Prediction from low-rank missing data

Elad Hazan  Roi Livni  Yishay Mansour
Princeton U  Hebrew U  Tel-Aviv U
& Microsoft Research  (all of us)
Recommendation systems

<table>
<thead>
<tr>
<th></th>
<th>18,000 movies</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>480,000 users</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>3</td>
</tr>
<tr>
<td>x</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>...</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>5</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>3</td>
</tr>
<tr>
<td>x</td>
<td>1</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>3</td>
</tr>
<tr>
<td>x</td>
<td>1</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Predicting from low-rank missing data

<table>
<thead>
<tr>
<th>x</th>
<th>1</th>
<th>1</th>
<th>x</th>
<th>...</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
<td>5</td>
<td>...</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>3</td>
<td>x</td>
<td>...</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>4</td>
<td>3</td>
<td>x</td>
<td>...</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>...</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>5</td>
<td>x</td>
<td>1</td>
<td>...</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>3</td>
<td>3</td>
<td>...</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>1</td>
<td>x</td>
<td>x</td>
<td>...</td>
<td>2</td>
</tr>
</tbody>
</table>

Gender? Annual income? Will buy “Halo4”? Likes cats or dogs?
Formally:
predicting w. low-rank missing data

Unknown distribution on vectors/rows $x'_i$ in $\{0,1\}^n$, missing data $x_i$ in $\{*,0,1\}^n$ (observed), $X$ has rank $k$, training data $y$ in $\{0,1\}$, every row has $\geq k$ observed entries

Find: efficient machine $M: \{*,0,1\}^n \rightarrow \mathbb{R}$
s.t. with $\text{poly}(\delta, \epsilon, k, n)$ samples, with probability $1-\delta$:

$$
E_i[(M(x_i) - y_i)^2] - \min_{\|w\| \leq 1} E_i[(w^\top x_i - y_i)^2] \leq \epsilon
$$

Kernel version:

$$
E_i[(M(x_i) - y_i)^2] - \min_{\|w\| \leq 1} E_i[(w^\top \phi(x_i) - y_i)^2] \leq \epsilon
$$
Difficulties

- Missing data (usually MOST data is missing)
- Structure in missing data (low rank)
- NP-hard (low-rank reconstruction is a special case)

- Can we use a non-proper approach? (distributional assumptions, convex relaxations for reconstruction)
Missing data (statistics & ML)

Statistics books: i.i.d missing entries. recovery from (large) constant percentage (MCAR, MAR)
Or generative model for missing-ness (MNAR) very different from what we need...
approach 1: Completion & prediction

[Goldberg, Zhu, Recht, Xu, Nowak ‘10]

Method: add predictions y as another column in X, use matrix completion to reconstruct & predict.
Can we use approach 1?
Completion & prediction
[Goldberg, Zhu, Recht, Xu, Nowak '10]
reconstruction is not sufficient nor necessary!!
Can we use approach 1?

Completion & prediction

[Goldberg, Zhu, Recht, Xu, Nowak ‘10]

Both are rank-2 completions
Can we use approach 1?
Completion & prediction

[Goldberg, Zhu, Recht, Xu, Nowak '10]

There is a recoverable k-dim subspace!!
Our results (approach 2)

- Agnostic learning – compete with the best linear predictor that knows all the data, assuming it is rank k (or close)

- Provable

- Efficient (theoretically & practically)

- Significantly improves prediction over standard datasets (Netflix, Jester, ...)

- Generalizes to kernel (non-linear) prediction
Our results (approach 2)
Formally:

Unknown distribution on rows $x'_i$ in $\{0,1\}^n$, missing data $x_i$ in $\{*,0,1\}^n$ (observed), $X'$ has rank $k$, training data $y$ in $\{0,1\}$, every row has $\geq k$ observed entries

We build efficient machine $M: \{*,0,1\}^n \rightarrow \mathbb{R}$ s.t. with $\text{poly}(\log \delta, k, n \log(1/\epsilon))$ samples, with probability $1-\delta$:

$$E_i[(M(x_i) - y_i)^2] - \min_{\|w\| \leq 1} E_i[(w^\top x_i - y_i)^2] \leq \epsilon$$

Extends to arbitrary kernels, # samples increases w. degree (polynomial kernels)
Warm up: agnostic, non-proper & useless (inefficient)

- Data matrix = $X$ of size $m \times n$ ($X'$ is full matrix, $X$ with hidden entries)
  rank = $k$
  every row has $k$ visible entries

- “Optimal predictor” = subspace + linear predictor (SVM)
  - $B$ = basis, $k \times n$ matrix
  - $w$ = predictor, vector in $\mathbb{R}^k$

- Given $x$ = row in $X$, unknown label $y$ predict according to:
  $$B\alpha = x$$
  $$\hat{y} = \alpha^\top w$$
Warm up: inefficient, agnostic

- Given $x = \text{row in } X$, unknown label $y$ predict according to:

$$B\alpha = x$$

$$\hat{y} = \alpha^\top w$$

Inefficiently: learn $B, w$ (bounded sample complexity/regret – compact sets)

(distributional world – bounded fat-shattering dimension)
Learning a hidden subspace is hidden-clique hard! [Berthet & Rigollet ‘13], any hope for efficient algorithms?

Hardness applies only for proper learning!!
Efficient agnostic algorithm

- Let $s$ be the set of $k$ coordinates that are visible in a certain $x$. Then:

$$B\alpha = x \iff \hat{y} = (B_s^{-1}x_s)^\top w$$

Where $B_s$ and $x_s$ are the submatrix (vector) corresponding to the coordinates $s$.

“2 operations” – subset of $s$ rows & inverse
Step 1: “rid of inverse”

Replace inverse by polynomial (need condition on the eigenvalues):

\[ w^\top B_s^{-1} x_s = w^\top \left[ \sum_{j=1}^{\infty} (I_s - B_s)^j \right] x_s \]

Let \( C = I - B \), and up to precision independent of \( k,n \):

\[ w^\top B_s^{-1} x_s = w^\top \left[ \sum_{j=1}^{q} C_s^j \right] x_s + O(\frac{1}{q}) \]

Thus, consider (non-proper) hypothesis class:

\[ g_{C,w}(x_s) = w^\top \left[ \sum_{j=1}^{q} C_s^j \right] x_s \]
Step 2: “rid of column selection”

Observation:

\[ g_{C,w}(x_s) = \sum_{\ell \subseteq s \mid \ell \mid \leq q} w_{\ell_1} C_{\ell_1,\ell_2} \times \ldots \times C_{\ell_{|\ell|-1},\ell_{|\ell|}} \cdot x_{\ell_{|\ell|}} \]

(polynomial in \(C,w\) multiplied by coefficients of \(x\))

Thus, there is a kernel mapping, and vector \(v = v(C,w)\) such that

\[ g_{C,w}(x_s) = v^\top \Phi(x_s) \]

\[ v = v(C, w) \in \mathcal{R}^{n^q} \]
Observation 3

Kernel inner products take the form:

\[ \phi(x_s^{(1)}) \cdot \phi(x_t^{(2)}) = \frac{|s \cap t|^q - 1}{|s \cap t| - 1} \sum_{k \in s \cap t} x_k^{(1)} x_k^{(2)} \]

Inner product \( \phi(x_s)^*\phi(x_t) \) –computed in time \( n^*q \)
Algorithm

Kernel function

\[ \phi(x_s^{(1)}) \cdot \phi(x_t^{(2)}) = \frac{|s \cap t|^q - 1}{|s \cap t| - 1} \sum_{k \in s \cap t} x_k^{(1)} x_k^{(2)} \]

Algorithm: SVM kernel with this particular kernel.

Guarantee – agnostic, non-proper, as good as best subspace embedding.

Nearly same algorithm for all degree q!
λ - regularity

To apply the Taylor series – eigenvalues need to be in unit circle.

Reduces to an assumption on appearance of missing data. This is provably necessary.

Regret bound (sample complexity) depend on this parameter – which is provably a constant independent of the rank/problem dimensions.

Running time – independent of this parameter.
Preliminary benchmarks
MAR data
Preliminary benchmarks
NMAR data (blocks)
Preliminary benchmarks
real data

<table>
<thead>
<tr>
<th></th>
<th>Karma</th>
<th>o-svm</th>
<th>Mcb0</th>
<th>Mcb1</th>
<th>Geom</th>
</tr>
</thead>
<tbody>
<tr>
<td>mamographic</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
<td>0.18</td>
<td>0.17</td>
</tr>
<tr>
<td>bands</td>
<td>0.24</td>
<td>0.34</td>
<td>0.41</td>
<td>0.40</td>
<td>0.35</td>
</tr>
<tr>
<td>hepatitis</td>
<td>0.23</td>
<td>0.17</td>
<td>0.23</td>
<td>0.21</td>
<td>0.22</td>
</tr>
<tr>
<td>wisconsin</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>Horses</td>
<td>0.35</td>
<td>0.36</td>
<td>0.55</td>
<td>0.37</td>
<td>0.36</td>
</tr>
<tr>
<td>Movielens (age)</td>
<td>0.16</td>
<td>0.22</td>
<td>0.25</td>
<td>0.25</td>
<td>NaN</td>
</tr>
</tbody>
</table>
Summary

Prediction from recommendation data:
- Reconstruction+relaxation approach doomed to fail
- Non-proper agnostic learning gives provable guarantees, efficient algorithm
- Benchmarks are promising
- Non-reconstructive approach for other types of missing data? Fully-polynomial alg?
- When does reconstruction fail and agnostic/non-proper learning work?

Thank you!