HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

Kernelized Bayesian Matrix Factorization (KBMF)

Mehmet Gönen[1], Muhammad Ammad-ud-din[1], Suleiman A. Khan[1], Samuel Kaski[1,2]

[1] Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University, Finland

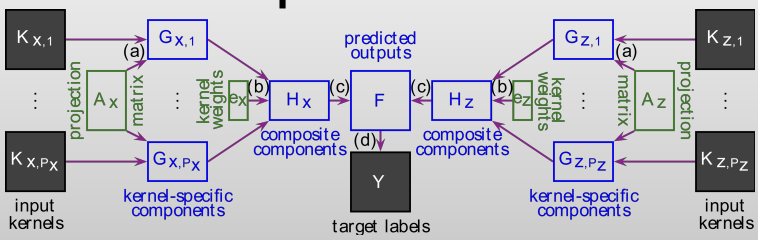
[2] Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Finland

http://research.ics.aalto.fi/mi/

Abstract

1. We extend **kernelized matrix factorization**
   - **with a fully Bayesian treatment,**
   - **with an ability to work with multiple side information sources.**
2. Side information is necessary for making out-of-matrix predictions (e.g., cold-start predictions in recommender systems).
3. We mainly discuss bipartite graph inference, where the output matrix is binary.
4. We show the performance of our method
   - **by predicting drug—protein interactions on two data sets.**

Proposed Method



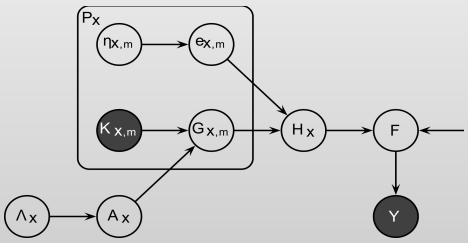
(a) Kernel based non-linear dimensionality reduction

(b) Multiple kernel learning

(c) Matrix factorization

(d) Binary classification (if data is binary)

Probabilistic Model



(a) kernel-based nonlinear dimensionality reduction

$$\lambda^i_{\mathbf{x},s} \sim \mathcal{G}(\lambda^i_{\mathbf{x},s}; \alpha_\lambda, \beta_\lambda) \qquad \forall(i,s)$$
$$a^i_{\mathbf{x},s} | \lambda^i_{\mathbf{x},s} \sim \mathcal{N}(a^i_{\mathbf{x},s}; 0, (\lambda^i_{\mathbf{x},s})^{-1}) \qquad \forall(i,s)$$
$$g^s_{\mathbf{x},m,i} | \boldsymbol{a}_{\mathbf{x},s}, \boldsymbol{k}_{\mathbf{x},m,i} \sim \mathcal{N}(g^s_{\mathbf{x},m,i}; \boldsymbol{a}^\top_{\mathbf{x},s} \boldsymbol{k}_{\mathbf{x},m,i}, \sigma_g^2) \qquad \forall(m,s,i)$$

(b) multiple kernel learning

$$\eta_{\mathbf{x},m} \sim \mathcal{G}(\eta_{\mathbf{x},m}; \alpha_\eta, \beta_\eta) \qquad \forall m$$
$$e_{\mathbf{x},m} | \eta_{\mathbf{x},m} \sim \mathcal{N}(e_{\mathbf{x},m}; 0, \eta^{-1}_{\mathbf{x},m}) \qquad \forall m$$
$$h^s_{\mathbf{x},i} | \{e_{\mathbf{x},m}, g^s_{\mathbf{x},m,i}\}_{m=1}^{P_{\mathbf{x}}} \sim \mathcal{N}\left(h^s_{\mathbf{x},i}; \sum_{m=1}^{P_{\mathbf{x}}} e_{\mathbf{x},m} g^s_{\mathbf{x},m,i}, \sigma_h^2\right) \qquad \forall(s,i)$$

(c) matrix factorization

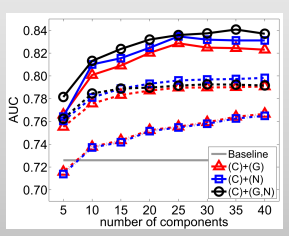
$$f^i_j | \boldsymbol{h}_{\mathbf{x},i}, \boldsymbol{h}_{\mathbf{z},j} \sim \mathcal{N}(f^i_j; \boldsymbol{h}^\top_{\mathbf{x},i} \boldsymbol{h}_{\mathbf{z},j}, 1) \qquad \forall(i,j)$$

(d) binary classification

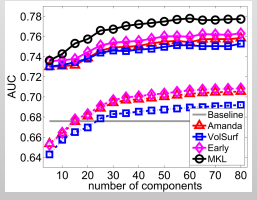
$$y^i_j | f^i_j \sim \delta(f^i_j y^i_j > \nu) \qquad \forall(i,j)$$

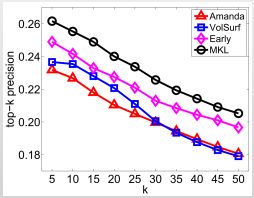
Drug–Protein Interaction Data Sets

1. A drug–protein network by Yamanishi et al. (2008)
   - 445 drugs, 664 proteins and 2926 validated interactions
   - C: chemical similarity for drugs
   - G: genomic similarity for proteins
   - N: network similarity for proteins
2. 5 replication of 5-fold CV over drugs



1. Another drug–protein interaction network by Khan et al. (2012)
   - 855 drugs, 800 proteins, and 4659 validated Interactions
2. Two standard 3D chemical structure descriptors for drugs:
   - Amanda (Duran et al., 2008) and VolSurf (Cruciani et al., 2000)
3. Gaussian kernel whose width is selected as √D
4. 5 replications of 5-fold cross validation over drugs
5. An extra task of finding or retrieving drugs with similar functions





KBMF is statistically significantly better than KPMF of Zhou et al. (2012) according to paired t-test ($p < 0.01$) on both data sets.

(1) M. Gönen, S. A. Khan, and S. Kaski. Kerelized Bayesian Matrix Factorization. In Proceedings of ICML 2013, the 30th International Conference on Machine Learning, volume 28 of JMLR W&CP, pages 864–872. JMLR, 2013. Implementations in Matlab are available at http://research.ics.aalto.fi/mi/ (2) S. A. Khan, A. Faisal, J. P. Mpindi, J. A. Parkkinen, T. Kalliokoski, A. Poso, O. P. Kallioniemi, K. Wennerberg, and S. Kaski. Comprehensive data-driven analysis of the impact of chemoinformatic structure on the genome-wide biological response profiles of cancer cells to 1159 drugs.BMC Bioinformatics, 13(112), 2012.