Bayesian Factor Modeling for Interest-Based Advertising with Count Data

Alvin Kaule
02459 Machine Learning for Signal Processing, May 19th 2010
DTU Informatics · Technical University of Denmark, Kgs. Lyngby, Denmark

Introduction

• We explore latent factor Bayesian models for dyadic data in the domain of online interest-based advertising which uses ad clicks and ad impressions from user history.

• The models we explore exploits the fact that the dyadic data are based on event observations and thus can be treated as count data.

• We compare different methods for estimating the factors in the models, by computing the maximum likelihood (ML), maximum posteriori (MAP) estimate as well a perform inference using Gibbs sampling to obtain the posterior distribution of the factors.

• We show how to use the model for predicting ad performance on user basis in form of click-through rate (CTR) which can be viewed as p(click|ad,user).

Factor Model for Count Data

The observed dyadic data can be represented as an $N \times M$ matrix $\mathbf{X}$, whose elements $x_{i,j}$ is the observed count of event j by user i. The matrix can be represented as an approximation $\mathbf{X} \approx \mathbf{Y} = \mathbf{U}\mathbf{V}^T$

$$\begin{pmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,M} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N,1} & y_{N,2} & \cdots & y_{N,M} \end{pmatrix} = \begin{pmatrix} u_{1,1} & \cdots & u_{1,K} \\ u_{2,1} & \cdots & u_{2,K} \\ \vdots & \ddots & \vdots \\ u_{N,1} & \cdots & u_{N,K} \end{pmatrix} \times \begin{pmatrix} v_{1,1} & v_{2,1} & \cdots & v_{M,1} \\ \vdots & \vdots & \ddots & \vdots \\ v_{1,K} & v_{2,K} & \cdots & v_{M,K} \end{pmatrix},$$

where $\mathbf{U}$ is a $N \times K$ matrix and $\mathbf{V}$ is a $K \times M$ matrix. The row vector $\boldsymbol{u}_i^T$ of $\mathbf{U}$ is a representation of user i in a latent space of lower dimensionality K and the elements $u_{i,k}$ can be see as a measurement of interest of user i to factor k as the total number of occurrences of all events contributing to factor k, while the elements in row vector $\boldsymbol{v}_k^T$ of $\mathbf{V}$ provides the latent factor for a particular event (e.g. ad click). We further constrain the factors of $\mathbf{U}$ and $\mathbf{V}$ to consists of non-negative elements.

Since the data is count data a natural assumption is that $\mathbf{X}$, element-wise follow Poisson distributions with corresponding mean parameters in $\mathbf{Y}$ such that

$$x_{i,j} \sim Poisson\left(y_{i,j}\right).$$

Each element $x_{i,j}$ can be seen as the sum of events (e.g. number of clicks by user i on ad j) where each event is a Bernoulli random variable with small probability of "success" and thus the sum is approximately Poisson distributed.

Maximum likelihood with EM

Assuming each element $x_{i,j}$ to be i.i.d., the Poisson likelihood of observing $\mathbf{X}$ is given by

$$p(\mathbf{X}\,|\,\mathbf{U},\mathbf{V}) = \prod_{i,j} Poisson\left((\mathbf{U}\mathbf{V}^T)_{i,j}\right) = \prod_{i,j} \frac{(\mathbf{U}\mathbf{V}^T)_{i,j}^{\,x_{i,j}} \exp\left(-(\mathbf{U}\mathbf{V}^T)_{i,j}\right)}{x_{i,j}!}.$$

Given our data $\mathbf{X}$ we wish to find the model parameters ($\mathbf{U}$, $\mathbf{V}$) by maximizing the log likelihood w.r.t. $\mathbf{X}$

$$\ln p(\mathbf{X}\,|\,\mathbf{U},\mathbf{V}) = \sum_i \ln p(\mathbf{x}_i\,|\,\mathbf{u}_i,\mathbf{V}) = \sum_{i,j} \ln\left(\frac{\left(\sum_k u_{i,k}v_{j,k}\right)^{x_{i,j}} \exp\left(-\sum_k u_{i,k}v_{j,k}\right)}{x_{i,j}!}\right)$$

$$= \sum_{i,j}\left(-\sum_k u_{i,k}v_{j,k}\right) + x_{i,j}\ln\left(\sum_k u_{i,k}v_{j,k}\right) - \ln\left(x_{i,j}!\right),$$

is equivalent to minimizing the KL-divergence

$$D(\mathbf{A}\|\mathbf{B}) = \sum_{i,j}\left(A_{i,j}\log\frac{A_{i,j}}{B_{i,j}} - A_{i,j} + B_{i,j}\right),$$

for which [1] gave following EM algorithm with multiplicative update rules

$$u_{i,k} \leftarrow u_{i,k}\frac{\sum_j \frac{v_{j,k}x_{i,k}}{(\mathbf{U}\mathbf{V}^T)_{i,j}}}{\sum_j v_{j,k}} \qquad v_{j,k} \leftarrow v_{j,k}\frac{\frac{u_{i,k}x_{i,k}}{(\mathbf{U}\mathbf{V}^T)_{i,j}}}{\sum_i u_{i,k}}.$$

Maximum a posteriori with EM

Introducing prior distributions over the factors in the model we can compute the MAP estimate of the factors.

We assume that the elements of $u_{i,k}$ is independently gamma distributed with shape parameters α_k and scale parameters β_k

$$u_{i,k} \sim Gamma\left(\alpha_k, \beta_k\right) \qquad \text{for } k = 1,..,K$$

Using Bayes rule, the posterior is proportional to

$$p(\mathbf{U},\mathbf{V}\,|\,\mathbf{X}) \propto p(\mathbf{X}\,|\,\mathbf{U},\mathbf{V})\,p(\mathbf{U},\mathbf{V}) = \prod_i \left(\prod_j \frac{\left(\sum_k u_{i,k}v_{j,k}\right)^{x_{i,j}} \exp\left(-\sum_k u_{i,k}v_{j,k}\right)}{x_{i,j}!} \prod_k \frac{\beta_k^{\alpha_k} u_{i,k}^{\alpha_k - 1} \exp\left(-\beta_k u_{i,k}\right)}{\Gamma(\alpha_k)} \right)$$

This model is called the GaP (Gamma-Poisson) model and was introduced by [2]. Following EM GaP algorithm based on [1] was derived to compute a MAP estimate of the model parameters

$$u_{i,k} \leftarrow u_{i,k}\frac{\sum_j \frac{v_{j,k}x_{i,k}}{(\mathbf{U}\mathbf{V}^T)_{i,j}} + \frac{(\alpha_k - 1)}{u_{i,k}}}{\sum_j v_{j,k} + \frac{1}{\beta_k}} \qquad v_{j,k} \leftarrow v_{j,k}\frac{\sum_i \frac{u_{i,k}x_{i,k}}{(\mathbf{U}\mathbf{V}^T)_{i,j}}}{\sum_i u_{i,k}}.$$

Markov Chain Monte Carlo with Gibbs sampling

In a Bayesian treatment we seek the posterior distribution over parameters instead of a point estimate such as given by ML and MAP. Unfortunately Bayesian inference is often not feasible instead we turn to an approximation method for Bayesian inference based on Markov Chain Monte Carlo (MCMC) sampling.

Discrete Component Analysis (DCA) for which GaP is a special case of was introduced by [3] and presented a Gibbs sampling algorithm. Introducing a new latent matrix $\mathbf{H}$ the Bayesian model is specified as follows

$$u_{i,k} \sim Gamma\left(\alpha_k, \beta_k\right)$$
$$c_{i,k} \sim Poisson\left(u_{i,k}\right)$$
$$x_{i,j} = \sum_k h_{j,k,(i)}, \qquad \text{where } h_{j,k,(i)} \sim Multinomial\left(c_{i,k}, \mathbf{v}_k\right)$$
$$\mathbf{v}_k \sim Dirichlet\left(\gamma\right),$$

where $\boldsymbol{c}_i$ is a discrete latent K-dimensional vector, where $c_{i,k}$ gives the number of ad events by user i in the k-th factor. The latent vector $\boldsymbol{c}_i$ is derived from a new latent $J \times K$ matrix $\mathbf{H}_{(i)}$ for each user i with elements $h_{j,k,(i)}$. The sum of the row of $\mathbf{H}_{(i)}$ is the count $x_{i,j}$ in the observed data and the sum of the column is $c_{i,k}$. The column vector $\mathbf{v}_k$ of $\mathbf{V}$ is normalised across the features such that $\sum_j v_{j,k} = 1$. The joint posterior is easily derived after introducing the new latent matrix and is given by

$$p(\mathbf{U},\mathbf{H},\mathbf{V}\,|\,\mathbf{X},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}) = \prod_i p(\mathbf{c}_i\,|\,\mathbf{u}_i)\prod_i p(\mathbf{u}_i\,|\,\boldsymbol{\alpha},\boldsymbol{\beta})\prod_{i,k} p\left(\mathbf{h}_{k,(i)}\,|\,\mathbf{c}_i,\mathbf{v}_k\right)\prod_k p(\mathbf{v}_k\,|\,\boldsymbol{\gamma})$$

$$= \prod_i \left(\prod_k \frac{u_{i,k}^{c_{i,k}} \exp\left(-u_{i,k}\right)}{c_{i,k}!} \prod_k \frac{\beta_k^{\alpha_k} u_{i,k}^{\alpha_k - 1} \exp\left(-\beta_k u_{i,k}\right)}{\Gamma(\alpha_k)} \prod_{j,k} \frac{c_{i,k}!}{h_{j,k,(i)}!} v_{j,k}^{h_{j,k,(i)}} \right) \prod_{j,k} v_{j,k}^{\gamma_j}$$

$$= \prod_i \left(\prod_k \frac{\beta_k^{\alpha_k} u_{i,k}^{c_{i,k}} u_{i,k}^{\alpha_k - 1} \exp\left(-u_{i,k}\right)\exp\left(-\beta_k u_{i,k}\right)}{\Gamma(\alpha_k)} \prod_{j,k} \frac{v_{j,k}^{h_{j,k,(i)}}}{h_{j,k,(i)}!} \right) \prod_{j,k} v_{j,k}^{\gamma_j}$$

$$= \prod_i \left(\prod_k \frac{\beta_k^{\alpha_k} u_{i,k}^{c_{i,k} + \alpha_k - 1} \exp\left(-\left(\beta_k + 1\right)u_{i,k}\right)}{\Gamma(\alpha_k)} \prod_{j,k} \frac{v_{j,k}^{h_{j,k,(i)}}}{h_{j,k,(i)}!} \right) \prod_{j,k} v_{j,k}^{\gamma_j}.$$

With Gibbs sampling we need the posterior conditional distributions, which in our case can all be represented by well known distributions and are given as follows

$$p\left(u_{i,k}\,|\,\mathbf{X},\mathbf{U}_{\setminus u_{i,k}},\mathbf{H}_{(i)},\mathbf{V},\boldsymbol{\alpha},\boldsymbol{\beta}\right) = Gamma\left(c_k + \alpha_k, 1 + \beta_k\right)$$

$$p\left(\mathbf{h}_{k,(i)}\,|\,\mathbf{X},\mathbf{U},\mathbf{H}_{(i)\setminus\mathbf{h}_{k,(i)}},\mathbf{V},\boldsymbol{\alpha},\boldsymbol{\beta}\right) = Multinomial\left(x_{i,j}, \left\{\frac{u_{i,k}v_{j,k}}{\sum_k u_{i,k}v_{j,k}}\right\} : k\right)$$

$$p\left(\mathbf{v}_k\,|\,\mathbf{X},\mathbf{U},\mathbf{H},\mathbf{V}_{\setminus\mathbf{v}_k},\boldsymbol{\alpha},\boldsymbol{\beta}\right) = Dirichlet\left(\sum_i \mathbf{h}_{k,(i)}\right),$$

where $\mathbf{U}_{\setminus u_{i,k}}$ denotes all elements of $\mathbf{U}$ except $u_{i,k}$. By iteratively sampling the posterior conditional distributions we obtain an approximation to the joint posterior distribution.

CTR Prediction

In order to predict the CTR for a given user i and ad j we use following

$$CTR_{i,ad(s)} = \frac{\mathbf{u}_i\mathbf{v}_{click(s)} + a}{\mathbf{u}_i\mathbf{v}_{impression(s)} + b},$$

where *click*(*s*) and *impression*(*s*) are the indices corresponding to the click/impression feature pair of ad *s*, respectively, by user *i.* The smoothing constants *a* and *b* are the total clicks and impressions for ad *j* such that a/b gives the CTR of a user with no history.

Evaluation and Results

We evaluate the predictive performance of the model for count data on advertising data.

• Each ad is represented by 3 features, impression, click and conversion.

• Data was collected over a period of 1 month and consists of *N*=960 users and 110 ads (*M*=330).

• The data matrix contains 2548 elements representing a total of 6052 ad events.

• The data is split into 80% training set and 20% test set.

• We compare inference with Gibbs GaP with algorithms for ML using NMF [1], MAP using the EM GaP algorithm [4] and a Gibbs sampler for Bayesian NMF [5].

• We compare the algorithms for a number of latent dimensions, *K*=1, *K*=3, *K*=5, *K*=10, *K*=20, *K*=50. The performance is reported in RMSE computed over the test set.

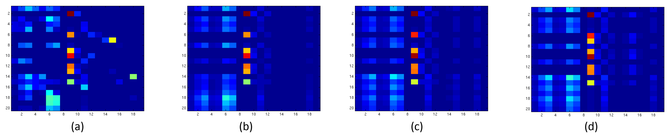

FIGURE 1: Toy sample set for illustrative purpose (k=4). (a) Data matrix, (b) NMF, (c) EM GaP, (d) Gibbs GaP.

K	Gibbs GaP	Gibbs NMF	NMF	EM GaP
	α=1.1, β=0.01			α=1.1, β=0.1
5	2.979	3.062	2.936	3.014
10	3.151	3.183	3.072	3.080
20	3.119	3.187	3.151	3.240

TABLE 1: RMSE on the test set for the different algorithms. Using the mean over the entire training set results in a baseline model with RMSE 3.206.

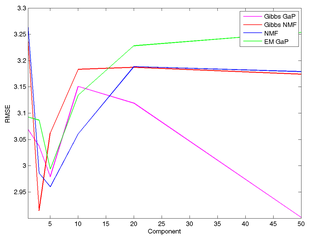

FIGURE 2: RMSE on the test set as a function of *K*.

Conclusion

• Presented a Bayesian model for count data and applied it to real world advertising data.

• Implemented Gibbs sampler in Matlab and compared it to other algorithms for count data.

• Results show good results on the advertising data for the Gibbs GaP sampler.

References

[1] D. D. Lee and H. S. Seung. Algorithms for Non-negative Matrix Factorization. *In Advances in Neural Information Processing 13 (Proc. NIPS 2000).* MIT Press, 2001.

[2] J. F. Canny. GaP: A Factor Model for Discrete Data. *ACM Conference on Information Retrieval SIGIR 2004,* (pp. 122–129), 2004.

[3] W. Buntine, A. Jakulin. Discrete Component Analysis. In *Subspace, Latent Structure and Feature Selection*, vol. 3940/2006, pp. 1–33, Springer (LNCS), 2006.

[4] Y. Chen, M. Kapralov, D. Pavlov, J. F. Canny. Factor Modeling for Advertisement Targeting, *NIPS 2009*, 2009.

[5] M. N. Schmidt, O. Winther, L. K. Hansen. Bayesian non-negative matrix factorization, *Independent Component Analysis and Signal Separation, International Conference on*, 2009.