PhD Course August 2012 - Advanced Signal Processing (02901) Morning Session Day 5: Mining Graphs by Relational Modeling

Morten Mørup

Cognitive Systems Group, DTU informatics,
e-mail: mm@imm.dtu.dk

Learning objective: *The aim of this exercise is to understand the infinite relational model (IRM). In particular, how the model generates graph and how the model parameters can be inferred and interpreted in terms of identifying structure in graphs.*

The Infinite Relational Model

Generative Model: The generative model for the Infinite Relational Model (IRM) for unipartite graphs is given by

$$\boldsymbol{Z} \sim CRP(\alpha) \tag{1}$$

$$\eta_{lm} \sim Beta(\beta_{lm}^+, \beta_{lm}^-) \tag{2}$$

$$A_{ij} \sim Bernoulli(\boldsymbol{z}_i^\top \boldsymbol{\eta} \boldsymbol{z}_j) \tag{3}$$

Where $\boldsymbol{Z}$ is a binary assignment matrix such that $\boldsymbol{z}_i$ is a vector indicating which cluster the ith node belongs to. Writing down the joint posterior and analytically integrating out $\boldsymbol{\eta}$ we obtain the following joint posterior for the observed graph data $\boldsymbol{A}$ and current estimate of the assignment matrix $\boldsymbol{Z}$

$$
\begin{aligned}
P(\boldsymbol{A}, \boldsymbol{Z} | \alpha, \boldsymbol{\beta}^+, \boldsymbol{\beta}^-) &= \int P(\boldsymbol{A}, \boldsymbol{Z}, \boldsymbol{\eta} | \alpha, \boldsymbol{\beta}^+, \boldsymbol{\beta}^-) \partial \boldsymbol{\eta} \\
&= [\prod_{lm} \frac{\Gamma(\beta_{lm}^+ + \beta_{lm}^-)}{\Gamma(\beta_{lm}^+)\Gamma(\beta_{lm}^-)} \int \eta_{lm}^{N_{lm}^+ + \beta_{lm}^+ - 1}(1 - \eta_{lm})^{N_{lm}^- + \beta_{lm}^- - 1} \partial \eta_{lm}] [\frac{\alpha^K \Gamma(\alpha) \prod_k \Gamma(n_k)}{\Gamma(J + \alpha)}] \\
&= [\prod_{lm} \frac{Beta(N_{lm}^+ + \beta_{lm}^+, N_{lm}^- + \beta_{lm}^-)}{Beta(\beta_{lm}^+, \beta_{lm}^-)}][\frac{\alpha^K \Gamma(\alpha) \prod_k \Gamma(n_k)}{\Gamma(J + \alpha)}]
\end{aligned}
$$

Where n_l gives the number of nodes assigned to group l, $N_{lm}^+ = [\boldsymbol{Z}\boldsymbol{A}\boldsymbol{Z}^\top]_{lm}$ gives the number of links between group l and m whereas $N_{lm}^- = n_k n_l - \delta_{lm} n_k - N_{lm}^+$ gives the number of non-links between group l and m (here for simplicity given for the case of directed graphs). $Beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is the Beta function and $\Gamma(a) = (a-1)!$.

Inference: Making use of Bayes' theorem we obtain the posterior distribution for each node's assignment conditioned on the remaining nodes assignment as

$$P(z_{li} = 1|\boldsymbol{A}, \boldsymbol{Z}_{\setminus i}, \alpha, \boldsymbol{\beta}^+, \boldsymbol{\beta}^-) = \frac{P(\boldsymbol{A}, \boldsymbol{Z}_{\setminus i}, z_{li} = 1|\alpha, \boldsymbol{\beta}^+, \boldsymbol{\beta}^-)}{\sum_{l'} P(\boldsymbol{A}, \boldsymbol{Z}_{\setminus i}, z_{l'i} = 1|\alpha, \boldsymbol{\beta}^+, \boldsymbol{\beta}^-)}$$

By (Gibbs) sampling each node assignment $\boldsymbol{z}_i$ in turn from the above posterior distribution we can infer $\boldsymbol{Z}$. Notice, the expected value of the relations $\boldsymbol{\eta}$ given the node assignments $\boldsymbol{Z}$ is defined by $\langle \boldsymbol{\eta}_{lm} \rangle = \frac{N_{lm}^+ + \beta_{lm}^+}{N_{lm}^+ + \beta_{lm}^+ + N_{lm}^- + \beta_{lm}^-}$. (Apart from Gibbs sampling we will also use so-called split-merge moves to infer the model parameters although these details has been left out, see also [*Kemp et al. AAAI 2006*].)

Mutual Information: When ground truth is available, i.e. we know the parameters used to generate the data or we have some additional information about the clustering structure of the graph, we can evaluate how well the estimated clustering correspond to the true clustering used to generate the data. We will use the normalized mutual information (NMI) to evaluate how closely the assignment matrix (i.e., clustering) $\boldsymbol{Z}^{estimated}$ inferred from the generated graph $\boldsymbol{A}$ is to the true assignment matrix (i.e., clustering) $\boldsymbol{Z}^{true}$ used to generate the graph. The normalized mutual information between $\boldsymbol{Z}^{true}$ and $\boldsymbol{Z}^{estimated}$ is given by

$$NMI(\boldsymbol{Z}^{true}, \boldsymbol{Z}^{estimated}) = \frac{2 \cdot MI(\boldsymbol{Z}^{true}, \boldsymbol{Z}^{estimated})}{MI(\boldsymbol{Z}^{true}, \boldsymbol{Z}^{true}) + MI(\boldsymbol{Z}^{estimated}, \boldsymbol{Z}^{estimated})},$$

where $MI(\boldsymbol{Z}^{estimated}, \boldsymbol{Z}^{estimated})$ is the mutual information as defined in the lecture. Notice, $0 \leq NMI \leq 1$ where 0 indicates no relationship between the two assignment matrices and 1 indicates a perfect correspondence.

Link prediction: In order to evaluate how well the model generalize to unobserved data we will use the IRM model to predict entries in the graph treated as missing. Treating entries as missing can be achieved by introducing an indicator matrix $\boldsymbol{W}$ where $W_{ij} = 0$ means that the observation A_{ij} is observed whereas $W_{ij} = 1$ indicates that the observation A_{ij} is unknown when inferring the model parameters. This changes the link and non-link counts by $N_{lm}^+ = \boldsymbol{Z}\boldsymbol{A}\boldsymbol{Z}_{lm}^\top - [\boldsymbol{Z}(\boldsymbol{A} \circ \boldsymbol{W})\boldsymbol{Z}^\top]_{lm}$ and $N_{lm}^- = n_k n_l - \delta_{lm} n_k - N_{lm}^+ - [\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}^\top]_{lm}$, where $\circ$ is the direct product, i.e. $(\boldsymbol{A} \circ \boldsymbol{W})_{ij} = A_{ij} W_{ij}$. We will evaluate the probability of observing a link in each missing entries from the IRM model according to $\pi_{ij} = \boldsymbol{z}_i^\top \langle \boldsymbol{\eta} \rangle \boldsymbol{z}_j$ and use the area under curve (AUC) of the receiver operator characteristic (ROC) to evaluate how well the model predicts, i.e. AUC=0.5 is equivalent to predicting by change while AUC=1 indicates that the predictions made by the IRM model perfectly separates links from non-links.

Part 1: Analysis of synthetic data where ground truth is known by the IRM

In this part of the exercise we will investigate how well the IRM model is able to infer the parameters of various graphs generated according to the model (i.e. generated according to equation (1)-(3)).

Q 1.1: Inspect and run the script *analData.m*. The script uses the function *generateGraphCRPUnipartite.m* to generate graphs according to the IRM model and infers by the script *IRMUnipartite.m* from the generated unsorted adjacency matrix $\boldsymbol{A}$ the group structure $\boldsymbol{Z}^{estimated}$ (notice that the

input W is a graph the size of A indicating entries treated as missing. The script *createValidation-Data.m* generates a W matrix of missing entries treating a given percentage of links and equivalent number of nonlinks as missing at random. The results of the analysis is visualized using the script *plotSyntheticResults.m*. This script automatically compares the estimated group structure to the true group structure using the script *calcNMI.m* while the link predictive performance is evaluated by the AUC score using the script *calcAUC.m*, see also lecture slides for details.

Q 1.2: Try vary the parameters J, *alpha bp* and *bn* used to generate the graphs in the script *generateGraphCRPUnipartite.m* and explain when the inference procedure for IRM is able to recover well the true structure of the graphs (try for instance bp=[1 1] bn=[1 1], bp=[1 10] bn=[10 1] and bp=[100 100] bn=[100 100]) (Notice, if you set $J > 500$ it may take too long to generate the graphs and infer the parameters). Try and explain your findings.

Part 2: Predicting Movie Viewing by the IRM

We will consider the Movielens 100k data containing the ratings of movies by users. This data set is a collection of users rating (1-5) of movies, together with demographic information for each user and detailed information about the movies. The MovieLens dataset is available from

<div align="center">

http://www.grouplens.org/node/73

</div>

We will attempt to predict whether a user saw a given movie based on the bipartite version of the infinite relational model given in the script *IRMBipartite.m*. We are trying to solve a so-called collaborative filtering problem, i.e. from the preferences of users with similar taste predict the preferences of a given user for a given movie (this entry is then treated as missing). Rather than predicting how much a user likes a given movie we are now interested in solely predicting whether a user would like to see a given movie or not. The bipartite IRM model groups the users according to the assignment matrix $\boldsymbol{Z}^{(user)}$ and movies according to the assignment matrix $\boldsymbol{Z}^{(movie)}$ such that the inferred probability of user i watching movie j is given by $\pi_{ij} = \boldsymbol{z}_i^{(user)\top} \boldsymbol{\eta} \boldsymbol{z}_j^{(movie)}$.

Below is a short summary of the most important information in the data:

The user by movie matrix denoted by A, is a binary matrix indicating which of the 943 users have watched which of the 1682 movies. There are a total of 100K movie viewings and each user has seen at least 20 movies.

User information *user* is a cell array containing user id, age, gender, occupation and zip code for all users.

Movie information *movie* is a cell array containing movie id, movie title, release date, video release date, IMDb URL and genre by 1 out of K coding and by name. The genres are: unknown, Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War and Western and a movie can be in more than one genre.

Q 2.1: Inspect and run the script *analMovieLensDataBipartite.m*. The script analyze the MovieLens data treating 2.5% of the links and an equivalent number of non-links as missing and displays the results and the distribution of movie genres of the movie clusters and distribution of gender and age in of the extracted user groups. Try and understand Figure 1 and Figure 2. How well does the model predict whether a user saw a given movie? Does the model appear to identify prominent structure in the graph?

Q 2.2: How would you interpret the extracted clusters? (Hint: Take a look at Figure 3-5: The first bar plot indicate the distribution of movie genre for each movie cluster. The second bar-plot the distribution in gender (left side of red vertical line) and age (right side of vertical line) for the extracted user groups.

Part 3: Analysis of the original data of [Kemp et al., 2006] by the IRM

In this part of the exercise we will consider the data analyzed in [*Kemp et al. AAAI 2006*].

Q 3.1: Take a look inside the folder *irmdata*. The folder contains the data described in the original paper on the IRM model by [*Kemp et al. AAAI 2006*] . Open the *READMEDatasets.txt* file and understand what each of the datasets represent from the original article provided as *KempTGYU06.pdf* in the folder.

Q 3.2: Write a script that can analyze the *50animalbindat* data. Use link-prediction to evaluate how well the IRM model account for structure in the relational data. (this type of analysis was not considered in the original paper by Kemp et al.). Does the IRM model predict better than chance? Does the extracted group structures appear to be reasonable? Discuss your findings.

Q 3.3: (Extra Challenge) Contrary to the analysis in Kemp et al. for multi-relational data (i.e. forming multiple-graphs on the same vertex sets) where the model clustered also in the relations we will presently specify individual relation parameters for each graph, i.e. $\boldsymbol{A}^{(n)} \sim Bernoulli(\boldsymbol{z}_i \boldsymbol{\eta}^{(n)} \boldsymbol{z}_j)$. Furthermore, we remove self-links from the Unipartite graphs. Try analyze some of the other data sets considered in [*Kemp et al. AAAI 2006*] (Notice, if the data form unipartite directed graphs you have to specify *opts.type='Directed'* when using the script *IRMUnipartite.m*. Furthermore, if you choose to analyze the *dnations* data note that we presently only consider one dataset type at a time, i.e. we will either analyze the multiple relations between countries or the feature data in separate analysis. Notice also that the *dnations* data have unknown entries that should be treated as missing in the IRM inference given by *WmissingA* and *WmissingB*.)

Mini-project suggestion: Write a small reports based on your findings in exercise part 1-3 and include for instance an additional analysis of one the datasets in the extra challenge or one of the many graphs available from `http://www.cise.ufl.edu/research/sparse/mat/Pajek/`.