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*

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

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 $\texttt{F}_1, \dots, \texttt{F}_d$, we define a new feature:

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

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

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

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

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

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:

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

where $C = D - \mathbf{1}\boldsymbol{\mu}_{\texttt{F}}^\top$ 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?

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?

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:

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

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.

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

F_1	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: $\sigma_{F_3}^2 = \frac{1}{n}\|(D - \mathbf{1}\mu_F^\top)\alpha\|^2$

(c) Plot the data points and the vector α which defines the new feature 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.

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 $C = D - \mathbf{1}\boldsymbol{\mu}_F^\top$ indicates a new feature $F_{V_{\cdot k}} = V_{1k}F_1 + \ldots + V_{dk}F_d$ whose sample variance is given by the corresponding squared singular value divided by the number of samples σ_k^2/n.

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.

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