ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 1-6

A Generative Model for Sparse Discrete Binary Data with Non-Uniform Categorical Priors

Mark Girolami

Applied Computational Intelligence Research Unit
Department of Computing and Information Systems
University of Paisley
Paisley PA1 2BE
Scotland

Abstract. The Generative Topographic Mapping (GTM) was developed and introduced as a *principle*d alternative to the Self-Organising Map for, principally, visualising high dimensional continuous data. There are many cases where the observation data is ordinal and discrete and the application of methods developed specifically for continuous data is inappropriate. Based on the continuous GTM data model a non-linear latent variable model for modeling sparse high dimensional binary data is presented. The primary motivation for this work is the requirement for a dense and low dimensional representation of sparse binary vector space models of text documents based on the multivariate Bernoulli event model. The method is however applicable to binary data in general.

1. Introduction

Generative latent variable models have recently been proposed as tools for the visualisation and analysis of high dimensional data [1, 5]. The Generative Topographic Mapping (GTM) [5] was proposed as a means of visualising high dimensional continuous data on a two-dimensional grid. The GTM is essentially based on a nonlinear mapping from an $L = 2$ dimensional uniform grid of points in latent space $\mathbf{x} \in \Re^{L=2}$ to a D-dimensional observation space $\mathbf{t} \in \Re^{D}$. The probability of the observations conditioned on the latent variables $p(\mathbf{t}|\mathbf{x})$ is chosen as an isotropic Gaussian, indicating that $p(\mathbf{t}|\mathbf{x}) = \prod_{i=1}^{D} p(t_i|\mathbf{x})$. The choice of the uniform prior of each point in latent space gives the observation data probability as an equally weighted mixture of univariate Gaussians. An efficient EM algorithm is then developed for the parameter estimation of each of the Gaussian mixture components within the GTM data model[5].

This work is supported by the British Library and Information Commission, Grant Number LIC/GC/991 'Improved Online Information Access'

More recently the notion of dealing with binary and categorical data types has been considered in [1]. As with the GTM the observation data density is given as the following expression.

$$p(\mathbf{t}) = \int \left\{ \prod_{i=1}^{D} p(t_i|\mathbf{x}, \theta) \right\} p(\mathbf{x}) d\mathbf{x} \tag{1}$$

In this case however as the observations are binary then the conditional distribution in (1) is given as a univariate single trial binomial distribution $P(t_i|\mathbf{x}) = p_i^{t_i}(1-p_i)^{1-t_i}$ where p_i is the expected value of the binomial variable. The key assumption of this work is that the latent space is distributed as a zero mean, unit variance Gaussian. By introducing a variational approximation for the conditional element of (1) it is noted that the exponential introduced will be quadratic in $\mathbf{x}$ and as the Gaussian prior is also quadratic in $\mathbf{x}$ the integral of the variational approximation can be solved, very elegantly, and so the required parameter estimates can be given in closed form. However by imposing a Gaussian prior on the continuous latent space the model defined in (1) will effectively provide a singular value decomposition of the binary observations onto the plane associated with the two most significant singular values.

In contrast to [1, 5] this work considers the latent space to consist of discrete categorical data points which will each be defined by a prior probability $P(\mathbf{x}_k)$ such that $\sum_k^K P(\mathbf{x}_k) = 1$. By defining the latent space in such a manner then we are effectively creating a binary latent trait model[1] [3]. The primary motivation for this is to be able to define an alternative *semantically dense* [8] representation of a vector space document model based on the multivariate Bernoulli event model [2]. The restructuring of a vector space document model using Independent Component Analysis (ICA) [6] has been proposed in [7]. However the current linear ICA data models appear to be inappropriate for identifying the mapping from *concept* space into word space. We therefore turn to probabilistic generative models for defining the text generation mechanism.

2. Algorithm

Let us consider a latent space representation of ordinal categorical points $\mathbf{x}_k \in \mathcal{K} = \{\mathbf{x}_1, \cdots, \mathbf{x}_{|\mathcal{K}|}\}$. Each $\mathbf{x}_k$ may, in some sense, represent a class or topic label. The integral in (1) now reduces to a summation and so the observation probability is given by.

$$p(\mathbf{t}) = \sum_{k=1}^{K} \left\{ \prod_{i=1}^{D} p(t_i|\mathbf{x}_k, \theta) \right\} P(\mathbf{x}_k) \tag{2}$$

Where θ defines the model parameters. It is interesting to note that the conditional independence of the observation probability density embodied in the

[1] Defining the latent space as a hyper-cube provides a latent class model

product of marginal probabilities is a direct match to the multivariate conditional independence of the *Bag-of-words* document representation [2]. In developing an Expectation-Maximisation procedure for the estimation of the model parameters the standard approach of forming a relative likelihood between *old* parameter estimates and *new* ones yields the following expression when employing Jensens approximation [4].

$$\tilde{Q} = \sum_{n,k}^{N,K} P^{old}(\mathbf{x}_k|\mathbf{t}_n) log \left\{ \left\{ \prod_{i=1}^{D} P^{new}(t_{in}|\mathbf{x}_k, \theta) \right\} P^{new}(\mathbf{x}_k) \right\} \tag{3}$$

Where each $P(t_i|\mathbf{x})$ is defined by the Binomial distribution $p_i^{t_i}(1-p_i)^{1-t_i}$. The generalised linear model [3] for Bernoulli response variables proposes the use of the logit based link function from the covariates, which in this case correspond to the latent variables $\mathbf{x}_k$. A nonlinear transformation from the latent space is denoted here in keeping with the general form of transformation proposed in [5]. The logit transformation is then given as the following.

$$E\{t_i|\mathbf{x}_k\} = p_{ik} = \frac{exp(\mathbf{w}_i^T \phi(\mathbf{x}_k))}{1 + exp(\mathbf{w}_i^T \phi(\mathbf{x}_k))} \tag{4}$$

The estimation of the *new* parameters can be found in a straightforward manner [4]. The latent prior probability is updated in the standard way, $P^{new}(\mathbf{x}_k) = \frac{1}{N}\sum_n P^{old}(\mathbf{x}_k|\mathbf{t}_n)$. Due to the nonlinear link function the estimation of the new conditional probabilities requires an inner iterative loop to estimate the parameters $\mathbf{w}_i$. The following derivatives with respect to each w_{mi} are required.

$$\frac{\partial \tilde{Q}}{\partial w_{mi}} = \sum_{n,k} P^{old}(\mathbf{x}_k|\mathbf{t}_n) \{t_{in} - p_{ik}\} \phi_m(\mathbf{x}_k) \tag{5}$$

If we are to use Newton or pseudo-Newton type optimisation approaches which require the inverse of the Hessian then the second order derivatives with respect to w_{mi} require to be computed. Note that as the expression for the posterior probabilities $P^{old}(\mathbf{x}_k|\mathbf{t}_n)$ is fixed within this step it does not form part of the expression for the derivative.

$$\frac{\partial^2 \tilde{Q}}{\partial w_{mi} \partial w_{m'i}} = -\sum_{n,k} P^{old}(\mathbf{x}_k|\mathbf{t}_n) p_{ik} \{1 - p_{ik}\} \phi_m(\mathbf{x}_k) \phi_{m'}(\mathbf{x}_k) \tag{6}$$

Both of these expressions can be given in convenient matrix format by defining the following set of matrices [5]. The $K \times M$ matrix $\mathbf{\Phi}$ has individual elements $\phi_m(\mathbf{x}_k)$ and $\mathbf{W}$ is a $M \times D$ matrix with elements w_{mi}. $\mathbf{M}$ is a $K \times D$ matrix whose elements are p_{ik} and the $N \times D$ matrix $\mathbf{T}$ represents the binary data matrix.

The $K \times K$ diagonal matrix whose k^{th} diagonal element is $\sum_n P^{old}(\mathbf{x}_k|\mathbf{t}_n)$ is defined as $\mathbf{G}$. $\mathbf{P}$ is a $K \times N$ matrix with elements $P^{old}(\mathbf{x}_k|\mathbf{t}_n)$. $\mathbf{Z}_i$ is the $K \times K$

diagonal matrix with elements $p_{ik}(1-p_{ik})$. As stated above the estimation of the parameters $\mathbf{W}$, which define each $P^{new}(t_{in}|\mathbf{x}_k) = p_{ik}^{t_{in}}(1-p_{ik})^{(1-t_{in})}$, in this M-step is iterative and a number of optimisation techniques are available to us. In the case of a simple gradient ascent then the matrix of parameters $\mathbf{W}$, can be updated using the following iterative procedure.

$$\mathbf{W}^{n+1} = \mathbf{W}^n + \delta_n \left[\mathbf{\Phi}^T\mathbf{P}\mathbf{T} - \mathbf{\Phi}^T\mathbf{G}\mathbf{M}^n\right] \qquad (7)$$

where δ_n is the step size at each n^{th} update iteration. Note that only the matrix of binomial distribution means p_{ik} defined by (4) requires to be updated at each step. The new value of this matrix is denoted in the usual way by $\mathbf{M}^n$.The iterative re-weighted least-squares approach [3] can also be used noting that we can write the Hessian matrix for each independent element t_i as $\mathbf{H}_i = -\mathbf{\Phi}^T\mathbf{G}\mathbf{Z}_i\mathbf{\Phi}$ then using the standard second order Newton type method for parameter update

$$\mathbf{W}_i^{n+1} = \mathbf{W}_i^n + \left[\mathbf{\Phi}^T\mathbf{G}\mathbf{Z}_i\mathbf{\Phi}\right]^{-1}\left[\mathbf{\Phi}^T\mathbf{P}\mathbf{T}_i - \mathbf{\Phi}^T\mathbf{G}\mathbf{M}_i^n\right] \qquad (8)$$

where $\mathbf{T}_i$, $\mathbf{M}_i$ and $\mathbf{W}_i$ denote the i^{th} column of the respective matrices. This completes the M-step estimation of the *new* parameters.

Specifically, for modeling text documents our prior domain knowledge indicates that the distribution of word occurrences is sparse and so $E\{t_i|\mathbf{x}_k\} \rightarrow 0$. This suggests that the $\mathbf{w}_i$ are distributed in a skew edmanner concentrated about *large* negative values. A prior over the coefficients can be imposed in (1) such that, for example, $p(\mathbf{w}_i) = \beta exp(-\beta\{\|\mathbf{w}_i\| \sum_j w_{ij}\})$ with $\beta \geq 0$. Setting $\beta = 1$ yields the additional gradient term $-\mathbf{1}diag(\mathbf{W}^T\mathbf{W}) - \mathbf{W}diag(\mathbf{1}^T\mathbf{W})$ in (7). Where the $M \times d$ matrix of unit values is denoted by $\mathbf{1}$ and the operator *diag* sets all the off-diagonal terms of the argument to zero.

3. Simulations

The main motivation for this work is to develop probabilistic models which will be able to model the word generation process within text documents and provide a means to represent documents in an efficient manner. Latent Semantic Analysis (LSA) [8] performs a singular value decomposition on the term document matrix representations of document corpora and the dimensions associated with the most significant singular values are retained. The absence of a probabilistic model and the difficulty in interpreting the transformed representation motivate the search for other representations. Both the SOM and ICA [7, 9] have been applied to this domain, it can be argued that the shortcomings of LSA are also found with these methods.

For this experiment an arbitrary collection of three thousand documents, one thousand from each of three newsgroups *comp.graphics, alt.atheism, talk.politics.guns*, was taken from the **CMU-Newsgroups**[2] collection. A vocabulary size of one hundred terms was used throughout in the document representations. The number of latent points, (*concepts* in information retrieval

[2] http://www.cs.cmu.edu/˜ textlearning

comp.graphics (50)	alt.atheism (38)	alt.atheism (10)	politics.guns (85)
graphics	god	people	gun
program	people	objective	guns
image	religion	morality	weapons
computer	fact	moral	firearms
file	atheism	evidence	law

Table 1: The five most probable words i.e. $P(t_i = graphics|\mathbf{x}_{k=90})$ from each latent point with the maximum posterior probability.

parlance) was taken as one hundred. The variational approach proposed in [1] was employed with the results being identical to those found with LSA (SVD) and this was unable to uncover the latent structure (or latent topics) of the document collection. The EM was run for 20 iterations and a fixed number of five inner iterations were employed for the estimation of the $\mathbf{W}$ matrix.

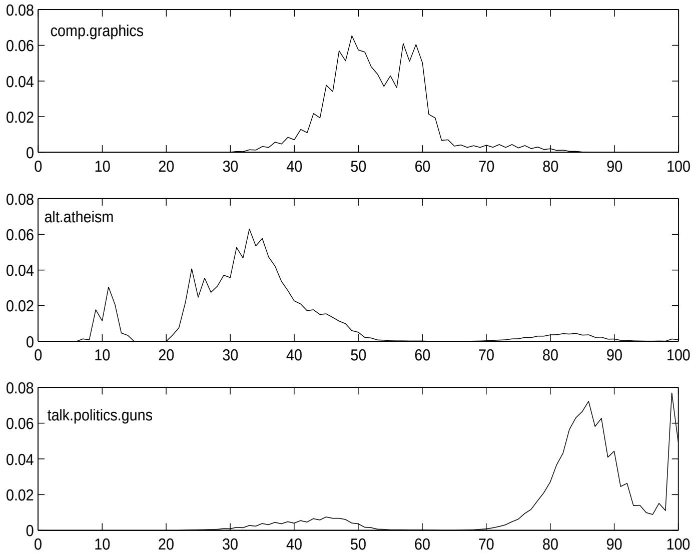

Figure 1: Distribution of posterior probability of latent points (topics) given a document $P(\mathbf{x}_k|\mathbf{t}_n)$ for each group of 1000 documents.

Figure 1. shows the posterior probability that each of the 1000 documents was generated from the various latent points. It is clear that the algorithm has been able to define a topical grouping of the documents which reflect the groupings within the corpora. It can be seen that the posterior for *alt.atheism* has two modes, inspection of the five most probable words ranked for each document by the point associated with the maximum posterior probability shows that one set of documents could be associated with discussion on religion (point 38) and another discusses objective morality (point 10). We can now provide a reduced dimensional representation of multivariate Bernoulli event models of documents where each axis denotes a mode of the posterior and each element is a readily interpretable probability of topic relevance.

4. Conclusion

Based on the theoretical basis of the GTM a latent trait model for sparse high dimensional binary data has been proposed. In this contribution the method has been applied to the analysis of binary representations of documents which may have a number of dimensions in excess of 1000 and yet only a small number of elements will be non-zero. The results have been most encouraging and further work into Latent Trait/Class document representations is underway. This proposed method is, however, suitable for binary data in general.

References

[1] Tipping, M.E., Probabilistic Visualisation of High-Dimensional Binary Data. *NIPS 11*, pp. 592-598, 1999.

[2] McCallum, A., and Nigam, K. A comparison of event models for naive Bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorisation*. http://www.cs.cmu.edu/ mccallum, 1998.

[3] McCullagh, P., and Nelder, J.A. *Generalised Linear Models*. Chapman and Hall. 1985.

[4] Bishop, C.M. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[5] Bishop, C.M., Svensén, M., and Williams, C.K.I. GTM: The Generative Topographic Mapping. *Neural Computation*, 10(1), 1998.

[6] Girolami, M. Cichocki. A., and Amari, S, I. A Common Neural Network Model for Exploratory Data Analysis and Independent Component Analysis. *IEEE Trans on Neural Networks*, Vol 9, No.6, pp 1495 - 1501, 1998.

[7] Isbell, B.L., Viola, P. Restructuring Sparse High Dimensional Data for Effective Retrieval. *Advances in Neural Information Processing Systems* **11**, 480-486, 1999.

[8] Deerwester, S., Dumais, S,T., Furnas, G,W., Landatter, T,K., Harshman, R. Indexing By Latent Semantic Analysis. *J. Amer. Soc. fo Inf. Science* **41**, 391-407, 1990.

[9] Kohonen, T. *Self-Organising Maps*, Springer-Verlag, 1995.