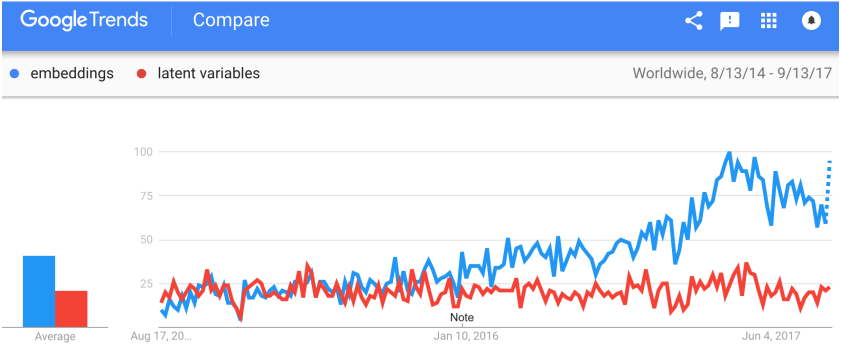
Multi-Target Prediction Using Low-Rank Embeddings: Theory & Practice

Inderjit S. Dhillon
UT Austin & Amazon

ECML PKDD 2017
Skopje, Macedonia
Sept 20, 2017

Embeddings



Google Trends
Compare
embeddings
latent variables
Worldwide, 8/13/14 - 9/13/17
100
75
50
25
Note
Average
Aug 17, 20...
Jan 10, 2016
Jun 4, 2017

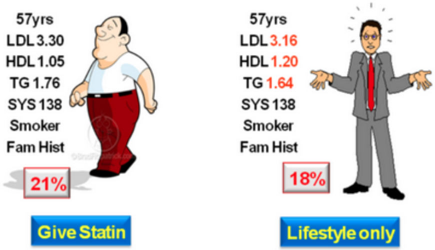
Outline

1. Motivation
2. Multi-Target Prediction
   - Real-world Applications
   - Linear Prediction
   - Bilinear Prediction: Inductive Matrix Completion (IMC)
   - Bilinear Prediction with Noisy Features
   - Positive-Unlabeled Learning
3. Nonlinear Multi-Target Prediction
   - Goal-Directed IMC: Two-layer Neural Network
   - Deep Neural Networks
4. Conclusions and Future Work

Sample Prediction Problems

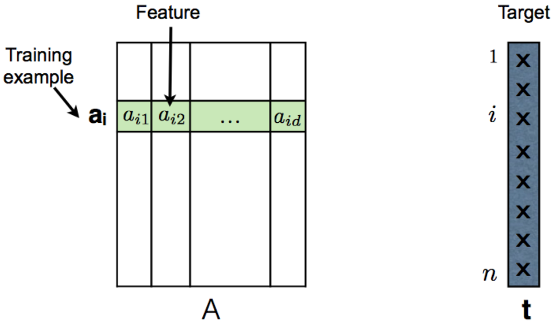
Predicting stock prices

Predicting risk factors in healthcare



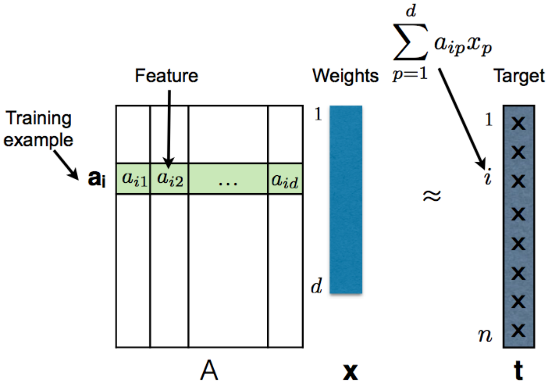
Single-Target Regression

- Real-valued responses (target) **t**
- Predict response for given input data (features) **a**



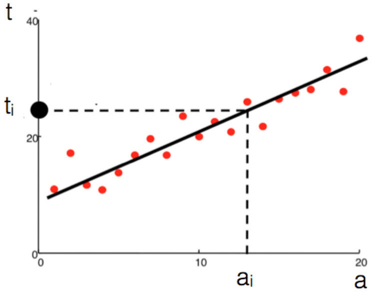
Linear Regression

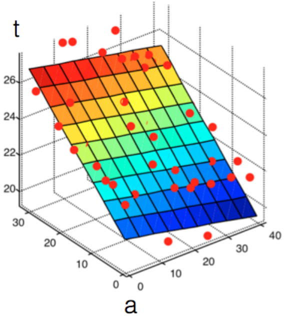
- Estimate target by a linear function of given data $\mathbf{a}$, i.e. $\mathbf{t} \approx \hat{\mathbf{t}} = \mathbf{a}^\top \mathbf{x}$.



Linear Regression: Least Squares

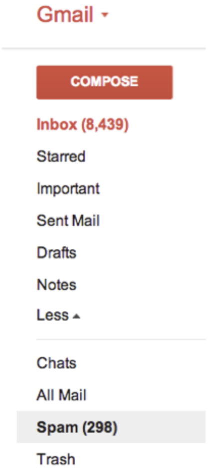
- Choose $\mathbf{x}$ that minimizes
$$J_{\mathbf{x}} = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{a}_i^\top \mathbf{x} - t_i)^2$$
- Closed-form solution: $\mathbf{x}^* = (A^\top A)^{-1}A^\top \mathbf{t}$.



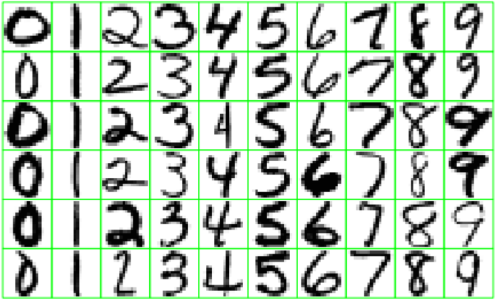


Classification

Spam detection

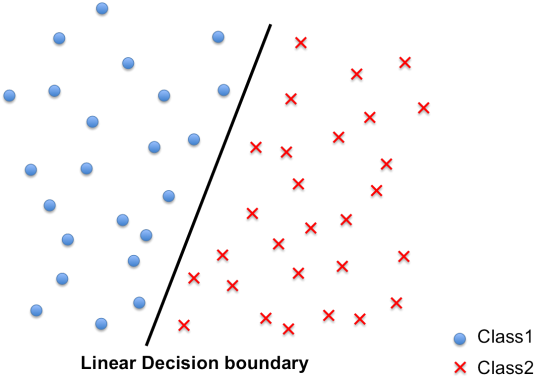


Character Recognition



Binary Classification

- Categorical responses (target) $\mathbf{t}$
- Predict response for given input data (features) $\mathbf{a}$
- Linear methods — decision boundary is a linear surface or hyperplane



Linear Methods for Prediction Problems

Regression:

- Ridge Regression: $J_{\mathbf{x}} = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{a}_i^\top \mathbf{x} - t_i)^2 + \lambda\|\mathbf{x}\|_2^2$.
- Lasso: $J_{\mathbf{x}} = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{a}_i^\top \mathbf{x} - t_i)^2 + \lambda\|\mathbf{x}\|_1$.

Classification:

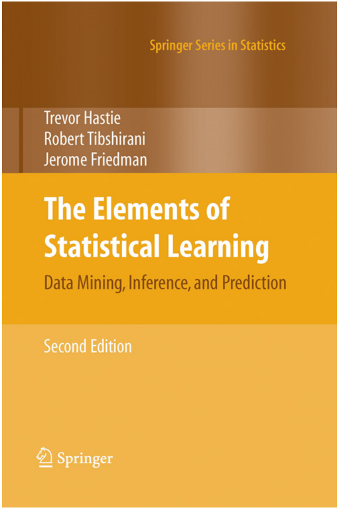
- Linear Support Vector Machines

$$J_{\mathbf{x}} = \frac{1}{2}\sum_{i=1}^{n}\max(0, 1 - t_i\mathbf{a}_i^\top \mathbf{x}) + \lambda\|\mathbf{x}\|_2^2.$$

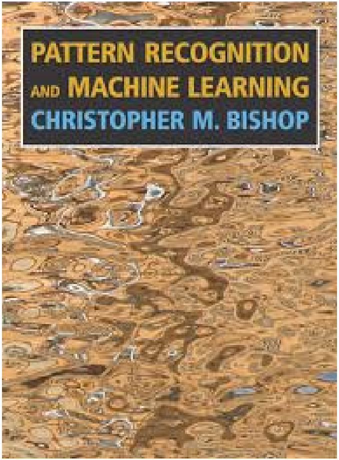
- Logistic Regression

$$J_{\mathbf{x}} = \frac{1}{2}\sum_{i=1}^{n}\log(1 + \exp(-t_i\mathbf{a}_i^\top \mathbf{x})) + \lambda\|\mathbf{x}\|_2^2.$$

Linear Prediction



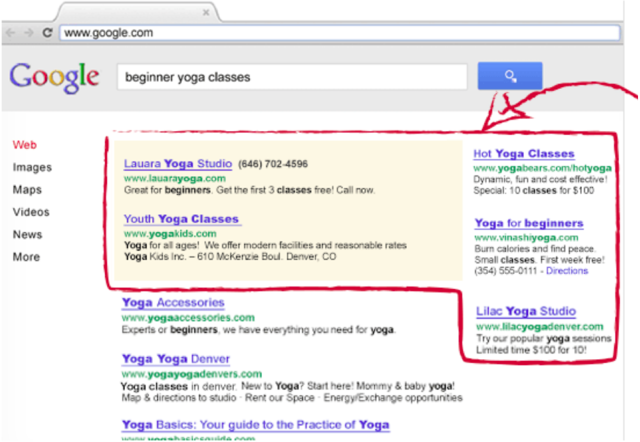
Linear Prediction



Multi-Target Prediction

Modern Prediction Problems

Ad-word Recommendation



Modern Prediction Problems

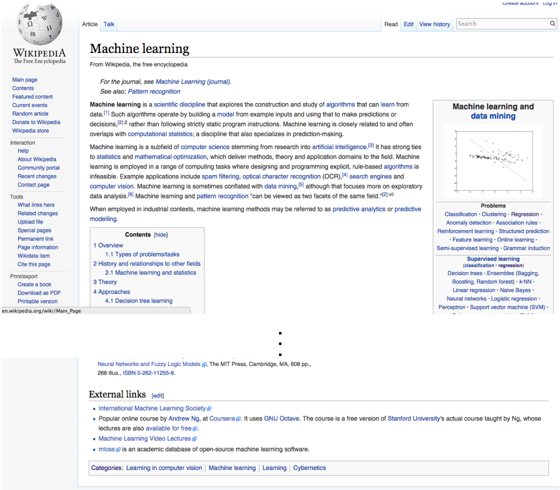
Ad-word Recommendation

- geico auto insurance
- geico car insurance
- car insurance
- geico insurance
- need cheap auto insurance
- geico com
- car insurance coupon code

Modern Prediction Problems

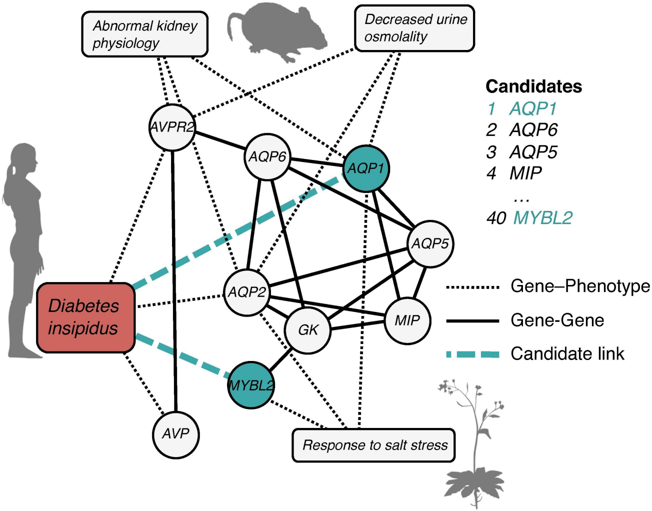
Wikipedia Tag Recommendation

- Learning in computer vision
- Machine learning
- Learning
- Cybernetics



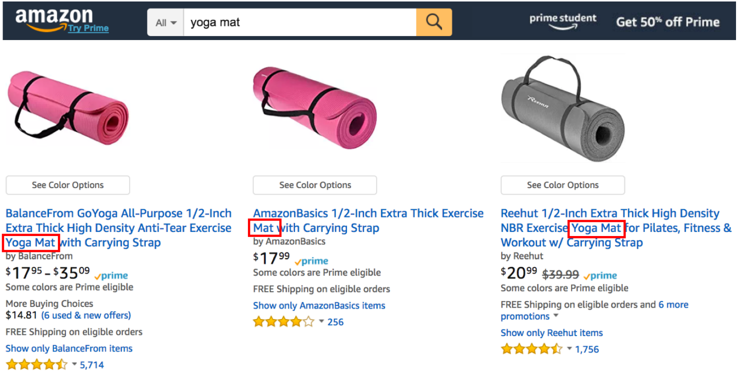
Modern Prediction Problems

Predicting associated disease genes



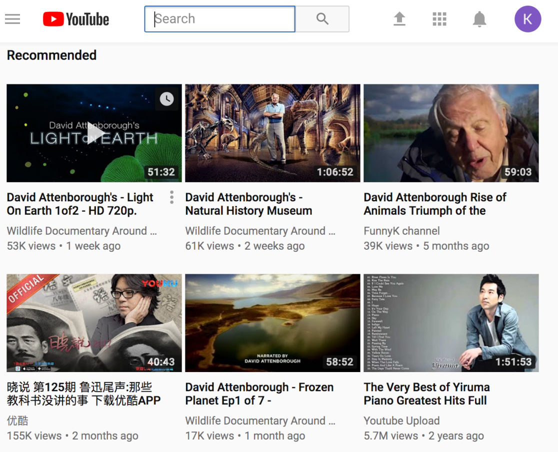
Modern Prediction Problems

Product Search



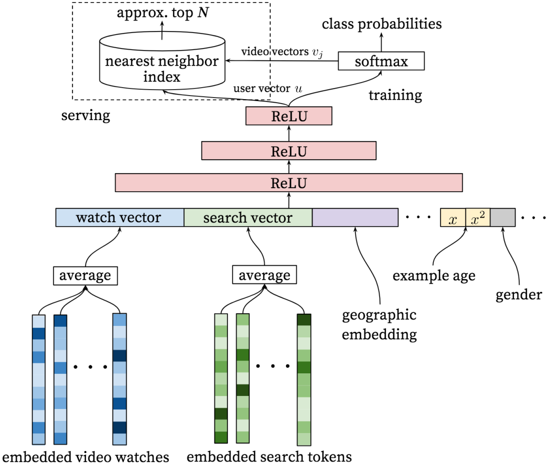
Modern Prediction Problems

Youtube recommendation



Modern Prediction Problems

Model architecture for youtube recommendation



P. Covington, J. Adams, and E. Sargin. *Deep neural networks for youtube recommendations.* Proceedings of the 10th ACM Conference on Recommender Systems. ACM, 2016.

Modern Challenges in Multi-Target Prediction

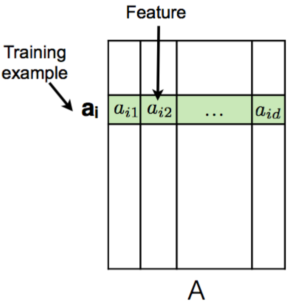
- Millions of correlated targets, and missing target values
- Targets have features
- Noisy Features
- Positive-unlabeled (PU) target values
- Non-linear Structure

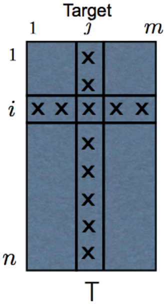
Modern Challenges in Multi-Target Prediction

- Millions of correlated targets, and missing target values
  - Low-rank + Alternating Least Squares
- Targets have features
  - Bilinear Prediction: Inductive Matrix Completion (IMC)
- Noisy Features
  - Dirty IMC
- Positive-unlabeled (PU) target values
  - PU learning for IMC
- Non-linear Structure
  - Deep Learning for IMC

Prediction with Multiple Targets

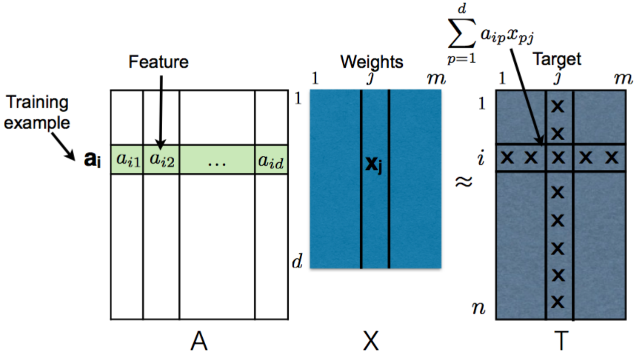
- Input data $\mathbf{a}_i$ is associated with m targets, $\mathbf{t}_i = (t_i^{(1)}, t_i^{(2)}, \ldots, t_i^{(m)})$





Multi-Target Linear Prediction

- Basic model: Treat targets independently
- Estimate regression coefficients $\mathbf{x}_j$ for each target j



Multi-Target Linear Prediction

- Assume targets $\mathbf{t}^{(j)}$ are independent
- Linear predictive model: $\mathbf{t}_i \approx \mathbf{a}_i^\top X$
- Objective for multi-target regression:

$$\min_X \; \|T - AX\|_F^2$$

Multi-Target Linear Prediction

- Assume targets $\mathbf{t}^{(j)}$ are independent
- Linear predictive model: $\mathbf{t}_i \approx \mathbf{a}_i^\top X$
- Objective for multi-target regression:

$$\min_X \ \|T - AX\|_F^2$$

- Closed-form solution:

$$V_A \Sigma_A^{-1} U_A^\top T = \arg\min_X \ \|T - AX\|_F^2$$

where $A = U_A \Sigma_A V_A^\top$ is the thin SVD of A

Multi-Target Linear Prediction

- Assume targets $\mathbf{t}^{(j)}$ are independent
- Linear predictive model: $\mathbf{t}_i \approx \mathbf{a}_i^\top X$
- Objective for multi-target regression:

$$\min_X \; \|T - AX\|_F^2$$

- Closed-form solution:

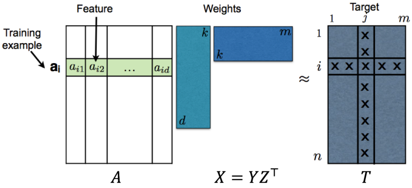
$$V_A \Sigma_A^{-1} U_A^\top T = \arg\min_X \; \|T - AX\|_F^2$$

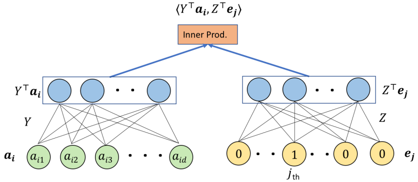
where $A = U_A \Sigma_A V_A^\top$ is the thin SVD of A

In multi-label classification: Binary Relevance (independent binary classifier for each label)

Multi-Target Linear Prediction: Low-rank Model

- Exploit correlations between targets T, where $T \approx AX$
- Reduced-Rank Regression [A.J. Izenman, 1974] — model the coefficient matrix X as *low-rank*





A. J. Izenman. *Reduced-rank regression for the multivariate linear model*. Journal of Multivariate Analysis 5.2 (1975): 248-264.

Multi-Target Linear Prediction: Low-rank Model

- X is rank-k
- Linear predictive model: $\mathbf{t}_i \approx \mathbf{a}_i^\top X$
- Objective for low-rank multi-target regression:

$$\min_{X: rank(X) \leq k} \quad \|T - AX\|_F^2$$

Multi-Target Linear Prediction: Low-rank Model

- X is rank-k
- Linear predictive model: $\mathbf{t}_i \approx \mathbf{a}_i^\top X$
- Objective for low-rank multi-target regression:

$$\min_{X:rank(X)\leq k} \quad \|T - AX\|_F^2$$

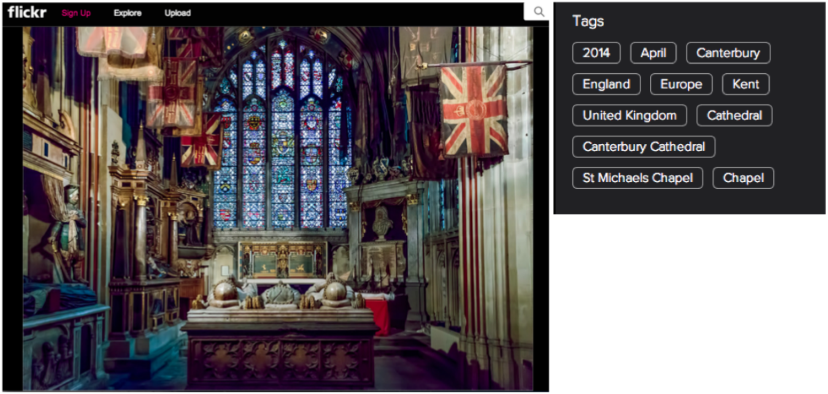
- Closed-form solution:

$$X^* = \arg \min_{X:rank(X)\leq k} \quad \|T - AX\|_F^2$$
$$= \begin{cases} V_A \Sigma_A^{-1} U_A^\top T_k & \text{if } A \text{ is full row rank,} \\ V_A \Sigma_A^{-1} M_k & \text{otherwise,} \end{cases}$$

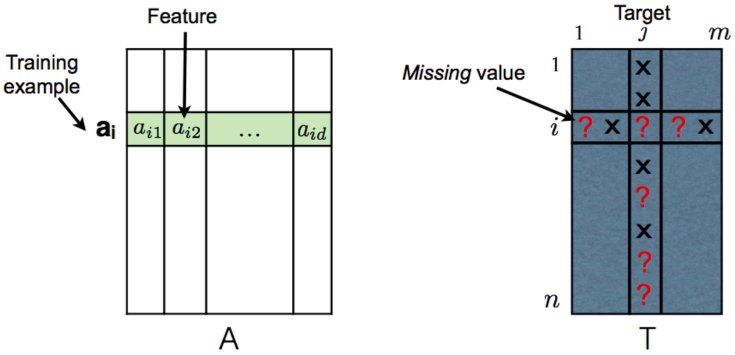
where $A = U_A \Sigma_A V_A^\top$ is the thin SVD of A, $M = U_A^\top T$, and T_k, M_k are the best rank-k approximations of T and M respectively.

Multi-Target Prediction with Missing Target Values

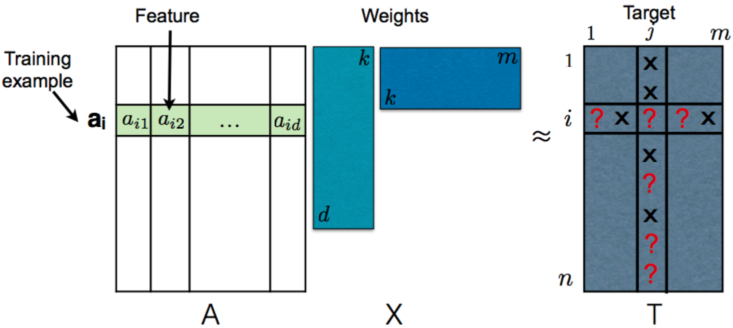
- In many applications, several observations (targets) may be *missing*
- E.g. Recommending tags for images and wikipedia articles



Multi-Target Prediction with Missing Target Values



Multi-Target Prediction with Missing Target Values



- Low-rank model: $\mathbf{t}_i = \mathbf{a}_i^\top X$ where X is low-rank

H. F. Yu, P. Jain, P. Kar, and I. S. Dhillon. *Large-scale Multi-label Learning with Missing Labels*. In Proceedings of The 31st International Conference on Machine Learning, pp. 593-601 (2014).

Multi-Target Prediction with Missing Target Values

- X is rank-k
- Linear predictive model: $\mathbf{t}_i \approx \mathbf{a}_i^\top X$
- Objective for low-rank multi-target regression with missing target values:

$$\min_{X:rank(X)\leq k} \sum_{(i,j)\in\Omega} (\boldsymbol{a}_i^\top X \boldsymbol{e}_j - T_{ij})^2,$$

where Ω is the set of observed targets.

Multi-Target Prediction with Missing Target Values

- X is rank-k
- Linear predictive model: $\mathbf{t}_i \approx \mathbf{a}_i^\top X$
- Objective for low-rank multi-target regression with missing target values:

$$\min_{X:rank(X)\leq k} \sum_{(i,j)\in\Omega} (\boldsymbol{a}_i^\top X \boldsymbol{e}_j - T_{ij})^2,$$

  where Ω is the set of observed targets.
- No closed-form solution

Multi-Target Prediction with Missing Values: Algorithms

- Algorithm 1 (LEML(Nuclear)): Nuclear-norm constraint objective

$$\min_{\|X\|_* \leq \mathcal{X}} \sum_{(i,j)\in\Omega} (\mathbf{a}_i^\top X \boldsymbol{e}_j - T_{ij})^2$$

  - Convex Relaxation

- Algorithm 2 (LEML(ALS)): Alternating Least Squares

$$\min_{Y\in\mathbb{R}^{d\times k}, Z\in\mathbb{R}^{m\times k}} \sum_{(i,j)\in\Omega} (\mathbf{a}_i^\top Y Z^\top \boldsymbol{e}_j - T_{ij})^2 + \lambda(\|Y\|_F^2 + \|Z\|_F^2)$$

  - Alternately minimize w.r.t. Y and Z
  - Non-convex optimization
  - Computationally cheaper than nuclear-norm method

Application: Image Tag Recommendation

NUS-Wide Image Dataset

- 161,780 training images
- 107,879 test images
- 1,134 features
- 1,000 tags

Application: Image Tag Recommendation

- Low-rank Model with $k = 50$:

	time(s)	prec@1	prec@3	AUC
LEML(ALS)	**574**	**20.71**	**15.96**	**0.7741**
WSABIE	4,705	14.58	11.37	0.7658

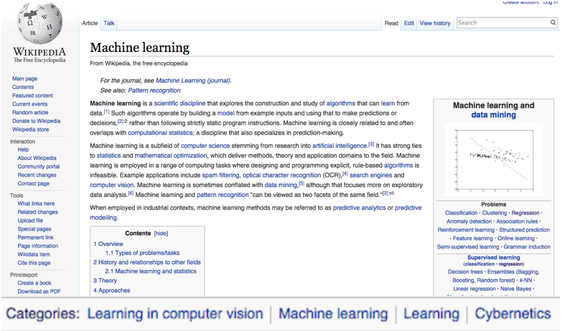
- Low-rank Model with $k = 100$:

	time(s)	prec@1	prec@3	AUC
LEML(ALS)	**1,097**	**20.76**	**16.00**	**0.7718**
WSABIE	6,880	12.46	10.21	0.7597

LEML: H. F. Yu, P. Jain, P. Kar, and I. S. Dhillon. *Large-scale Multi-label Learning with Missing Labels*. In ICML (2014).
WSABIE: J. Weston, S. Bengio, and N. Usunier. *Wsabie: Scaling up to large vocabulary image annotation.* in IJCAI (2011).

Application: Wikipedia Tag Recommendation

Wikipedia Dataset



- 881,805 training wiki pages
- 10,000 test wiki pages
- 366,932 features
- 213,707 tags

Application: Wikipedia Tag Recommendation

- Low-rank Model with $k = 250$:

	time(s)	prec@1	prec@3	AUC
LEML(ALS)	**9,932**	**19.56**	14.43	**0.9086**
WSABIE	79,086	18.91	**14.65**	0.9020

- Low-rank Model with $k = 500$:

	time(s)	prec@1	prec@3	AUC
LEML(ALS)	**18,072**	**22.83**	**17.30**	**0.9374**
WSABIE	139,290	19.20	15.66	0.9058

LEML: H. F. Yu, P. Jain, P. Kar, and I. S. Dhillon. *Large-scale Multi-label Learning with Missing Labels*. In ICML (2014).
WSABIE: J. Weston, S. Bengio, and N. Usunier. *Wsabie: Scaling up to large vocabulary image annotation.* in IJCAI (2011).

Modern Challenges in Multi-Target Prediction

- Millions of correlated targets, and missing target values
  - Low-rank + Alternating Least Squares
- Targets have features
- Noisy Features
- Positive-unlabeled (PU) target values
- Non-linear Structure

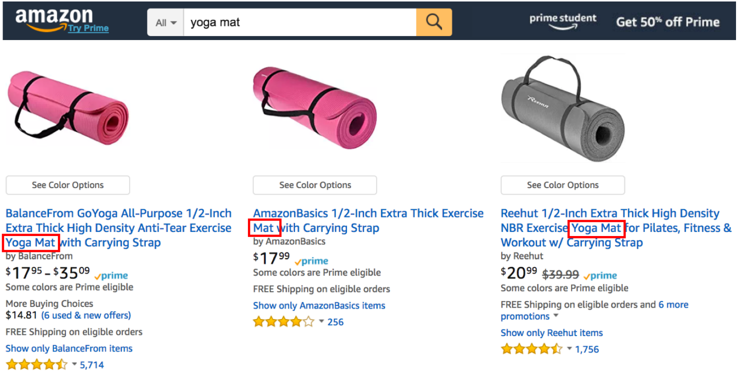
Modern Challenges in Multi-Target Prediction

- Millions of correlated targets, and missing target values
  - Low-rank + Alternating Least Squares
- Targets have features
  - Bilinear Prediction: Inductive Matrix Completion (IMC)
- Noisy Features
- Positive-unlabeled (PU) target values
- Non-linear Structure

Bilinear Prediction: Inductive Matrix Completion (IMC)

Modern Prediction Problems

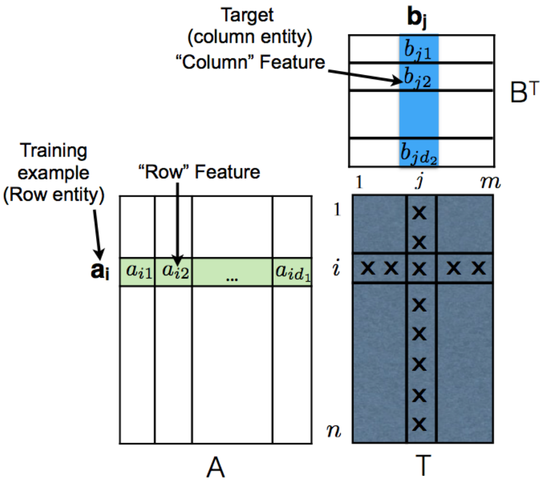
Product Search



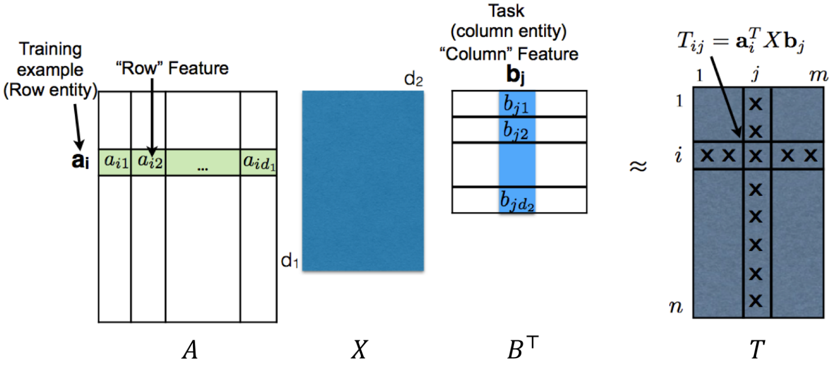
Bilinear Prediction

- Augment multi-target prediction with *features* on targets as well
- Motivated by modern applications of machine learning — bioinformatics, recommendation system with item features
- Need to model *dyadic* or *pairwise* interactions
- Move from linear models to *bilinear* models — linear in input features *as well as* target features

Bilinear Prediction



Bilinear Prediction



Bilinear Prediction

- Bilinear predictive model: $T_{ij} \approx \mathbf{a}_i^\top X \mathbf{b}_j$
- Objective for bilinear predictive model

$$\min_{X} \; \|T - AXB^\top\|_F^2$$

Bilinear Prediction

- Bilinear predictive model: $T_{ij} \approx \mathbf{a}_i^\top X \mathbf{b}_j$
- Objective for bilinear predictive model

$$\min_X \; \|T - AXB^\top\|_F^2$$

- Closed-form solution:

$$V_A \Sigma_A^{-1} U_A^\top T U_B \Sigma_B^{-1} V_B^\top = \arg\min_X \; \|T - AXB^\top\|_F^2$$

where $A = U_A \Sigma_A V_A^\top$, $B = U_B \Sigma_B V_B^\top$ are the thin SVDs of A and B

Bilinear Prediction: Low-rank Model

- X is rank-k
- Bilinear predictive model: $T_{ij} \approx \mathbf{a}_i^\top X \mathbf{b}_j$
- Objective for low-rank bilinear model:

$$\min_{X:rank(X)\leq k} \quad \|T - AXB^\top\|_F^2$$

Bilinear Prediction: Low-rank Model

- X is rank-k
- Bilinear predictive model: $T_{ij} \approx \mathbf{a}_i^\top X \mathbf{b}_j$
- Objective for low-rank bilinear model:

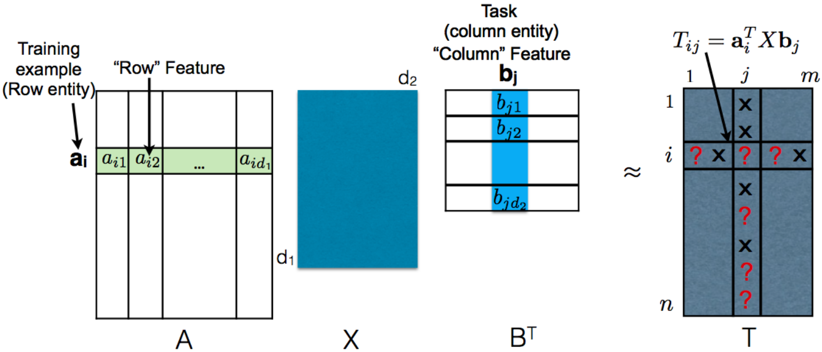
$$\min_{X:rank(X)\leq k} \quad \|T - AXB^\top\|_F^2$$

- Closed-form solution:

$$\begin{aligned} X^* &= \min_{X:rank(X)\leq k} \quad \|T - AXB^\top\|_F^2 \\ &= \begin{cases} V_A \Sigma_A^{-1} U_A^\top T_k U_B \Sigma_B^{-1} V_B^\top & \text{if } A, B \text{ are full row rank,} \\ V_A \Sigma_A^{-1} M_k \Sigma_B^{-1} V_B^\top & \text{otherwise,} \end{cases} \end{aligned}$$

where $A = U_A \Sigma_A V_A^\top$, $B = U_B \Sigma_B V_B^\top$ are the thin SVDs of A and B, $M = U_A^\top T U_B$, and T_k, M_k are the best rank-k approximations of T and M

Bilinear Prediction with Missing Target Values



Bilinear Prediction with Missing Target Values: Algorithms

- Algorithm 1: Nuclear-norm constraint objective

$$\min_{\|X\|_* \leq \mathcal{X}} \sum_{(i,j)\in\Omega} (\mathbf{a}_i^\top X \boldsymbol{b}_j - T_{ij})^2$$

  - Convex Relaxation

- Algorithm 2: Alternating Least Squares (ALS)

$$\min_{Y\in\mathbb{R}^{d\times k}, Z\in\mathbb{R}^{d\times k}} \sum_{(i,j)\in\Omega} (\mathbf{a}_i^\top Y Z^\top \boldsymbol{b}_j - T_{ij})^2 + \lambda(\|Y\|_F^2 + \|Z\|_F^2)$$

  - Non-convex optimization

Bilinear Prediction with Missing Target Values: Algorithms

- Algorithm 1: Nuclear-norm constraint objective

$$\min_{\|X\|_* \leq \mathcal{X}} \sum_{(i,j)\in\Omega} (\mathbf{a}_i^\top X \boldsymbol{b}_j - T_{ij})^2$$

  - Convex Relaxation

- Algorithm 2: Alternating Least Squares (ALS)

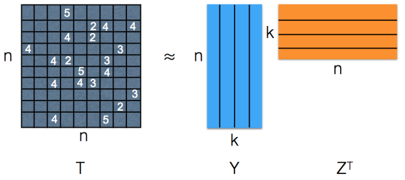
$$\min_{Y\in\mathbb{R}^{d\times k}, Z\in\mathbb{R}^{d\times k}} \sum_{(i,j)\in\Omega} (\mathbf{a}_i^\top Y Z^\top \boldsymbol{b}_j - T_{ij})^2 + \lambda(\|Y\|_F^2 + \|Z\|_F^2)$$

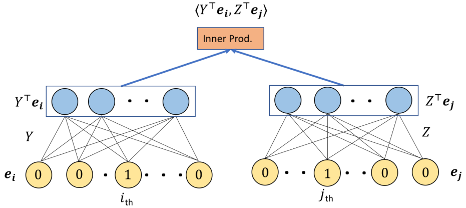
  - Non-convex optimization

- Can we recover the model? How many observations are required?

Recovery Guarantees: Matrix Completion

- Matrix Completion:
  - Recover a low-rank matrix from partially observed entries
- Exact recovery requires $\tilde{O}(kn)$ observed entries



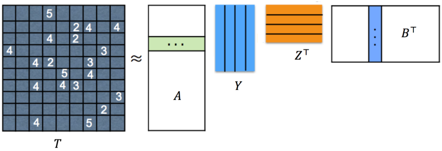


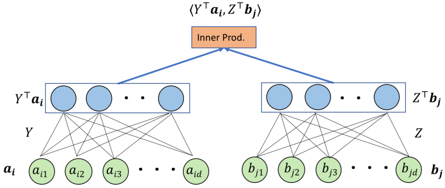
$\tilde{O}(n)$ hides polylog(n)

E. J. Candes and B. Recht. *Exact matrix completion via convex optimization.* Foundations of Computational mathematics (2009).

Inductive Matrix Completion

- Inductive Matrix Completion:
  - Recover a low-rank bilinear model from partially obtained targets
- Degrees of freedom in X are $O(kd)$
- Can we get better sample complexity (than $\tilde{O}(kn)$)?





Recovery Guarantees

Theorem (Recovery Guarantees for Nuclear-norm Minimization)

Let $X_* = U_*\Sigma_* V_*^\top \in \mathbb{R}^{d\times d}$ be the SVD of X_* with rank k, and $T = AX_*B^\top$. Let $\mathcal{X} = \|X_*\|_*$. Assume A, B are orthonormal matrices w.l.o.g., satisfying the incoherence conditions. Then if Ω is uniformly observed with

$$|\Omega| \geq O(kd \log d \log n),$$

the solution of nuclear-norm minimization problem is unique and equal to X_* with high probability.

The incoherence conditions are

$$\textbf{C1.} \ \max_{i\in[n]} \|\mathbf{a}_i\|_2^2 \leq \frac{\mu d}{n}, \ \max_{j\in[n]} \|\mathbf{b}_j\|_2^2 \leq \frac{\mu d}{n}$$

$$\textbf{C2.} \ \max_{i\in[n]} \|U_*^\top \mathbf{a}_i\|_2^2 \leq \frac{\mu_0 k}{n}, \ \max_{j\in[n]} \|V_*^\top \mathbf{b}_j\|_2^2 \leq \frac{\mu_0 k}{n}$$

K. Zhong, P. Jain, I. S. Dhillon. *Efficient Matrix Sensing Using Rank-1 Gaussian Measurements*. In ALT (2015).

Recovery Guarantees

Theorem (Convergence Guarantees for ALS)

Let X_* be a rank-k matrix with condition number β and $T = AX_*B^\top$. Assume A, B are orthogonal w.l.o.g. and satisfy the incoherence conditions. Then if Ω is uniformly sampled with

$$|\Omega| \geq O(k^4\beta^2 d \log d),$$

then after H iterations of ALS, $\|Y_H Z_{H+1}^\top - X_*\|_2 \leq \epsilon$, where $H = O(\log(\|X_*\|_F/\epsilon))$.

The incoherence conditions are:

C1. $\max_{i\in[n]} \|\mathbf{a}_i\|_2^2 \leq \frac{\mu d}{n}, \ \max_{j\in[n]} \|\boldsymbol{b}_j\|_2^2 \leq \frac{\mu d}{n}$

C2'. $\max_{i\in[n]} \|Y_h^\top \mathbf{a}_i\|_2^2 \leq \frac{\mu_0 k}{n}, \ \max_{j\in[n]} \|Z_h^\top \boldsymbol{b}_j\|_2^2 \leq \frac{\mu_0 k}{n}, \forall h = 1, 2, \cdots, H$

K. Zhong, P. Jain, I. S. Dhillon. *Efficient Matrix Sensing Using Rank-1 Gaussian Measurements*. In ALT (2015).

Sample Complexity for Recovery Guarantees

- Sample complexity of Inductive Matrix Completion (IMC) and Matrix Completion (MC).

methods	IMC	MC
Nuclear-norm	$\tilde{O}(kd)$	$\tilde{O}(kn)$ (Recht &Cands, 2009)
ALS	$\tilde{O}(k^4\beta^2 d)$	$\tilde{O}(k^3\beta^2 n)$ (Hardt, 2014)

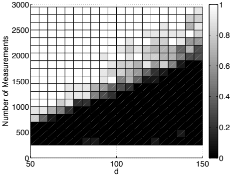
where β is the condition number of X

- In most cases, $n \gg d$
- Incoherence conditions on A, B are required
  - Satisfied e.g. when A, B are Gaussian (no assumption on X needed)

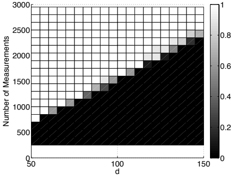
K. Zhong, P. Jain, I. S. Dhillon. *Efficient Matrix Sensing Using Rank-1 Gaussian Measurements*. In ALT (2015).
E. J. Cands and B. Recht. *Exact matrix completion via convex optimization.* Foundations of Computational mathematics (2009).
M. Hardt. *Understanding alternating minimization for matrix completion.* Foundations of Computer Science (2014).

Sample Complexity Results

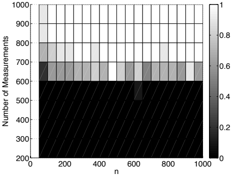
- All matrices are sampled from Gaussian random distribution.
- Left two figures: fix $k = 5$, $n = 1000$ and change d.
- Right two figures: fix $k = 5$, $d = 50$ and change n.
- Darkness of the shading is proportional to the number of failures (repeated 10 times).



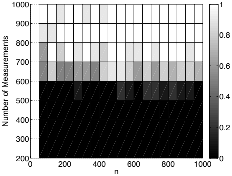
$|\Omega|$ vs. d (ALS)



$|\Omega|$ vs. d (Nuclear)



$|\Omega|$ vs. n (ALS)



$|\Omega|$ vs. n (Nuclear)

- Sample complexity is proportional to d while almost independent of n for both Nuclear-norm and ALS methods.

Modern Challenges in Multi-Target Prediction

- Millions of correlated targets, and missing target values
  - Low-rank + Alternating Least Squares
- Targets have features
  - Bilinear Prediction: Inductive Matrix Completion (IMC)
- Noisy Features
- Positive-unlabeled (PU) target values
- Non-linear Structure

Modern Challenges in Multi-Target Prediction

- Millions of correlated targets, and missing target values
  - Low-rank + Alternating Least Squares
- Targets have features
  - Bilinear Prediction: Inductive Matrix Completion (IMC)
- Noisy Features
  - Dirty IMC
- Positive-unlabeled (PU) target values
  - PU learning for IMC
- Non-linear Structure
  - Deep Learning for IMC

Bilinear Prediction with Noisy Features

Noisy Features

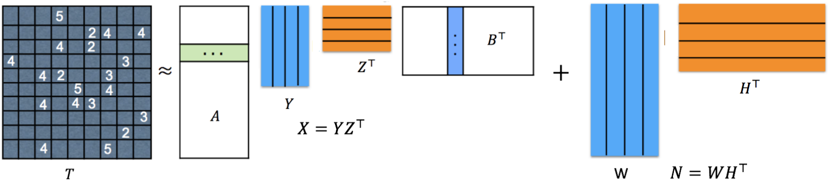
- IMC implicitly assumes that features are good, i.e.,

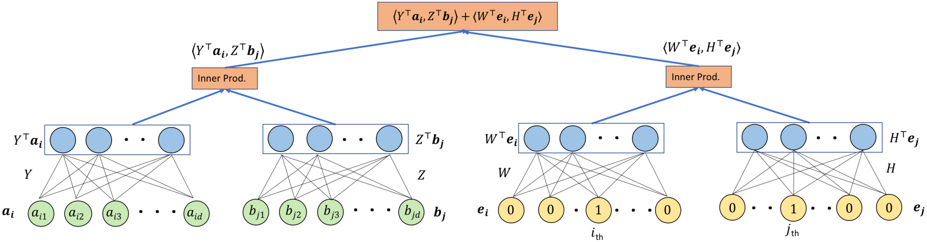
$$\mathrm{col}(T) \subseteq \mathrm{col}(A) \quad \text{and} \quad \mathrm{row}(T) \subseteq \mathrm{col}(B)$$

- When features are not good, learn underlying matrix T jointly from two parts:
  1. Feature-covered part: $AXB^\top$.
  2. Residual part: N.

     $\Rightarrow$ Estimate T as $AXB^\top + N$. (DirtyIMC)
- Both X and N are preferred to be low-rank.
  1. $X = YZ^\top$, where $Y \in \mathbb{R}^{d\times k}, Z \in \mathbb{R}^{d\times k}$
  2. $N = WH^\top$, where $W \in \mathbb{R}^{n\times k}, H \in \mathbb{R}^{n\times k}$

Dirty Inductive Matrix Completion





Dirty Inductive Matrix Completion

- Algorithm 1: Nuclear-norm constraint objective

$$\min_{X,N} \sum_{(i,j)\in\Omega} ((\boldsymbol{a}_i^\top X \boldsymbol{b}_j + N_{ij}) - T_{ij})^2$$
$$\text{s.t.} \|X\|_* \leq \mathcal{X}, \|N\|_* \leq \mathcal{N}$$

- Algorithm 2: Alternating Least Squares (ALS)

$$\min_{Y,Z,W,H} \sum_{(i,j)\in\Omega} ((\mathbf{a}_i^\top YZ^\top \boldsymbol{b}_j + \boldsymbol{e}_i^\top WH^\top \boldsymbol{e}_j) - T_{ij})^2$$
$$\text{s.t. } Y \in \mathbb{R}^{d\times k}, Z \in \mathbb{R}^{d\times k}, W \in \mathbb{R}^{n\times k}, H \in \mathbb{R}^{n\times k}$$

Measuring Quality of Features

- Intuition: what is the meaning of good features?
  - T lies mostly in the space spanned by features.
- A formal measurement:
  - Define linear projection onto col(A) and col(B):

$$\bar{X} = \arg\min_X \|AXB^\top - T\|_F^2,$$

  - The trace norm of residual is used for measuring quality of features.

$$\mathcal{N} = \|T - A\bar{X}B^\top\|_*$$

  - Smaller $\mathcal{N}$ implies a better (linear) feature set.

Error Bound for Dirty IMC

Theorem

Consider nuclear-norm objective with $\mathcal{X} = \|\bar{X}\|_*$ *and* $\mathcal{N} = \|T - A\bar{X}B^\top\|_*$. *Then with probability at least* $1-\delta$, *the optimal solution* $(\widehat{N}, \widehat{X})$ *satisfies:*

$$\frac{1}{n^2}\|\widehat{N} + A\widehat{X}B^\top - T\|_F^2 \leq \min\left\{L_\ell \mathcal{N}\sqrt{\frac{\log n}{|\Omega|}}, \sqrt{CL_\ell \mathcal{B}\frac{\mathcal{N}\sqrt{n}}{|\Omega|}}\right\} + \frac{L_\ell d^2}{\gamma^2}\sqrt{\frac{\log n}{|\Omega|}} + \mathcal{B}\sqrt{\frac{\log \frac{1}{\delta}}{|\Omega|}},$$

where $\mathcal{B}, C, \gamma, L_\ell$ *are all numerical constants.*

Error Bound for Dirty IMC

What does this error bound mean?

- The feature quality measure, $\mathcal{N} := \|T - A\bar{X}B^\top\|_*$

  Case 1: when features are perfect, $\mathcal{N} = 0$

  Case 2: when features contain no information, $\mathcal{N} = O(n)$

  Case 3: when features are noisy but still informative, $\mathcal{N} = o(n)$
- Consider Case 3, i.e., $\mathcal{N} = o(n)$,
  1. if use IMC,

  $$\lim_{n\to\infty} \frac{1}{n^2}\|A\widehat{X}B^\top - T\|_F^2 \not\to 0$$

  2. if use DirtyIMC,

  $$\lim_{n\to\infty} \frac{1}{n^2}\|\widehat{N} + A\widehat{X}B^\top - T\|_F^2 \to 0,$$

  as long as $|\Omega| = o(n^{1.5})$

Applications: Sign Prediction

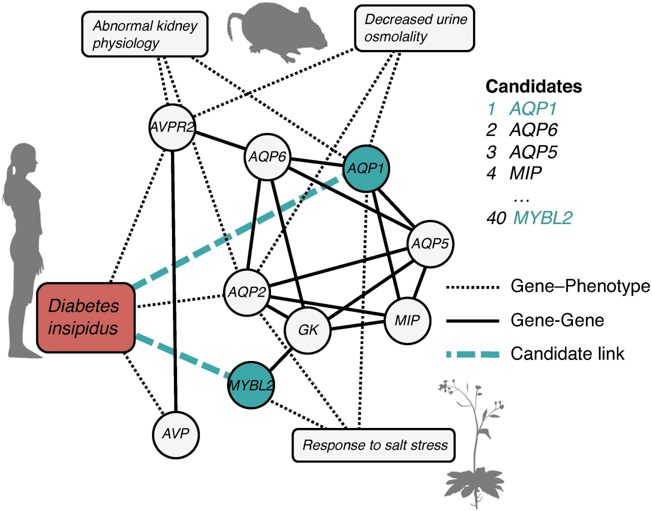
- Sign prediction task on a review sharing site Epinions ($n \approx 105K, |\Omega| \approx 807K$), where people can trust or distrust others.
- Low-rank MC and IMC yield state-of-the-art results on these problems.
- We also collect features $\boldsymbol{a}_i \in \mathbb{R}^{41}$ for each person i from review history.
- Thus, we can replace MC with DirtyIMC by encoding side information.
- DirtyIMC performs the best in terms of both accuracy and AUC.

Method	Accuracy	AUC
DirtyIMC	**0.9474**$\pm$0.0009	**0.9506**
MC	0.9412$\pm$0.0011	0.9020
IMC	0.9139$\pm$0.0016	0.9109

Positive-Unlabeled Learning

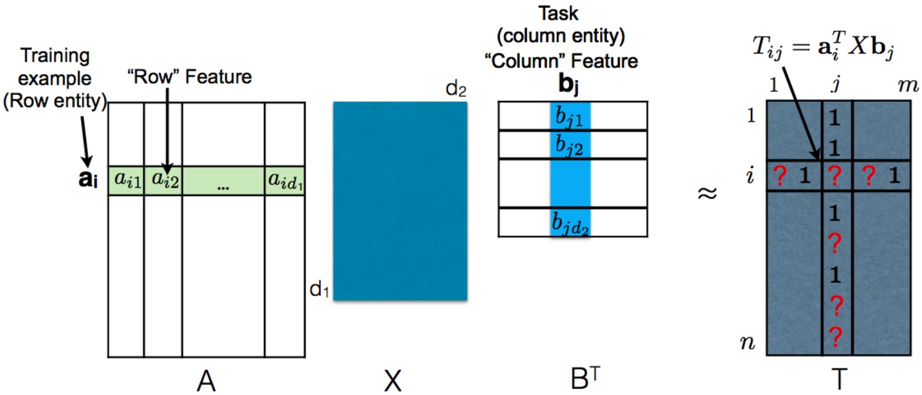
Modern Prediction Problems

Predicting related disease genes



Bilinear Prediction: PU Learning

In many applications, only "positive" labels are observed



PU Learning

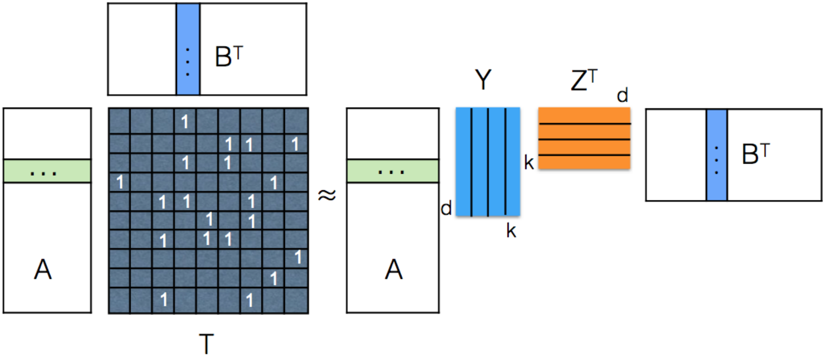
Learning Task	"Positives"	"Negatives"	"Unlabeled"
Supervised	✓	✓	
Semi-supervised	✓	✓	✓
Positive-Unlabeled (PU)	✓		✓
Unsupervised			✓

- No observations of the "negative" class available

$(X, Y) \sim D$ Training data

PU Inductive Matrix Completion

- Guarantees so far assume observations are sampled uniformly
- What can we say about the case when observations are all 1's ("positives")?
- Typically, 99% entries are missing ("unlabeled")



PU Inductive Matrix Completion

- Inductive Matrix Completion:

$$\min_{\|X\|_* \leq \mathcal{X}} \sum_{(i,j)\in\Omega} (\mathbf{a}_i^\top X \mathbf{b}_j - T_{ij})^2$$

- Commonly used PU strategy: Biased Matrix Completion

$$\min_{\|X\|_* \leq \mathcal{X}} \alpha \sum_{(i,j)\in\Omega} (\mathbf{a}_i^\top X \mathbf{b}_j - T_{ij})^2 + (1-\alpha) \sum_{(i,j)\notin\Omega} (\mathbf{a}_i^\top X \mathbf{b}_j - 0)^2$$

Typically, $\alpha > 1 - \alpha$ ($\alpha \approx 0.9$).

V. Sindhwani, S. S. Bucak, J. Hu, A. Mojsilovic. *One-class matrix completion with low-density factorizations*. ICDM, pp. 1055-1060. 2010.

PU Inductive Matrix Completion

- Inductive Matrix Completion:

$$\min_{\|X\|_* \leq \mathcal{X}} \sum_{(i,j)\in\Omega} (\mathbf{a}_i^\top X \mathbf{b}_j - T_{ij})^2$$

- Commonly used PU strategy: Biased Matrix Completion

$$\min_{\|X\|_* \leq \mathcal{X}} \alpha \sum_{(i,j)\in\Omega} (\mathbf{a}_i^\top X \mathbf{b}_j - T_{ij})^2 + (1-\alpha) \sum_{(i,j)\notin\Omega} (\mathbf{a}_i^\top X \mathbf{b}_j - 0)^2$$

Typically, $\alpha > 1 - \alpha$ ($\alpha \approx 0.9$).

- We can show theoretical guarantees for the biased formulation

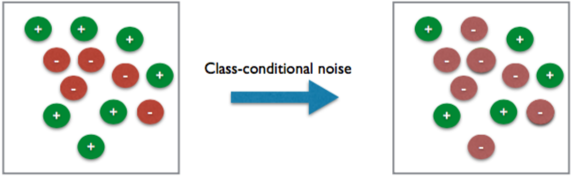
V. Sindhwani, S. S. Bucak, J. Hu, A. Mojsilovic. *One-class matrix completion with low-density factorizations*. ICDM, pp. 1055-1060. 2010.

PU Learning: Random Noise Model

- Can be formulated as learning with "class-conditional" noise

$$P(\tilde{Y} = -1 | Y = +1) = \rho_{+1}$$
$$P(\tilde{Y} = +1 | Y = -1) = \rho_{-1}$$

Becomes PU learning when $\rho_{-1} = 0$



$(X, Y) \sim D$

$(X, \tilde{Y}) \sim D_\rho$

Noisy training data

N. Natarajan, I. S. Dhillon, P. Ravikumar, and A.Tewari. *Learning with Noisy Labels*. In Advances in Neural Information Processing Systems, pp. 1196-1204. 2013.

PU Inductive Matrix Completion

A deterministic PU learning model

M

0.2	0.1	0	0.8
0	0.6	0.1	0.9
0	0	0.8	0.1
0.9	0	0.2	0.1
0	0.6	0	1

T

0	0	0	1
0	1	0	1
0	0	1	0
1	0	0	0
0	1	0	1

$$T_{ij} = \begin{cases} 1 & \text{if } M_{ij} > 0.5, \\ 0 & \text{if } M_{ij} \leq 0.5 \end{cases}$$

PU Inductive Matrix Completion

A deterministic PU learning model

M

0.2	0.1	0	0.8
0	0.6	0.1	0.9
0	0	0.8	0.1
0.9	0	0.2	0.1
0	0.6	0	1

T

0	0	0	1
0	1	0	1
0	0	1	0
1	0	0	0
0	1	0	1

$\tilde{T}$

?	?	?	1
?	1	?	?
?	?	1	?
1	?	?	?
?	?	?	1

- $P(\tilde{T}_{ij} = 0 | T_{ij} = 1) = \rho$ and $P(\tilde{T}_{ij} = 0 | T_{ij} = 0) = 1$.
- We are given *only* $\tilde{T}$ but *not* T or M
- Goal: Recover T given $\tilde{T}$ (recovering M is not possible!)

PU Inductive Matrix Completion

Theorem (Error Bound for PU IMC)

Assume ground-truth X satisfies $\|X\|_* \leq \mathcal{X}$ (where $M = AXB^\top$). Define $\widehat{T}_{ij} = I\big[(A\widehat{X}B^\top)_{ij} > 0.5\big]$, $\mathcal{A} = \max_i \|\mathbf{a}_i\|$ and $\mathcal{B} = \max_i \|\mathbf{b}_i\|$. If $\alpha = \frac{1+\rho}{2}$, then with probability at least $1-\delta$,

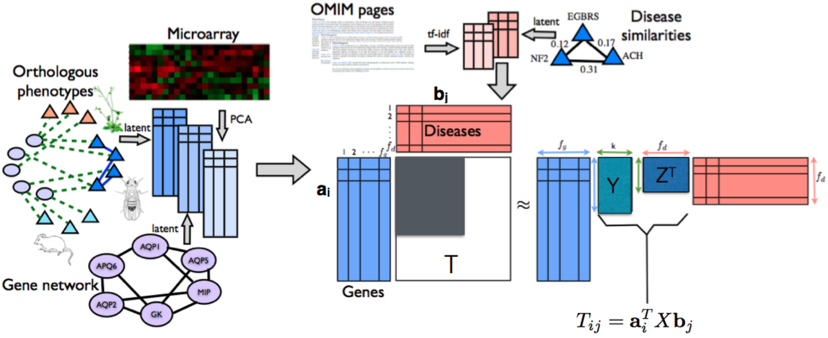
$$\frac{1}{n^2}\|T - \widehat{T}\|_F^2 = O\left(\frac{\sqrt{\log(2/\delta)}}{n(1-\rho)} + \frac{\mathcal{X}\mathcal{A}\mathcal{B}\sqrt{log 2d}}{(1-\rho)n^{3/2}}\right)$$

- In other words, as long as $1-\rho = O(\frac{\log n}{n})$,

$$\lim_{n\to\infty} \frac{1}{n^2}\|T - \widehat{T}\|_F^2 \to 0$$

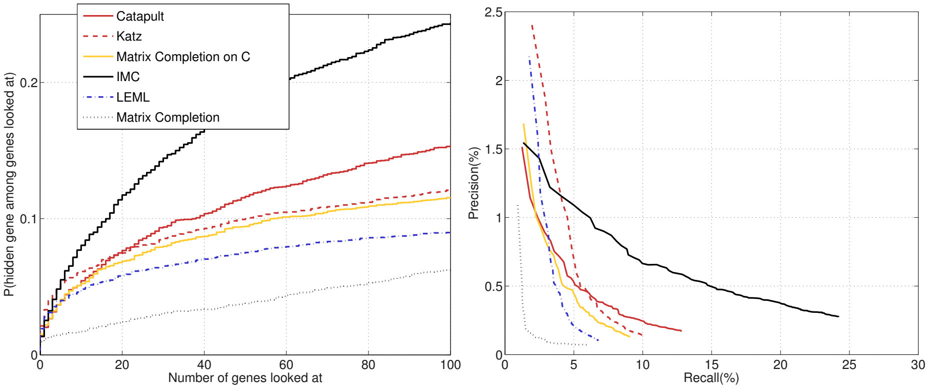
C-J. Hsieh, N. Natarajan, and I. S. Dhillon. *PU Learning for Matrix Completion*. In Proceedings of The 32nd International Conference on Machine Learning, pp. 2445-2453 (2015).

PU Inductive Matrix Completion: Gene-Disease Prediction



N. Natarajan, and I. S. Dhillon. *Inductive matrix completion for predicting gene disease associations*. Bioinformatics, 30(12), i60-i68 (2014).

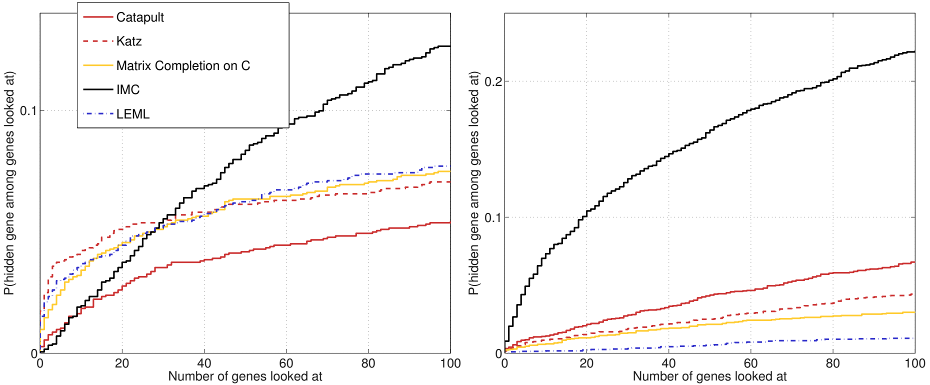
PU Inductive Matrix Completion: Gene-Disease Prediction



Predicting gene-disease associations in the OMIM data set (`www.omim.org`).

N. Natarajan, and I. S. Dhillon. *Inductive matrix completion for predicting gene disease associations*. Bioinformatics, 30(12), i60-i68 (2014).

PU Inductive Matrix Completion: Gene-Disease Prediction



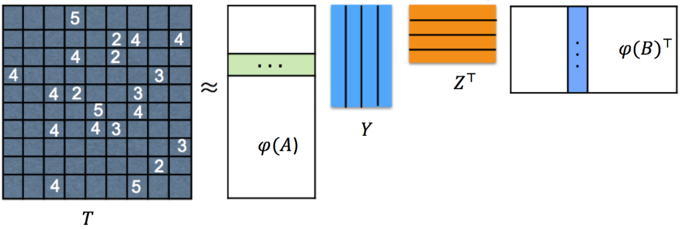
Predicting genes for diseases with *no* training associations.

N. Natarajan, and I. S. Dhillon. *Inductive matrix completion for predicting gene disease associations*. Bioinformatics, 30(12), i60-i68 (2014).

Nonlinear Prediction

Goal-Directed IMC: Two-layer Neural Network

- Key ideas for Goal-Directed IMC (GIMC):
  - Non-linear feature mapping.
    - $A \to \varphi(A)$ using non-linear mapping.
  - Learn the model and non-linear mapping simultaneously.
    - learn model and features jointly in a framework.
    - alternating minimization.

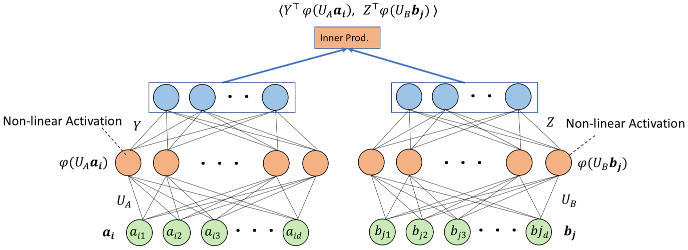


Goal-Directed IMC: Two-layer Neural Network

- GIMC framework:

$$\min_{Y,Z,U_A,U_B} \sum_{(i,j)\in\Omega} (T_{ij} - (\varphi_{U_A}(A) YZ^\top \varphi_{U_B}(B)^\top)_{ij})^2 + \lambda(\|Y\|_F^2 + \|Z\|_F^2).$$

- U_A and U_B are parameters for the non-linear feature mapping.
- Y, Z are the model.



S. Si, K.-Y. Chiang, C.-J. Hsieh, N. Rao, and I.S.Dhillon. *Goal-Directed Inductive Matrix Completion.* In KDD, 2016.

Experimental Results: Multi-Label Learning

- Multi-label learning:
  - LEML[Yu et al. 2014]: an embedding based technique.
  - FastXML[Prabhu and Varma, 2014]: a random forest approach.
  - SLEEC[Bhatia et al. 2015]: an ensemble of local distance preserving embeddings.
- Results (Precision @ 1, 3, 5):

	Delicious			NUS-WIDE			Delicious-large		
	P1 (%)	P3 (%)	P5 (%)	P1 (%)	P3 (%)	P5 (%)	P1 (%)	P3 (%)	P5 (%)
GIMC	**71.40**	**65.16**	**59.79**	**22.49**	**17.40**	**14.70**	46.13	40.32	38.15
SLEEC	68.38	61.50	56.35	17.67	14.20	12.07	**47.03**	**41.67**	**38.88**
FastXML	69.65	63.93	59.36	21.00	16.32	13.66	42.81	38.76	36.34
LEML	65.66	61.15	56.08	20.76	16.00	13.11	40.30	37.76	36.66

- On Delicious and NUS-WIDE, GIMC performs the best.
- On Delicious-large dataset, GIMC has similar accuracy with SLEEC, but takes much less time: 4,724 vs 25,289 seconds.

Experimental Results: Semi-Supervised Clustering

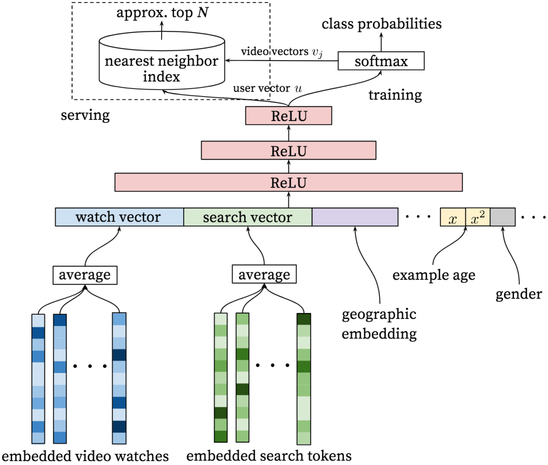
- Goal: find a clustering of n items given:
  - feature matrix $A \in \mathbb{R}^{n \times d}$.
  - pairwise constraints $\{T_{ij} \mid (i,j) \in \Omega\}$ describing similar/dissimilar pairs.
- Earlier state-of-the-art: MCCC algorithm [Yi et. al, ICML 2013]:
  1. Learn a low rank similarity matrix S by conducting IMC on Ω.
  2. Do k-means clustering on top-k eigenvectors of S.
- Thus, we can replace IMC step in MCCC with GIMC.
- Results on clustering error rate:

Dataset	n	d	k	# constraints $\lvert\Omega\rvert$	k-means	MCCC	IMC	GIMC
Segment	2319	19	7	n	0.1433	0.0891	**0.0683**	0.0724
				$5n$	0.1347	0.0800	0.0580	**0.0570**
				$10n$	0.1363	0.0809	0.0650	**0.0446**
				$15n$	0.1362	0.0872	0.0678	**0.0402**
				$20n$	0.1330	0.0837	0.0564	**0.0380**
Covtype-sub	1711	54	7	n	0.2523	0.2498	**0.1840**	0.1898
				$5n$	0.2112	0.1772	0.1930	**0.1592**
				$10n$	0.2068	0.1708	0.1722	**0.1388**
				$15n$	0.2203	0.1677	0.1687	**0.1262**
				$20n$	0.2124	0.1607	0.1561	**0.1078**

Deep Neural Networks

YouTube recommendation

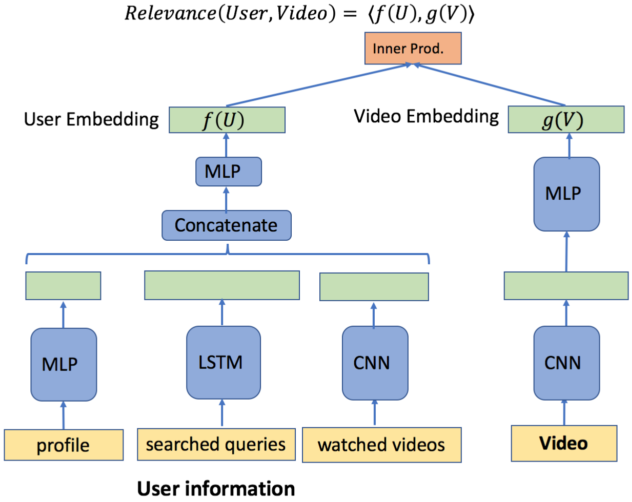
Recall model architecture:



P. Covington, J. Adams, and E. Sargin. *Deep neural networks for youtube recommendations.* Proceedings of the 10th ACM Conference on Recommender Systems. ACM, 2016.

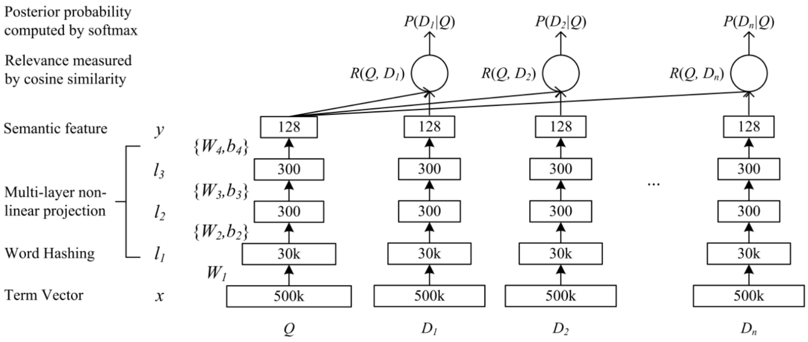
Multi-Target Prediction with Deep Neural Networks

Video Recommendation



Multi-Target Prediction with Deep Neural Networks

Web Search (DSSM Model)



P. S. Huang, X. He, J. Gao, L. Deng, A. Acero, & L. Heck, *Learning deep structured semantic models for web search using clickthrough data.* In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management (2013)

DSSM Model

Web Search

- Training data: 100 million query-document pairs with rich click information sampled from one-year query log files of a commercial search engine.
- Evaluation Data: 16k queries, each with about 15 documents whose relevance scores are labeled as 0-4 by human.

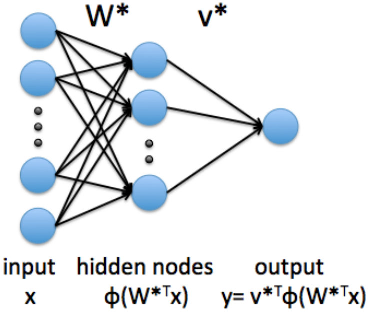
#	Models	NDCG@1	NDCG@3	NDCG@10
1	TF-IDF	0.319	0.382	0.462
2	BM25	0.308	0.373	0.455
3	WTM	0.332	0.400	0.478
4	LSA	0.298	0.372	0.455
5	PLSA	0.295	0.371	0.456
6	DAE	0.310	0.377	0.459
7	BLTM-PR	0.337	0.403	0.480
8	DPM	0.329	0.401	0.479
9	DNN	0.342	0.410	0.486
10	L-WH linear	0.357	0.422	0.495
11	L-WH non-linear	0.357	0.421	0.494
12	**L-WH DNN**	**0.362**	**0.425**	**0.498**

Table 2: Comparative results with the previous state of the art approaches and various settings of DSSM.

P. S. Huang, X. He, J. Gao, L. Deng, A. Acero, & L. Heck, *Learning deep structured semantic models for web search using clickthrough data.* In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management (2013)

Neural Networks: Theoretical Guarantees

- Small steps: we have shown recovery guarantees for single-target regression problems with one-hidden-layer neural network.



K. Zhong, Z. Song, P. Jain, P. L. Bartlett & I. S. Dhillon. *Recovery Guarantees for One-hidden-layer Neural Networks.* ICML 2017

Neural Networks: Theoretical Guarantees?

- Objective function

$$\widehat{f}_S(W) = \frac{1}{2n} \sum_{j \in [n]} \left(\sum_{i=1}^{k} v_i^* \varphi(\boldsymbol{w}_i^\top \boldsymbol{x}_j) - y_j \right)^2 .$$

Neural Networks: Theoretical Guarantees?

- Objective function

$$\widehat{f}_S(W) = \frac{1}{2n} \sum_{j \in [n]} \left(\sum_{i=1}^{k} v_i^* \varphi(\boldsymbol{w}_i^\top \boldsymbol{x}_j) - y_j \right)^2 .$$

- We show local strongly convexity near the ground truth.
- Standard gradient descent will converge linearly to the ground truth when initialized appropriately.
- Future work: extend to multi-target with missing values.

K. Zhong, Z. Song, P. Jain, P. L. Bartlett & I. S. Dhillon. *Recovery Guarantees for One-hidden-layer Neural Networks.* ICML 2017

Summary

- Millions of correlated targets, and missing target values
  - Low-rank + Alternating Least Squares
- Targets have features
  - Bilinear Prediction: Inductive Matrix Completion (IMC)
- Noisy Features
  - Dirty IMC
- Positive-unlabeled (PU) target values
  - PU learning for IMC
- Non-linear Structure
  - Deep Learning for IMC

Summary

- Inductive Matrix Completion:
  - Scales to millions of targets
  - Captures correlations among targets
  - Overcomes missing target values
  - Handles noisy features and non-linear features
  - Extends to PU learning

Future Work

- Much work to do:
  - Other structures: low-rank+sparse, low-rank+column-sparse (outliers)?
  - Different loss functions?
  - Handling "time" as one of the dimensions — incorporating smoothness through graph regularization?
  - Efficient (parallel) implementations?
  - Guarantees for deep inductive matrix completion?

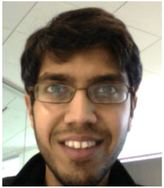
Collaborators

Kai-Yang Chiang

Cho-Jui Hsieh

Prateek Jain

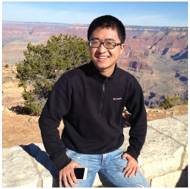
Nagarajan Natarajan



Nikhil Rao

Si Si

Hsiang-fu Yu



Kai Zhong

References

[1] P. Jain, and I. S. Dhillon. *Provable inductive matrix completion*. In arXiv preprint arXiv:1306.0626 (2013).

[2] K. Zhong, P. Jain, I. S. Dhillon. *Efficient Matrix Sensing Using Rank-1 Gaussian Measurements*. In ALT (2015).

[3] N. Natarajan, and I. S. Dhillon. *Inductive matrix completion for predicting gene disease associations*. In Bioinformatics, 30(12), i60-i68 (2014).

[4] H. F. Yu, P. Jain, P. Kar, and I. S. Dhillon. *Large-scale Multi-label Learning with Missing Labels*. In ICML (2014).

[5] C-J. Hsieh, N. Natarajan, and I. S. Dhillon. *PU Learning for Matrix Completion*. In ICML (2015).

[6] S. Si, K.-Y. Chiang, C.-J. Hsieh, N. Rao, and I.S.Dhillon *Goal-Directed Inductive Matrix Completion* In KDD, 2016.

[7] K.-Y. Chiang, C.-J. Hsieh and I. S. Dhillon. *Matrix Completion with Noisy Side Information* In NIPS (2015).

[8] K. Zhong, Z. Song, P. Jain, P. L. Bartlett & I. S. Dhillon. *Recovery Guarantees for One-hidden-layer Neural Networks.* In ICML (2017)