The Infinite Latent Attribute Model

Konstantina Palla[1] David A. Knowles Zoubin Ghahramani
University of Cambridge University of Cambridge University of Cambridge

[1]kp376@cam.ac.uk

UNIVERSITY OF CAMBRIDGE

Abstract

Latent variable models for network data extract a summary of the relational structure underlying an observed network. The simplest possible models subdivide nodes into clusters; the probability of a link between any two nodes then depends only on their cluster assignment. Currently available models can be classified by whether clusters are disjoint or are allowed to overlap. These models can explain a "flat" clustering structure. We propose a model in which entities are characterised by a latent feature vector. Each feature is itself partitioned into disjoint groups (subclusters), corresponding to a second layer of hierarchy. In experimental comparisons, the model achieves significantly improved predictive performance on social and biological link prediction tasks.

Latent Class & Latent Feature Models

Latent Class models assume a number of clusters K and each entity belongs to a *single* cluster. The link probability between two entities depends only on their cluster assignments. The Infinite Relational Model (IRM) [3] belongs to this category. In the IRM:

- The cluster assignments $c_i = k$, $k =\in \{1, 2, ..., K\}$ are drawn from the Chinese Restaurant Process (CRP).
- A $K \times K$ weight matrix $\mathbf{W}$ contains the link probability between each pair of clusters.
- To generate a link $Y(i, j) \in \{0, 1\}$, draw

$$Y(i, j) \sim \text{Bernoulli}(W(c_i, c_j))$$

Latent Feature models relate each entity with a vector of M features and determine the link probability based on feature interactions. Such a model is the Nonparametric Latent Feature Relational Model (NLFRM) [4]:

- each entity i is assigned a binary feature vector $\mathbf{z}_i$. The $N \times M$ latent feature matrix, $\mathbf{Z}$ is drawn from the Indian Buffet Process.
- a $M \times M$ weight matrix $\mathbf{W}$ contains the real valued weights between each pair of features.
- to generate a link $Y(i, j) \in \{0, 1\}$, draw

$$Y(i, j) \sim \text{Bernoulli}\big(\sigma(\mathbf{z}_i^T \mathbf{W} \mathbf{z}_j)\big)$$

Motivation

An example: a friendship network at a collegiate University. A person might belong to more than one cluster e.g. a *college*, a *department* and a *sport team*. A latent class model would need a new cluster for each combination of the types of cluster, e.g. *'Gryffindor College, Department of Mathematics, Football'*. A latent feature model uses the feature vector representation to implicitly account for the possible combination of clusters.

Motivation existing models only account for a flat clustering. The 'college' feature might be divided into subclusters, e.g. 'Slytherin College', 'Gryffindor College' etc. A latent feature model must represent each cluster with a new feature.

Infinite Latent Attribute Model

Proposal: allow an explicit representation of the partitioning of each general feature into subclusters.

ILA model: Links are generated as follows:

- every entity is assigned a binary vector $\mathbf{z}_i$ indicating which features it has active. Draw the $N \times M$ latent feature matrix $\mathbf{Z}$ from the Indian Buffet Process.

$$\mathbf{Z}|\alpha \sim \text{IBP}(\alpha)$$

- all the members of the m^{th} feature, are assigned to $K^{(m)}$ subclusters, with each entity belonging to a single subcluster in that feature. $\mathbf{c}^{(m)}$ is a vector of length N and $c_i^{(m)}$ denotes the subcluster the i^{th} entity belongs to in the m^{th} feature.

$$\mathbf{c}^{(m)}|\gamma \sim \text{CRP}(\gamma)$$

- Each feature m has a real-valued $K^{(m)} \times K^{(m)}$ weight matrix $\mathbf{W}^{(m)}$. $w_{kk'}^{(m)} \equiv W^{(m)}(k, k')$ is the weight that affects the probability of there being a link from entity i to entity j, given that entity i belongs to subcluster k and entity j belongs to subcluster k' of the m^{th} feature.

$$w_{kk'}^{(m)}|\sigma_w \sim N(0, \sigma_w^2)$$

- to generate a link $Y(i, j) \in \{0, 1\}$, from entity i to entity j, draw

$$Y_{ij}|\mathbf{Z}, \mathbf{C}, \mathbf{W} \sim \text{Bernoulli}\left(\sigma\left(\sum_m z_{im} z_{jm} w_{c_i^m c_j^m}^{(m)} + s\right)\right).$$

where s is a bias parameter. Only classes that are on for both entities influence the probability of a link between them.

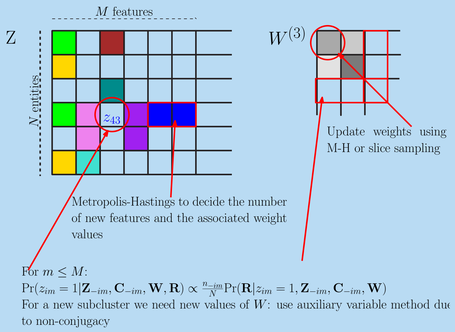
Inference

Sampling Z:

- for $m \leq M$: Bernoulli sample. Integrate over $c_i^{(m)}$, including case of new subcluster. $P(w)$ non-conjugate to likelihood so use auxiliary variable approach [5] (Algorithm 8)
- for $m > M$: sample the number of new features and the associated weights. Due to non-conjugacy, use Metropolis-Hastings.

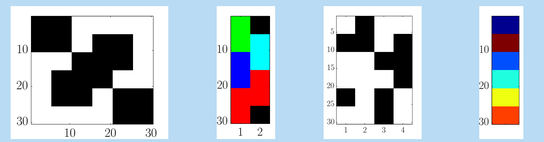
Sampling C: included in sampling **Z**.

Sampling W: Non-conjugacate so use Metropolis-Hastings or slice sampling.



Results

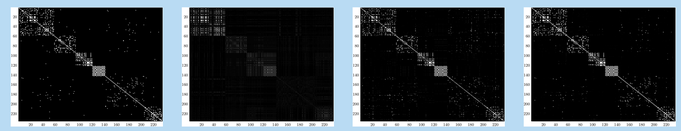
Synthetic data



NIPS Coauthorship network

We used the NIPS 1-17 coauthorship dataset [1]. We kept only the 234 most connected authors, ran 10 repeats, holding out 20% of the data.

	IRM	LFRM	ILA ($M = 6$)	ILA ($M = \infty$)
Test error (0-1 loss)	0.0440 ± 0.0014	0.0228 ± 0.0041	0.0141 ± 0.0012	$\mathbf{0.0106 \pm 0.0007}$
Test log likelihood	-0.0859 ± 0.0043	-0.0547 ± 0.0079	$\mathbf{-0.0322 \pm 0.0058}$	$\mathbf{-0.0318 \pm 0.0094}$
AUC	0.9565 ± 0.0037	0.9631 ± 0.0150	$\mathbf{0.9908 \pm 0.0048}$	$\mathbf{0.9910 \pm 0.0056}$



Gene Interactions network

We used a subset of the interaction data by [2]. We used 156 genes.

	IRM	LFRM	ILA ($M = 6$)	ILA ($M = \infty$)
Test error (0-1 loss)	0.3608 ± 0.0031	0.2661 ± 0.0086	0.2284 ± 0.0077	$\mathbf{0.0735 \pm 0.0047}$
Test log likelihood	-0.4669 ± 0.0097	-0.4223 ± 0.0147	-0.3596 ± 0.0156	$\mathbf{-0.2654 \pm 0.0447}$
AUC	0.8654 ± 0.0057	0.8471 ± 0.0132	0.9401 ± 0.0046	$\mathbf{0.9924 \pm 0.0037}$

Conclusion

- ILA is able to capture the complex nature of real world networks, with corresponding gains in empirical performance.
- ILA could be made even more flexible by allowing multiple membership of subclusters within a feature, corresponding to a nested IBP.

References

[1] A. Globerson, G. Chechik, F. Pereira, and N. Tishby. Euclidean embedding of co-occurrence data. *The Journal of Machine Learning Research*, 8, 2007.

[2] Martin C. Jonikas, Sean R. Collins, Vladimir Denic, Eugene Oh, Erin M. Quan, Volker Schmid, Jimena Weibezahn, Blanche Schwappach, Peter Walter, Jonathan S. Weissman, and Maya Schuldiner. Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum. *Science*, 323(5922):1693–1697, 2009.

[3] Charles Kemp and Joshua B. Tenenbaum. Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st National Conference on Artificial Intelligence*, 2006.

[4] Kurt Miller, Thomas Griffiths, and Michael Jordan. Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems (NIPS) 22*. 2009.

[5] Radford M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.