Online Matrix Completion with Side Information

Mark Herbster, Stephen Pasteris, Lisa Tse

NEURAL INFORMATION PROCESSING SYSTEMS

UCL

Abstract

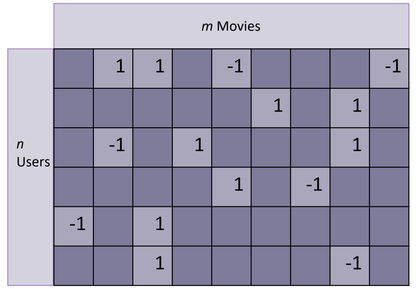
We present online transductive and inductive algorithms for binary matrix completion with side information. For these algorithms, we prove novel mistake and expected regret bounds. In the case of no side information, the bounds scale with the dimensionality of the matrix. In the case of ideal side information, these scale with the number of underlying latent factors.

The Model

Online Matrix Completion

On a trial $t = 1, \ldots, T$:

1. the learner is queried by the *environment* to predict matrix entry (i_t, j_t)
2. the learner predicts a label $\hat{y}_t \in \{-1, 1\}$
3. the learner receives a label $y_t \in \{-1, 1\}$ from the environment and
4. a mistake is incurred if $y_t \neq \hat{y}_t$.



Side Information

Assume that we are given additional side information about each row and column. For instance with movies, we might have “genre” information and for the users, we might have “demographics”. For our purposes, we assume that this can be summarised as $m \times m$ positive definite matrix M for the row side information and $n \times n$ positive definite matrix N for the column side information.

Previous results in online learning

Matrix completion on real-valued matrices [1]

Binary matrix completion [2]

Mistake bound:

$$\sum_{t\in[T]} [y_t \neq \hat{y}_t] \leq \tilde{\mathcal{O}}\left((m+n)\,\mathsf{mc}^2(\boldsymbol{U})\right).$$

No side information

Matrix completion with graph side information on specific matrix classes with only mistake bounds [3,4]

Main Results

General regret and mistake bounds

Mistake bound:

$$\sum_{t\in[T]} [y_t \neq \hat{y}_t] \leq \tilde{\mathcal{O}}\left(\mathsf{mc}^2(\mathbf{U})\mathcal{D}\right).$$

Expected regret bound:

$$\sum_{t\in[T]} \mathbb{E}[y_t \neq \hat{y}_t] - \sum_{t\in[T]} [y_t \neq U_{i_t j_t}] \leq \tilde{\mathcal{O}}\left(\sqrt{\mathcal{D}\|\boldsymbol{U}\|^2_{\mathsf{max}} T}\right)$$

for all $\boldsymbol{U} \in \{-1, 1\}^{m\times n}$.

The *max-norm*:

$$\|\boldsymbol{U}\|_{\mathsf{max}} := \min_{\boldsymbol{P}\boldsymbol{Q}^\top = \boldsymbol{U}} \left\{ \max_{1\leq i\leq m} \|\boldsymbol{P}_i\| \times \max_{1\leq j\leq n} \|\boldsymbol{Q}_j\| \right\},$$

where the minimum is over all matrices $\boldsymbol{P}$, $\boldsymbol{Q}$ and every integer d.

The *margin complexity*:

$$\mathsf{mc}(\boldsymbol{U}) := \min_{\boldsymbol{V}\in\mathsf{SP}^1(\boldsymbol{U})} \|\boldsymbol{V}\|_{\mathsf{max}} = \min_{\boldsymbol{P}\boldsymbol{Q}^\top\in\mathsf{SP}(\boldsymbol{U})} \max_{ij} \frac{\|\boldsymbol{P}_i\|\,\|\boldsymbol{Q}_j\|}{|\langle \boldsymbol{P}_i, \boldsymbol{Q}_j\rangle|}$$

where the minimum is over all matrices $\boldsymbol{P}$, $\boldsymbol{Q}$ and every integer d. Here, $\mathsf{SP}(\boldsymbol{U}) = \{\boldsymbol{V} \in \Re^{m\times n} : \forall_{ij} V_{ij}U_{ij} > 0\}$ and $\mathsf{SP}^1(\boldsymbol{U}) = \{\boldsymbol{V} \in \Re^{m\times n} : \forall_{ij} V_{ij}U_{ij} \geq 1\}$

The *quasi-dimension*:

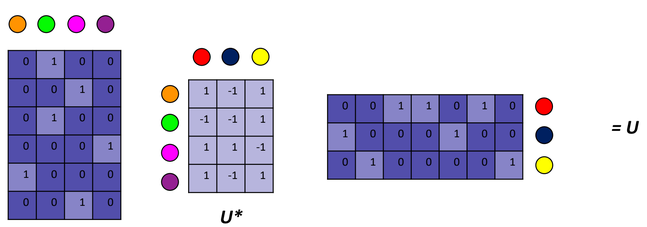
$$\mathcal{D}^\gamma_{M,N}(\boldsymbol{U}) := \min_{\hat{P}\hat{Q}^\top=\gamma\boldsymbol{U}} \mathcal{R}_M\,\mathsf{tr}\left(\hat{\boldsymbol{P}}^\top M\hat{\boldsymbol{P}}\right) + \mathcal{R}_N\,\mathsf{tr}\left(\hat{\boldsymbol{Q}}^\top N\hat{\boldsymbol{Q}}\right),$$

where the minimum is over all row-normalized matrices $\hat{\boldsymbol{P}}$, $\hat{\boldsymbol{Q}}$ and every integer d and where $\mathcal{R}_M := \max_{i\in[m]} M^+_{ii}$. If the infimum does not exist then $\mathcal{D}^\gamma_{M,N}(\boldsymbol{U}) := +\infty$.

Biclustered matrices

The class of (k, ℓ)-binary-biclustered matrices is defined as

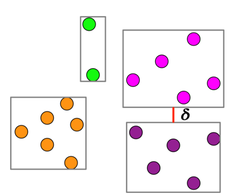
$$\mathbb{B}^{m,n}_{k,\ell} = \{\boldsymbol{U} \in \{-1, 1\}^{m\times n} : \boldsymbol{r} \in [k]^m, \boldsymbol{c} \in [\ell]^n, \boldsymbol{U}^* \in \{-1, 1\}^{k\times\ell}, U_{ij} = U^*_{r_i c_j}, i \in [m], j \in [n]\}$$



$$\mathsf{mc}(\boldsymbol{U}) \leq ||\boldsymbol{U}||_{\mathsf{max}} = ||\boldsymbol{U}^*||_{\mathsf{max}} \leq \min\left(\sqrt{k}, \sqrt{\ell}\right)$$

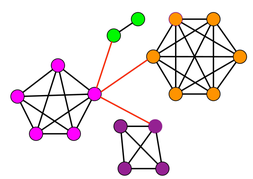
Side information Examples

Feature vectors in well-separated boxes with the min kernel



$\mathcal{D} \leq \mathcal{O}(k^2 + \ell^2)$

Graph Laplacians



$\mathcal{D} \leq \mathcal{O}(k + \ell)$

Algorithms

Transductive algorithm: instance of MEG [5]

Parameters: Learning rate: $0 < \eta$, quasi-dimension estimate: $1 \leq \widehat{\mathcal{D}}$, margin estimate: $0 < \gamma \leq 1$, non-conservative flag $[\textsc{non-conservative}] \in \{0, 1\}$ and side information matrices $\boldsymbol{M} \in \boldsymbol{S}^m_{++}$, $\boldsymbol{N} \in \boldsymbol{S}^n_{++}$ with $m + n \geq 3$

Initialization: $\mathbb{M} \leftarrow \emptyset$; $\tilde{\boldsymbol{W}}^1 \leftarrow \frac{\widehat{\mathcal{D}}}{(m+n)} \boldsymbol{I}^{m+n}$.

For $t = 1, \ldots, T$

- Receive pair $(i_t, j_t) \in [m] \times [n]$.
- Define

$$\tilde{\boldsymbol{X}}^t := \boldsymbol{x}^t(\boldsymbol{x}^t)^\top := \left[\frac{\sqrt{\boldsymbol{M}^+}\boldsymbol{e}^{i_t}_m}{\sqrt{2\mathcal{R}_M}}; \frac{\sqrt{\boldsymbol{N}^+}\boldsymbol{e}^{i_t}_n}{\sqrt{2\mathcal{R}_N}}\right]\left[\frac{\sqrt{\boldsymbol{M}^+}\boldsymbol{e}^{i_t}_m}{\sqrt{2\mathcal{R}_M}}; \frac{\sqrt{\boldsymbol{N}^+}\boldsymbol{e}^{i_t}_n}{\sqrt{2\mathcal{R}_N}}\right]^\top. \quad (3)$$

- Predict

$Y_t \sim \textsc{Uniform}(-\gamma, \gamma)\times[\textsc{non-conservative}]$; $\bar{y}_t \leftarrow \mathsf{tr}\left(\tilde{\boldsymbol{W}}^t\tilde{\boldsymbol{X}}^t\right) - 1$; $\hat{y}_t \leftarrow \mathsf{sign}(\bar{y}_t - Y_t)$.

- Receive label $y_t \in \{-1, 1\}$.
- If $y_t \neq \hat{y}_t$ then $\mathbb{M} \leftarrow \mathbb{M} \cup \{t\}$.
- If $y_t\bar{y}_t < \gamma \times [\textsc{non-conservative}]$ then

$$\tilde{\boldsymbol{W}}^{t+1} \leftarrow \exp\left(\log(\tilde{\boldsymbol{W}}^t) + \eta y_t\tilde{\boldsymbol{X}}^t\right).$$

- Else $\tilde{\boldsymbol{W}}^{t+1} \leftarrow \tilde{\boldsymbol{W}}^t$.

Inductive algorithm

For the inductive algorithm, we are given kernel functions instead of the matrices ***M*** and ***N***. On each trial, we are also given feature vectors for the row and column.

We can show that the transductive and inductive algorithms are prediction-equivalent, up to the value of $\Re_M$ and $\Re_N$.

References

[1] E. Hazan, S. Kale, and S. Shalev-Shwartz (2012)
[2] M. Herbster, S. Pasteris, and M. Pontil (2008)
[3] M. Herbster, S. Pasteris, and M. Pontil (2015)
[4] C. Gentile, M. Herbster, and S. Pasteris (2013)
[5] K. Tsuda, G. Rätsch, and M.K. Warmuth (2005)

DAIS ITA