Nonparametric approach for binary/ordinal matrix completion

Miaoyan Wang, Aug 27, 2020

1 Problem

Suppose that we observe a subset of entries from a binary matrix, $\{y_{ij} \in \{-1, 1\} : (i, j) \in \Omega\}$, where $\Omega \subset [d_1] \times [d_2]$ is the index set of observed entries. How to predict the unobserved entries $\{y_{ij} \in \{-1, 1\} : (i, j) \in \Omega^c\}$?

$$\begin{bmatrix} -1 & ? & ? & -1 & ? \\ ? & 1 & ? & ? & ? \\ -1 & ? & ? & -1 & ? \\ ? & ? & -1 & ? & 1 \end{bmatrix} \tag{1}$$

2 Earlier two-step solution

First, we perform probability estimation based on parametric models. Assume $\boldsymbol{Y}_{ij}$ are independent Bernoulli random variables with success probabilities $P(Y_{ij} = 1)$ for all $(i, j) \in [d_1] \times [d_2]$. We model the probability matrix using the GLM logistic model,

$$\mathbb{P}(Y_{ij} = 1) = \log\left(\frac{\theta_{ij}}{1 - \theta_{ij}}\right), \quad \text{where} \quad \Theta = [\![\theta_{ij}]\!] \in \mathbb{R}^{d_1 \times d_2} \text{ is a rank-}r\text{ matrix.}$$

The constrained maximum log-likelihood estimator is $\hat{\Theta} = [\![\hat{\theta}_{ij}]\!] = \arg\min_{\Theta \in \mathbb{R}^{d_1 \times d_2}, \text{rank}(\Theta) \leq r} L(\Theta)$, where

$$L(\Theta) = -\sum_{(i,j) \in \Omega} \left[\mathbb{1}\{y_{ij} = 1\} \log(e^{-\theta_{ij}} + 1) + \mathbb{1}\{y_{ij} = -1\} \log(e^{\theta_{ij}} + 1) \right].$$

Second, we perform prediction using plug-in estimates,

$$\hat{Y}_{ij} = \text{sign}(\hat{\theta}_{ij} - 0.5), \quad \text{for all } (i, j) \in \Omega^c.$$

3 Proposed nonparametric solution

If our goal is to predict the unobserved entries by two labels {-1,1}, there is no need to estimate the probability. We could directly perform the prediction in a nonparametric fashion. This scenario reduces to a special case of our matrix-valued classification problem.

1. Feature space:

$$\begin{aligned}\mathcal{X} &= \{\boldsymbol{X} \in \{0,1\}^{d_1 \times d_2} \big| \text{only one entry of } \boldsymbol{X} \text{ is one, and others are zero}\} \\ &= \{\boldsymbol{e}_i \otimes \boldsymbol{e}_j : (i,j) \in [d_1] \times [d_2]\}.\end{aligned}$$

2. Outcome space: $\mathcal{Y} \in \{0,1\}$.

3. Uniform marginal distribution $\mathcal{P}(\boldsymbol{X})$ over $\mathcal{X}$. No other distribution assumptions on $P(\boldsymbol{X}, y)$ over the space $(\mathcal{X}, \mathcal{Y})$;

4. i.i.d. training set: $\{(\boldsymbol{X}_{ij}, y_{ij}) : (i,j) \in \Omega\}$, where $\boldsymbol{X}_{ij} = \boldsymbol{e}_i \otimes \boldsymbol{e}_j \in \{0,1\}^{d_1 \times d_2}$ is an indicator matrix specifying the observed index, and $y_{ij} \in \{-1, 1\}$ is the observed label at index (i,j). For example, the features in the training sample for problem (1) are

$$\boldsymbol{X}_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \quad \boldsymbol{X}_2 = \begin{bmatrix} 0 & \cdots & 1 & 0 \\ 0 & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \end{bmatrix}, \cdots, \boldsymbol{X}_7 = \begin{bmatrix} 0 & \cdots & 0 & 0 \\ 0 & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 1 \end{bmatrix}.$$

5. Define low-rank large-margin estimator as $\hat{\Theta} = [\![\hat{\theta}_{ij}]\!] = \arg\min_{\Theta \in \mathbb{R}^{d_1 \times d_2}, \text{rank}(\Theta) \leq r} L(\Theta)$, where

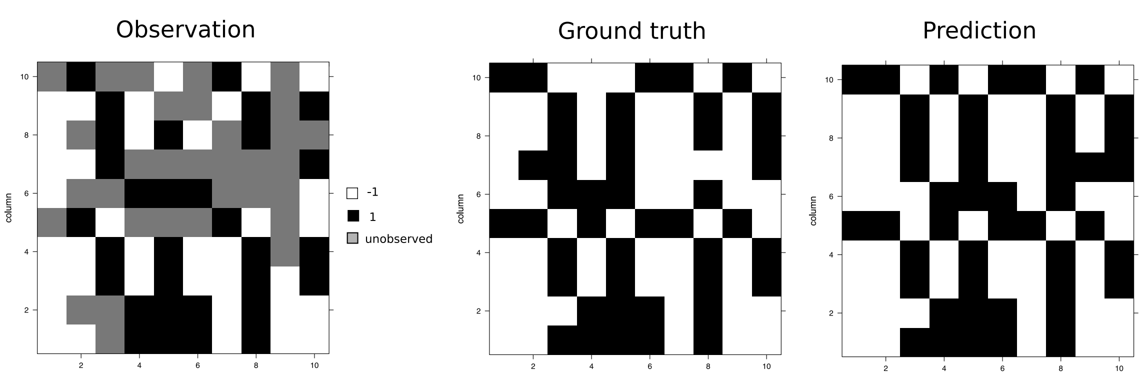
$$L(\Theta) = \sum_{(i,j) \in \Omega} [1 - (y_{ij} \langle \boldsymbol{X}_{ij}, \Theta + b_0 \rangle)]_+ + C\|\Theta\|_F^2. \tag{2}$$

6. Predict unobserved entries using $\hat{y}_{ij} = \text{sign}(\hat{\theta}_{ij})$.

7. Nonparametric probability estimation $\widehat{\mathbb{P}}(y_{ij} = 1 | \boldsymbol{X}_{ij})$ is also possible using a sequence of weighted low-rank classifications (2).

4 Numerical experiment

dimension $d_1 = d_2 = 10$; rank $= 2$; observation probability $p = 0.6$.



	Unobserved		Observed	
	pred = 1	pred = -1	pred = 1	pred = -1
true = 1	16	3	36	1
true = -1	1	12	1	30

5 Theory

Theorem 5.1 (Informally). *For any binary matrix* $\boldsymbol{Y} = [\![y_{ij}]\!] \in \{-1,1\}^{d_1 \times d_2}$, $\delta > 0$ *and integer* $r \geq 1$, *with probability at least* $1-\delta$ *over choosing a subset of* Ω *of entries in* $\boldsymbol{Y}$ *uniformly among all subsets of* $|\Omega|$ *entries, the 0-1 prediction error satisfies*

$$\frac{1}{d_1 d_2} \sum_{(i,j)\in[d_1]\times[d_2]} \mathbb{1}\{y_{ij} \neq \hat{y}_{ij}\} \leq \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \mathbb{1}\{y_{ij} \neq \hat{y}_{ij}\} + \sqrt{\frac{r(d_1+d_2) - \log\delta}{|\Omega|}},$$

where $\hat{y}_{ij} = sign(\hat{\theta}_{ij})$ *and* $\hat{\Theta} = [\![\hat{\theta}_{ij}]\!]$ *is the rank-r large-margin estimator from* (2).