top V_{\cdot d} \end{pmatrix}\right\|^2 = \frac{1}{n}\left\|\Sigma \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}\right\|^2 \\
&= \frac{1}{n}\sigma_k^2. && (\sigma_k = \Sigma_{kk})
\end{aligned}
$$

4. Show that the constraint $Z^\top Z = I$ for $Z \in \mathbb{R}^{n\times r}$, $r < n$ which is imposed for the objective of PCA implies that Z has orthogonal columns which all have a Euclidean norm of one.

Solution: The constraint $Z^\top Z = I$ implies that the entry s,t for $1 \le s,t \le r$ of $Z^\top Z$ is equal to zero if $s \neq t$ and equal to one otherwise. We have

$$
\begin{aligned}
Z^\top Z &= I \\
\Leftrightarrow (Z^\top Z)_{s,t} = Z_{\cdot s}^\top Z_{\cdot t} &= \begin{cases} 0 & \text{if } s \neq t \\ 1 & \text{otherwise} \end{cases} \\
Z_{\cdot s}^\top Z_{\cdot t} &= 0 && \text{for } 1 \le s \neq t \le r \qquad (7) \\
Z_{\cdot s}^\top Z_{\cdot s} &= \|Z_{\cdot s}\|^2 = 1 && \text{for } 1 \le s \le r \qquad (8)
\end{aligned}
$$

From Eq. (7) follows that the columns of Z are orthogonal and from Eq. (8) follows that the columns of Z have a Euclidean norm of one.

3 Recommended Literature

Linear Algebra and Optimization for Machine Learning by Charu C. Aggarwal

7.2 SVD: A Linear Algebra Perspective

7.2.4 Truncated Singular Value Decomposition

7.2.5 Two Interpretations of SVD

7.2.6 Is Singular Value Decomposition Unique?

7.2.7 Two-Way Versus Three-Way Decompositions

7.4 Applications of Singular Value Decomposition

7.4.1 Dimensionality Reduction

8.3 Unconstrained Matrix Factorization

8.3.2 Applications to Recommender Systems