Computational Intelligence Laboratory

Matrix Approximation & Reconstruction

Hadi Daneshmand, Gideon Dresdner

14-15 March 2019

ETH Zurich – cil.inf.ethz.ch

Overview

Motivation: Collaborative Filtering

Low-Rank Matrix Factorization

Alternating Least Squares

(Problem 1)

Exact Matrix Recovery

Matrix norms

Convex relaxations

Problems 2.1, 2.2

Section 1

Motivation: Collaborative Filtering

The problem of collaborative filtering

Given: rating matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ (e.g. user-movie matrix)

Not all users have rated all movies $\implies$ only a subset of indices of $\mathbf{A}$ are observed, $\mathcal{I} := \{(i, j) \mid \text{user } i \text{ rated movie } j\}$.

18,000 movies

480,000 users

x	1	1	x	...	x
x	x	x	5	...	x
x	x	3	x	...	x
x	4	3	x	...	2
...	x	x	x	...	x
x	5	x	1	...	x
x	x	3	3	...	x
x	1	x	x	...	2

Goal: fill in x's parsimoniously, i.e. figure out what people think of movies they haven't seen.

The problem of collaborative filtering 2

Goal: fill in x's parsimoniously, i.e. figure out what people think of movies they haven't seen.

Formalization 1: Low-rank matrix factorization

$$\min_{\mathbf{X}:\mathrm{rank}(\mathbf{X})\leq k} \|\mathbf{A}-\mathbf{X}\|_{\mathcal{I}}^2 \tag{1}$$

Formalization 2: Exact matrix recovery

$$\begin{aligned} &\min_{\mathbf{X}} \mathrm{rank}(\mathbf{X}) \\ &\text{such that} \|A-X\|_{\mathcal{I}}=0 \end{aligned} \tag{2}$$

Section 2

Low-Rank Matrix Factorization

Low-rank matrix factorization

- Original problem:

$$\min_{\mathbf{X}:\mathrm{rank}(\mathbf{X})\leq k} \|\mathbf{A}-\mathbf{X}\|_{\mathcal{I}}^2 \tag{3}$$

- Difficult to work with the domain $\{\mathbf{X} | \, \mathrm{rank}(\mathbf{X}) = k\}$ (non-convex, etc.)
- $\Longrightarrow$ reparameterize. Objective is no longer convex, but domain is basically unconstrained:

$$\min_{\mathbf{U},\mathbf{V}} \|\mathbf{A}-\mathbf{U}^\top\mathbf{V}\|_{\mathcal{I}}^2 + \lambda\|\mathbf{U}\|_F^2 + \lambda\|\mathbf{V}\|_F^2 \tag{4}$$

- Add additional regularization

Low rank matrix factorization: regularized

Improving the objective function

$$
\begin{aligned}
L(\mathbf{U}, \mathbf{V}) &:= \|\mathbf{A} - \mathbf{U}^\top \mathbf{V}\|_{\mathcal{I}}^2 + \lambda \|\mathbf{U}\|_F^2 + \lambda \|\mathbf{V}\|_F^2 \\
&= \sum_{(i,j)\in\mathcal{I}} (a_{ij} - \mathbf{u}_i^\top \mathbf{v}_j)^2 + \lambda \sum_{i=1}^{m} \|\mathbf{u}_i\|^2 + \lambda \sum_{j=1}^{n} \|\mathbf{v}_j\|^2,
\end{aligned} \tag{5}
$$

where $\lambda > 0$ is the regularization strength.

Effect of regularization: Prevent entries of $\mathbf{u}_i^\top, \mathbf{v}_j$ from becoming too large

Low rank matrix factorization: regularized

$$L(\mathbf{U}, \mathbf{V}) := \|\mathbf{A} - \mathbf{U}^\top \mathbf{V}\|_{\mathcal{I}}^2 + \lambda \|\mathbf{U}\|_F^2 + \lambda \|\mathbf{V}\|_F^2 \tag{6}$$

Remarks

- L is non-convex w.r.t. $(\mathbf{U}, \mathbf{V})$ even for $m = n = 1$ (check that the Hessian $\nabla^2 L$ is not positive semi-definite).
- However, convex w.r.t. each of $\mathbf{U}$ and $\mathbf{V}$.
- $\Longrightarrow$ devise a strategy to optimize w.r.t. $\mathbf{U}$ and $\mathbf{V}$ separately. One such strategy: the **alternating least squares** algorithm.

The alternating least squares (ALS) algorithm

The algorithm

Initialize $\mathbf{U}, \mathbf{V}$;
while not converged **do**
  for $i = 1, \ldots, m$ **do**
    $\mathbf{u}_i = (\sum_{j:(i,j)\in\mathcal{I}} \mathbf{v}_j \mathbf{v}_j^\top + \lambda \mathbf{I}_k)^{-1} \sum_{j:(i,j)\in\mathcal{I}} a_{ij} \mathbf{v}_j$
  for $j = 1, \ldots, n$ **do**
    $\mathbf{v}_j = (\sum_{i:(i,j)\in\mathcal{I}} \mathbf{u}_i \mathbf{u}_i^\top + \lambda \mathbf{I}_k)^{-1} \sum_{i:(i,j)\in\mathcal{I}} a_{ij} \mathbf{u}_i$

How to derive update rule?

- Differentiate the objective w.r.t. $\mathbf{u}_i$ holding $\mathbf{V}$ constant and set the gradient to zero (see later slide for details).
- Same for $\mathbf{v}_j$. It is symmetric.

Low-rank matrix factorization: using the model

How to utilize the obtained $\mathbf{U}$ and $\mathbf{V}$?

- For entries $(p, q) \notin \mathcal{I}$, complete the missing entries $a_{pq} := \mathbf{u}_p^\top \mathbf{v}_q$.
- Bonus: we have learned low-dimensional representations of users and items in the same space, $\mathbb{R}^k$, $\{\mathbf{u}_i\}_{i=1}^n$ and $\{\mathbf{v}_j\}_{j=1}^m$.

Subsection 2

(Problem 1)

Computing $\nabla_{\mathbf{u}_i} L$ (Problem 1)

$$\nabla_{\mathbf{u}_i} L = -2 \sum_{j:(i,j)\in\mathcal{I}} (a_{ij} - \mathbf{u}_i^\top \mathbf{v}_j)\mathbf{v}_j + 2\lambda\mathbf{u}_i \stackrel{!}{=} 0$$

Therefore,

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

Section 3

Exact Matrix Recovery

Exact matrix recovery: problem definition

Recall

$$\min_{\mathbf{X}} \operatorname{rank}(\mathbf{X}) \quad \text{such that} \|A - X\|_{\mathcal{I}} = 0 \tag{7}$$

where,

$$\operatorname{rank}(\mathbf{X}) = \dim(\operatorname{span}(\mathbf{X})) = \text{n. singular values} \tag{8}$$

and,

$$\|\mathbf{A} - \mathbf{X}\|_{\mathcal{I}} = \sum_{(i,j)\in\mathcal{I}} (a_{ij} - x_{ij})^2 \tag{9}$$

- The rank function is not convex, it is not smooth. Also, the constraint is very stringent. This is a difficult problem.

Subsection 1

Matrix norms

But there is hope: matrix norm zoo

Consider $\mathbf{A} \in \mathbb{R}^{m \times n}$. Let $\sigma(A) = [\sigma_1 \cdots \sigma_k]^\top$ be the vector of singular values. Then we can define some norms:

- $\|\sigma(\mathbf{A})\|_1 = \mathrm{trace}(\sqrt{\mathbf{A}^\top \mathbf{A}}) = \|\mathbf{A}\|_*$ a.k.a. "nuclear norm"
- $\|\sigma(\mathbf{A})\|_2 = \sqrt{\mathrm{trace}(\mathbf{A}^\top \mathbf{A})} = \sqrt{\sum_i \sum_j |a_{ij}|^2} = \|\mathbf{A}\|_F$, a.k.a. "Frobenius norm"
- $\|\sigma(\mathbf{A})\|_\infty = \|\mathbf{A}\|_2 = \sigma_{\max}(\mathbf{A})$, a.k.a. "spectral norm"

And, it turns out that rank is a member of this zoo also (though not formally, since the 0-norm, $\|\cdot\|_0$ does not follow the properties of a norm).

- $\|\sigma(\mathbf{A})\|_0 = \sum_k \sigma_k^0 = \sum_k 1 = \mathrm{rank}(\mathbf{A})$

The zoo has a name

These norms are called **the Schatten p-norms**

$$\|\mathbf{A}\|_p = \|\sigma(\mathbf{A})\|_p$$

and are computed as the p-norm of the vector of singular values of $\mathbf{A}$ where p-norm is defined as:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$$

Subsection 2

Convex relaxations

From 0-norm to 1-norm

Rather than minimize the rank, minimize the nuclear norm $\|\cdot\|_*$. It turns out that the nuclear norm is not only convex (!) but the largest convex function less than rank (and therefore the best). Our new optimization problem is:

$$\begin{aligned} &\min_{\mathbf{X}} \|\mathbf{X}\|_* \\ &\text{such that } \|A - X\|_{\mathcal{I}} = 0 \end{aligned} \tag{10}$$

Subsection 3

Problems 2.1, 2.2

Solution to Problem 2.1

You must know from linear algebra:

- $\mathsf{rank}(\mathbf{XY}) \leq \mathsf{rank}(\mathbf{X}) \quad \forall \mathbf{X} \in \mathbb{R}^{m \times n}, \mathbf{Y} \in \mathbb{R}^{n \times k}$
- $\mathsf{rank}(\mathbf{XY}) = \mathsf{rank}(\mathbf{X}) \quad \forall \mathbf{Y} \in \mathbb{R}^{n \times n},\ \mathsf{rank}(\mathbf{Y}) = n$

The SVD decomposition: $\mathbf{A} = \mathbf{UDV}^\top$

$$\mathsf{rank}(\mathbf{A}) = \mathsf{rank}(\mathbf{UDV}^\top) = \mathsf{rank}(\mathbf{D}) = \#\{\sigma_i > 0\}$$

On the other hand,

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

Therefore, if $\|\mathbf{A}\|_2 \leq 1$ then $\sigma_i \leq 1\ \forall i$. Thus,

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

Solution to Problem 2.2 (part 1)

Recall def. $f: X \to \mathbb{R}$ is convex if $\forall x, y \in X$

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

Want to show. $\forall \lambda \in [0,1],\ \forall \mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$

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

Write SVD decomposition of left side: $\lambda \mathbf{A} + (1-\lambda)\mathbf{B} = \mathbf{U}_\lambda \mathbf{D}_\lambda \mathbf{V}_\lambda^\top$

$$\|\lambda \mathbf{A} + (1-\lambda)\mathbf{B}\|_* = \text{trace}\left(\mathbf{D}_\lambda\right) \tag{11}$$

$$= \text{trace}\left((\mathbf{U}_\lambda^\top \mathbf{U}_\lambda)\mathbf{D}_\lambda(\mathbf{V}_\lambda^\top \mathbf{V}_\lambda)\right) \tag{12}$$

$$= \text{trace}\left(\mathbf{U}_\lambda^\top(\mathbf{U}_\lambda \mathbf{D}_\lambda \mathbf{V}_\lambda^\top)\mathbf{V}_\lambda\right) \tag{13}$$

$$= \text{trace}\left(\mathbf{U}_\lambda^\top(\lambda \mathbf{A} + (1-\lambda)\mathbf{B})\mathbf{V}_\lambda\right) \tag{14}$$

$$= \lambda\,\text{trace}\left(\mathbf{U}_\lambda^\top \mathbf{A} \mathbf{V}_\lambda\right) + (1-\lambda)\,\text{trace}\left(\mathbf{U}_\lambda^\top \mathbf{B} \mathbf{V}_\lambda\right) \tag{15}$$

Solution to Problem 2.2 (part 2)

Thus,

$$\|\lambda\mathbf{A}+(1-\lambda)\mathbf{B}\|_*=\lambda\,\mathrm{trace}\left(\mathbf{U}_\lambda^\top\mathbf{A}\mathbf{V}_\lambda\right)+(1-\lambda)\,\mathrm{trace}\left(\mathbf{U}_\lambda^\top\mathbf{B}\mathbf{V}_\lambda\right)$$

Write SVD decomp. of $\mathbf{A}$, $\mathbf{A}=\mathbf{U}_A\mathbf{D}_A\mathbf{V}_A^\top$.Thus,

$$\mathrm{trace}\left(\mathbf{U}_\lambda^\top\mathbf{A}\mathbf{V}_\lambda\right)=\sum_{i=1}^{\min(m,n)}\left[\mathbf{U}_\lambda^\top\mathbf{A}\mathbf{V}_\lambda\right]_i^i=\sum_{i=1}^{\min(m,n)}\left[\mathbf{U}_\lambda^\top\mathbf{U}_A\mathbf{D}_A\mathbf{V}_A^\top\mathbf{V}_\lambda\right]_i^i \tag{16}$$

$$=\sum_{i=1}^{\min(m,n)}\sum_{j=1}^{\min(m,n)}\left[\mathbf{U}_\lambda^\top\mathbf{U}_A\right]_j^i\sigma_j(\mathbf{A})\left[\mathbf{V}_A^\top\mathbf{V}_\lambda\right]_i^j \tag{17}$$

$$=\sum_{j=1}^{\min(m,n)}\sigma_j(\mathbf{A})\sum_{i=1}^{\min(m,n)}\left[\mathbf{U}_\lambda^\top\mathbf{U}_A\right]_j^i\left[\mathbf{V}_A^\top\mathbf{V}_\lambda\right]_i^j \tag{18}$$

Solution to Problem 2.2 (part 3)

$$= \sum_{j=1}^{\min(m,n)} \sigma_j(\mathbf{A}) \sum_{i=1}^{\min(m,n)} \left[\mathbf{U}_\lambda^\top \mathbf{U}_A\right]_j^i \left[\mathbf{V}_A^\top \mathbf{V}_\lambda\right]_i^j \tag{19}$$

$$\leq \sum_{j=1}^{\min(m,n)} \sigma_j(\mathbf{A}) \left\| \left[\mathbf{U}_\lambda^\top \mathbf{U}_A\right]_j \right\|_2 \left\| \left[\mathbf{V}_A^\top \mathbf{V}_\lambda\right]^j \right\|_2 \quad \text{(Cauchy-Schwartz)} \tag{20}$$

$$= \sum_{j=1}^{\min(m,n)} \sigma_j(\mathbf{A}) = \|\mathbf{A}\|_* \quad \text{(Frobenius norm invariant to rotation)} \tag{21}$$