Normalized Gaussian Networks with Mixed Feature Data

Shu-Kay Ng[1] and Geoffrey J. McLachlan[1,2]

[1] Department of Mathematics, University of Queensland,
Brisbane, QLD 4072, Australia
{skn, gjm}@maths.uq.edu.au
[2] Institute for Molecular Bioscience, University of Queensland,
Brisbane, QLD 4072, Australia

Abstract. With the mixed feature data, problems are induced in modeling the gating network of normalized Gaussian (NG) networks as assumption of multivariate Gaussian becomes invalid. In this paper, we propose an independence model to handle mixed feature data within the framework of NG networks. The method is illustrated using a real example of breast cancer data.

1 Introduction

Normalized Gaussian (NG) networks, such as the NG mixture of experts (NGME) nets [1], are of extensive interest due to their wide applicability, generalization capability, and the advantage of efficient learning via the expectation-maximization (EM) algorithm [2]; see for example [1, 3, 4]. For many applied problems in machine learning, there often involves both categorical and continuous feature variables [5]. With the mixed feature data, the input vector $\boldsymbol{x}_j$ on the j-th entity consists of q categorical variables in the vector $\boldsymbol{x}_{1j}$ in addition to p continuous variables represented by the vector $\boldsymbol{x}_{2j}$ for $j = 1, \ldots, n$, where n is the total number of observations. Problems are therefore induced in modeling the gating network with NG networks as the assumption of multivariate Gaussian becomes invalid when the data are mixed-mode. In this paper, we propose an independence model to handle mixed feature data within the framework of NG networks. The method bases on the NAIVE assumption that the categorical variables are independent of each other and of the continuous variables [6, 7].

2 Generalized NGME and learning via the EM algorithm

Normalized Gaussian networks softly partition the input space into, say M, regions by NG functions (the gating network)

$$\mathcal{N}_h(\boldsymbol{x}; \boldsymbol{\pi}, \boldsymbol{\alpha}) = \pi_h f_h(\boldsymbol{x}; \boldsymbol{\alpha}_h) / \sum_{l=1}^{M} \pi_l f_l(\boldsymbol{x}; \boldsymbol{\alpha}_l) \qquad (h = 1, \ldots, M), \quad (1)$$

where $\pi_h > 0$, $\sum_{h=1}^{M} \pi_h = 1$, and $f_h(\boldsymbol{x}; \boldsymbol{\alpha}_h) = \phi_h(\boldsymbol{x}; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)$ denotes the multivariate Gaussian function for input vector $\boldsymbol{x}$, with mean $\boldsymbol{\mu}_h$ and covariance

matrix $\boldsymbol{\Sigma}_h$. The local units (experts) approximate the distribution of the output $\boldsymbol{y}_j$ within the partition. The final output of the NGME network is given by the summation of these local outputs weighted by the NG functions (1):

$$f(\boldsymbol{y}_j|\boldsymbol{x}_j;\,\boldsymbol{\Psi}) = \sum_{h=1}^{M} \mathcal{N}_h(\boldsymbol{x}_j;\boldsymbol{\pi},\boldsymbol{\alpha}) f_h(\boldsymbol{y}_j|\boldsymbol{x}_j;\boldsymbol{\theta}_h), \tag{2}$$

where $\boldsymbol{\Psi}$ is the vector of all the unknown parameters and $f_h(\boldsymbol{y}_j|\boldsymbol{x}_j;\boldsymbol{\theta}_h)$ are local output densities, which are generally assumed to belong to the exponential family of densities [1, 8]. The unknown parameter vector $\boldsymbol{\Psi}$ can be estimated by the maximum likelihood approach via the EM algorithm [1]. In contrast to the ME networks [8], the learning of NGME networks does not require both the selection of a learning rate and the iterative inner loop in the EM algorithm [1, 4, 8]. Under the independence assumption, $f_h(\boldsymbol{x}_j;\boldsymbol{\alpha}_h)$ in (1) can be written as

$$f_h(\boldsymbol{x}_j;\boldsymbol{\alpha}_h) = \prod_{i=1}^{q} g_h(x_{1ij})\phi_h(\boldsymbol{x}_{2j};\boldsymbol{\mu}_h,\boldsymbol{\Sigma}_h) = \prod_{i=1}^{q}\prod_{v=1}^{n_i} \lambda_{hiv}^{\delta(x_{1ij},v)} \phi_h(\boldsymbol{x}_{2j};\boldsymbol{\mu}_h,\boldsymbol{\Sigma}_h), \tag{3}$$

where the h-th conditional density of the i-th categorical variable x_{1ij} $(i = 1,\ldots,q)$ in $\boldsymbol{x}_{1j}$, $g_h(x_{1ij})$, is given by a multinomial distribution consisting of one draw on n_i distinct values with probabilities $\lambda_{hi1},\ldots,\lambda_{hin_i}$, and where $\lambda_{hin_i} = 1-\sum_{l=1}^{n_i-1}\lambda_{hil}$ and $\delta(x_{1ij},v) = 1$ if $x_{1ij} = v$ and is zero otherwise $(v = 1,\ldots,n_i)$. The vector of unknown parameters $\boldsymbol{\alpha}_h$ thus consists of λ_{hiv} $(i = 1,\ldots,q;\ v = 1,\ldots,n_i-1)$, and the elements of $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}_h$ $(h = 1,\ldots,M)$.

To apply the EM algorithm to the generalized NGME networks, we introduce the indicator variables z_{hj}, where z_{hj} is one or zero according to whether $\boldsymbol{y}_j$ belongs or does not belong to the hth expert [4]. On the $(k+1)$th iteration, the E-step involves the calculation of $\tau_{hj}^{(k)}$

$$\tau_{hj}^{(k)} = \mathrm{pr}_{\boldsymbol{\Psi}^{(k)}}\{Z_{hj} = 1|\boldsymbol{y}_j,\boldsymbol{x}_j\} = \frac{\pi_h^{(k)} f_h(\boldsymbol{x}_j;\boldsymbol{\alpha}_h^{(k)}) f_h(\boldsymbol{y}_j|\boldsymbol{x}_j;\boldsymbol{\theta}_h^{(k)})}{\sum_{l=1}^{M}\pi_l^{(k)} f_l(\boldsymbol{x}_j;\boldsymbol{\alpha}_l^{(k)}) f_l(\boldsymbol{y}_j|\boldsymbol{x}_j;\boldsymbol{\theta}_l^{(k)})} \tag{4}$$

for $h = 1,\ldots,M$, with $f_h(\boldsymbol{x}_j;\boldsymbol{\alpha}_h^{(k)})$ is given by (3) based on the current estimate $\boldsymbol{\alpha}_h^{(k)}$. In the M-step, the updated estimates of $\boldsymbol{\Psi}$ are obtained as follows:

$$\pi_h^{(k+1)} = \sum_{j=1}^{n}\tau_{hj}^{(k)}/n, \quad \lambda_{hiv}^{(k+1)} = \frac{\sum_{j=1}^{n}\tau_{hj}^{(k)}\delta(x_{1ij},v)}{\sum_{j=1}^{n}\tau_{hj}^{(k)}},$$

$$\boldsymbol{\mu}_h^{(k+1)} = \frac{\sum_{j=1}^{n}\tau_{hj}^{(k)}\boldsymbol{x}_{2j}}{\sum_{j=1}^{n}\tau_{hj}^{(k)}}, \quad \boldsymbol{\Sigma}_h^{(k+1)} = \frac{\sum_{j=1}^{n}\tau_{hj}^{(k)}(\boldsymbol{x}_{2j}-\boldsymbol{\mu}_h^{(k+1)})(\boldsymbol{x}_{2j}-\boldsymbol{\mu}_h^{(k+1)})^T}{\sum_{j=1}^{n}\tau_{hj}^{(k)}}.$$

For binary classification problems, $f_h(y_j|\boldsymbol{x}_j;\boldsymbol{\theta}_h)$ are assumed to be Bernoulli distribution of possible binary outcomes of "failure" and "success" [8]. That is,

$$f_h(y_j|\boldsymbol{x}_j;\boldsymbol{\theta}_h) = \left(\frac{\exp(\boldsymbol{w}_h^T\boldsymbol{x}_j)}{1+\exp(\boldsymbol{w}_h^T\boldsymbol{x}_j)}\right)^{y_j}\left(\frac{1}{1+\exp(\boldsymbol{w}_h^T\boldsymbol{x}_j)}\right)^{1-y_j}, \tag{5}$$

Table 1. Leave-one-out error rates for the breast cancer data

Method	Error rate
NGME network on continuous variables	29.5%
independence model on mixed variables	19.2%

where $\boldsymbol{\theta}_h = \boldsymbol{w}_h$. For notational convenience, we still present the mixed-mode input vector as $\boldsymbol{x}_j$ in (5). Indeed, the categorical variables are replaced by n_i-1 dummy variables and contribute to the local output via the linear predictor $\eta_{hj} = \boldsymbol{w}_h^T \boldsymbol{x}_j$; see [4]. The updated estimate of $\boldsymbol{\theta}_h^{(k+1)}$ is obtained by solving

$$\sum_{j=1}^{n} \tau_{hj}^{(k)} \partial \log f_h(\boldsymbol{y}_j|\boldsymbol{x}_j;\boldsymbol{\theta}_h)/\partial\boldsymbol{\theta}_h = \sum_{j=1}^{n} \tau_{hj}^{(k)} \left(y_j - \frac{\exp(\boldsymbol{w}_h^T \boldsymbol{x}_j)}{1+\exp(\boldsymbol{w}_h^T \boldsymbol{x}_j)} \right) \boldsymbol{x}_j = \boldsymbol{0}$$

for $h = 1, \ldots, M$, which are M sets of nonlinear equations each with unknown parameter vector $\boldsymbol{w}_h$.

3 A real example: Breast cancer data

We illustrate the method using an example of classifying breast cancer patients on the basis of the gene expression-profile vector of tumor samples and categorical variables of patient's clinical characterisitcs. The original data set [9] consists of 5000 gene expression profiles and 6 binary variables of clinical indicators from 78 sporadic lymph-node-negative breast cancer patients. With these patients, 44 remained metastasis free after a period of more than 5 years (good prognosis) and 34 patients had developed distant metastases within 5 years (poor prognosis). In this study, we work on the data set with 6 binary variables of clinical indicators and 5 continuous variables representing the top 5 "metagenes" ranked in terms of the likelihood ratio statistic described in [10]. We first apply the NGME network of [1] on the continuous variables to classify the patients into good and poor prognosis subgroups; see Eqt. (5). This preliminary analysis provides the initial estimates and the determination of the number of experts M for the generalized NGME network. In addition, the improvement of the generalized NGME network by using additional binary clinical indicators can be assessed. Such evaluation is based on the misclassification error rate using the "leave-one-out" method for cross-validation. The number of experts M is determined based on a frequentist analog of the "worth index" on model selection [11]. A NGME network with $M = 2$ experts is selected. The leave-one-out error rate is provided in Table 1. We then apply the generalized NGME networks to classify the patients, using the independence model, on the mixed feature data. From Table 1, it can be seen that the generalized NGME network significantly reduce the error rate by using additional binary clinical indicators.

4 Discussion

We have extended the NGME network to incorporate the independence model for tackling problems with mixed feature data. Although the independence assumption is likely to be unrealistic for many problems, it often performs surprisingly well in practice as a way of handling problems with mixed feature data [6, 7]. One important reason is that the NAIVE method usually requires fewer parameters to be estimated and hence tends to have a lower variance for the estimates [6].

The error rates in Table 1 have been considered in a relative sense. However, caution should be exercised in interpreting these rates in an absolute sense. This is because the metagenes in the data set are determined using the expression profiles from the 78 cancer patients. Thus, the misclassification error rate is calculated without allowance for the selection bias [12]. The error rates given in Table 1 should therefore be interpreted as apparent error rates. An "external" cross-validation can be adopted to correct for the bias in estimating the error of a prediction rule [12].

References

1. Xu, L., Jordan, M.I., Hinton, G.E.: An alternative model for mixtures of experts. In: Cowan, J.D., Tesauro, G., Alspector, J. (eds.): Adv. in Neural Inf. Proc. Systems 7. MIT Press, Cambridge, Massachusetts (1995) 633–640
2. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. Ser. B **39** (1977) 1–38
3. Moody, J., Darken, C.J.: Fast learning in networks of locally-tuned processing units. Neural Comput. **1** (1989) 281–294
4. Ng, S.K., McLachlan, G.J.: Using the EM algorithm to train neural networks: Misconceptions and a new algorithm for multiclass classification. IEEE T. Neural Networ. **15** (2004) 738–749
5. McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
6. Hand, D.J., Yi, K.: Idiot's Bayes – not so stupid after all? Int. Stat. Rev. **69** (2001) 385–398
7. Titterington, D.M., Murray, G.D., Murray, L.S., et al.: Comparison of discrimination techniques applied to a complex data set of head injured patients. J. Roy. Stat. Soc. Ser. A **144** (1981) 145–175
8. Jordan, M.I., Jacobs, R.A.: Hierarchical mixtures of experts and the EM algorithm. Neural Comput. **6** (1994) 181–214
9. van't Veer, L.J., Dai, H., van de Vijver, M.J., et al.: Gene expression profiling predicts clinical outcomes of breast cancer. Nature **415** (2002) 530–536
10. McLachlan, G.J., Bean, R.W., Peel, D.: A mixture model-based approach to the clustering of microarray expression data. Bioinformatics **18** (2002) 413–422
11. Ng, S.K., McLachlan, G.J., Yau, K.K.W., Lee, A.H.: Modelling the distribution of ischaemic stroke-specific survival time using an EM-based mixture approach with random effects adjustment. Statist. Med. **23** (2004) 2729–2744
12. Ambroise, C., McLachlan, G.J.: Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. USA **99** (2002) 6562–6566