Neurocomputing 73 (2010) 2989–2997

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Variational inference for Student-*t* MLP models

Hang T. Nguyen *, Ian T. Nabney

The Non-linearity and Complexity Research Group (NCRG), School of Engineering and Applied Science, Aston University, Aston Triangle, Birmingham B4 7ET, UK

ARTICLE INFO

Article history:
Received 5 October 2009
Received in revised form
21 July 2010
Accepted 26 July 2010
Communicated by C. Fyfe
Available online 26 August 2010

Keywords:
Variational inference
Student-*t* noise
Multi-layer perceptrons
EM algorithm
Forecast

ABSTRACT

This paper presents a novel methodology to infer parameters of probabilistic models whose output noise is a Student-*t* distribution. The method is an extension of earlier work for models that are linear in parameters to non-linear multi-layer perceptrons (MLPs). We used an EM algorithm combined with variational approximation, an evidence procedure, and an optimisation algorithm. The technique was tested on two forecasting applications. The first one is a synthetic dataset and the second is gas forward contract prices data from the UK energy market. The results showed that forecasting accuracy is significantly improved by using Student-*t* noise models.

1. Introduction

In forecasting models, we generally assume that the data are corrupted by noise:

$$y_t = f(\mathbf{x}_t) + \varepsilon_t,$$

where ε_t is a zero-mean probability distribution. Normally, the noise is assumed to be Gaussian distribution either because of arguments derived from the central limit theorem or just to simplify calculations. For example, the log likelihood of a Gaussian noise model is a quadratic function of the output variables. This leads to the fact that in the training process, we can easily estimate the maximum likelihood solution using optimisation algorithms. Software and frameworks for training machine learning models such as radial basis functions (RBF), MLP, and linear regression (LR) with Gaussian noise can be found in [1]. Conversely, other noise models are much less tractable. So why use the Student-*t* distribution?

In our previous work [2,3], we used models with Gaussian noise to forecast gas and electricity forward prices in the UK energy market. In these experiments, the kurtosis, which is a measure of how outlier-prone a distribution is, of the residuals (i.e. the different between target and output of forecasting model) is between 16 and 17: the kurtosis of the Gaussian distribution is 3. Furthermore, $P(\mu-3\sigma < r < \mu+3\sigma) \approx 0.982$, where μ and σ are the mean and standard derivation of the residual, respectively. The equivalent probability for a Gaussian distribution is 0.997; therefore, the residual distribution has heavy tails. This means that the residual distributions are much more outlier-prone than the Normal distribution. The large number of outliers can make the training process unreliable and error bar estimates inaccurate, because Gaussians are sensitive to outliers. It is clear that these data are not modelled well by a Gaussian distribution as has often been noted for financial data.

As a consequence, a Student-*t* distribution can be considered as a good alternative to a Gaussian because it is a fat-tailed distribution and is more robust. Moreover, the Student-*t* distribution family contains the Normal distribution as a special case.

There are several previous studies of inference with Student-*t* models. Tipping and Lawrence proposed a framework for training an RBF model with fixed basis functions [4]. This study is a fully Bayesian treatment based on a variational approximation framework. A variational inference scheme was also used for unsupervised learning with mixture models: Bishop and Svensén presented an algorithm for automatically determining the number of components in a mixture of *t*-distribution using a Bayesian variational framework [5]. In order to obtain a tractable solution, it was assumed that the latent variables are independent, and thus posterior distributions of latent variables can be factorized. This means that the algorithm does not capture correlations among the latent variables. Archambeau and Verleysen introduced a new variational Bayesian learning algorithm for Student-*t* mixture models, in which they removed the assumption of variable independence [6]. Numerical experiments showed that their model had a greater robustness to outliers than Bishop and Svensén's method in [5].

* Corresponding author. Tel.: +44 121 257 7718; fax: +44 121 204 3685.
E-mail addresses: thihangn@aston.ac.uk, nthangd98vt2@yahoo.com (H.T. Nguyen).

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.07.009

This paper presents a novel methodology to infer parameters of Student-t probabilistic models. This methodology for MAP estimation is an extension of the technique introduced by Tipping and Lawrence [4], in which models are assumed to be linear in parameters. Both approaches are based on a variational approximation. The main advantage of our method is that it is not limited to models whose output is linearly dependent on model parameters. On the other hand, our approach provides only MAP estimates of parameters while Tipping and Lawrence give a fully Bayesian treatment in which predictions are made by integrating out all the parameters apart from those defining the t-distribution, which are optimised. Thus, although our algorithm can be applied to models that are linear in parameters, we would not expect it to outperform Tipping and Lawrence, so our discussion focusses on the MLP.

This paper is organised as follows. In Section 2, Student-t noise models are presented. Section 3 describes our inference technique for MLPs. Numerical results on two datasets are given in Section 4. Section 5 discusses some conclusions.

2. Student-t noise model

We assume that the output data are corrupted by noise with a Student-t distribution:

$$y_t = f(\mathbf{x}_t,\boldsymbol{\omega}) + \varepsilon_t,$$

where ε_t is a Student-t noise process, and $f(\mathbf{x}_t)$ is the output function of a forecast model, which can be a multi-layer perceptron (MLP), radial basis function (RBF), or linear regression (LR). In the case of MLP models, the output is non-linear in the parameters. Conversely, the output is linear in parameters when the model is LR or RBF. We are not investigating the case where the independent variables $\mathbf{x}_t$ are also noisy.

The Student-t distribution can be considered as a mixture of an infinite number of zero-mean Gaussians with different variances:

$$\begin{aligned} p(\varepsilon_t|c,d) &= \int_0^\infty p(\varepsilon_t|\beta_t)p(\beta_t|c,d)\,d\beta_t \\ &= \frac{d^c}{\Gamma(c)}\left(\frac{1}{2\pi}\right)^{1/2}\left[d+\frac{\varepsilon_t^2}{2}\right]^{-c-1/2}\Gamma(c+1/2), \end{aligned} \tag{1}$$

where

$$p(\varepsilon_t|\beta_t) = N(\varepsilon_t|0,\beta_t^{-1}),$$

$$p(\beta_t|c,d) = Gamma(\beta_t|c,d) = \frac{d^c}{\Gamma(c)}\beta_t^{c-1}\exp(-\beta_t d).$$

The mixture weight for a given β_t is specified by the Gamma distribution $p(\beta_t|c,d)$. $v=2c$ is called the "number of degrees of freedom" and $\sigma=\sqrt{d/c}$ is the scale parameter of the distribution. The degrees-of-freedom parameter v can be considered as a robustness tuning parameter [6]. When v tends to infinity, this distribution converges to a Gaussian. Therefore, the Student-t noise model still contains the Gaussian as a special case when v is very large.

3. MAP estimation for MLPs

The aim of our approach is to find maximum a posterior (MAP) estimates of network and noise model parameters. MAP estimation is not a fully Bayesian treatment because it finds the optimal parameters of the models instead of integrating over all unknown parameters. This is equivalent to the type-II maximum likelihood method [7].

In this paper, we will describe an EM algorithm for training a model with a Student-t noise model. This training framework can be use for both "non-linear in parameters" models and "linear in parameters" models.

Given a dataset $\mathbf{D}=\{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_T,y_T)\}$, our goal is to optimise parameters of a predictive model (i.e. MLP, LR or RBF) using MAP. To simplify the notation, let $\boldsymbol{\Omega}=\{\boldsymbol{\omega},c,d,\boldsymbol{\alpha}\}$ be the set of parameters/hyperparameters of the model and noise. The posterior density of the parameters given a dataset $\mathbf{D}$ is given by

$$p(\boldsymbol{\Omega}|\mathbf{D}) = \frac{p(\mathbf{D}|\boldsymbol{\Omega})p(\boldsymbol{\Omega})}{p(\mathbf{D})},$$

where $p(\mathbf{D}|\boldsymbol{\Omega})$ is the dataset likelihood, $p(\boldsymbol{\Omega})$ is the prior, and $p(\mathbf{D})$ is evidence. Because the denominator does not affect the MAP solution, we can ignore this term: $p(\boldsymbol{\Omega}|\mathbf{D}) \propto p(\mathbf{D}|\boldsymbol{\Omega})p(\boldsymbol{\Omega})$. The likelihood and the prior are given by

$$p(\mathbf{D}|\boldsymbol{\Omega}) = p(\mathbf{D}|\boldsymbol{\omega},c,d) = \prod_{t=1}^{T} p(y_t|\mathbf{x}_t,\boldsymbol{\omega},c,d),$$

$$p(y_t|\mathbf{x}_t,\boldsymbol{\Omega}) = \frac{d^c}{\Gamma(c)}\left(\frac{1}{2\pi}\right)^{1/2}\left[d+\frac{(y_t-f(\mathbf{x}_t,\boldsymbol{\omega}))^2}{2}\right]^{-c-1/2}\Gamma(c+1/2),$$

$$p(\boldsymbol{\Omega}) = p(\boldsymbol{\omega}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})p(c,d). \tag{2}$$

The weight prior $p(\boldsymbol{\omega}|\boldsymbol{\alpha})$ is a Gaussian. It is helpful to generalise the hyperparameter $\boldsymbol{\alpha}$ to multiple hyperparameters $\alpha_1,\ldots,\alpha_M$ corresponding to groups of weights $\mathcal{W}_1,\ldots,\mathcal{W}_M$. In theory, we can create groupings of the weights in any way that we want. However, weights in an MLP are normally divided into four groups: first-layer weights, first-layer biases, second-layer weights, and second-layer biases. In addition, the first-layer weights can be also divided into several groups: weights fanning out from a input variable are associated to a separate group. The latter grouping approach relates to automatic relevance determination (ARD) [8] and is used in our experiments. Denote group dimensions by $\mathbf{W}_1,\ldots,\mathbf{W}_M$ corresponding to the groups $\mathcal{W}_1,\ldots,\mathcal{W}_M$. Thus the dimension of $\boldsymbol{\omega}$ is $\mathbf{W}=\sum_{m=1}^{M}\mathbf{W}_m$.

$$p(\boldsymbol{\omega}|\boldsymbol{\alpha}) = \prod_{m=1}^{M} N(\mathcal{W}_m|0,\alpha_m^{-1}) = \prod_{m=1}^{M}\left(\frac{\alpha_m}{2\pi}\right)^{\mathbf{W}_m/2}\exp\left[\sum_{m=1}^{M}\left(-\frac{\alpha_m}{2}\sum_{\omega\in\mathcal{W}_m}\omega^2\right)\right]. \tag{3}$$

There are many possible choices for the densities $p(\boldsymbol{\alpha})$ and $p(c,d)$, but for simplicity we assume that they are uniform distributions. Therefore, they will be ignored in the subsequent analysis. Hence

$$\begin{aligned} \log p(\boldsymbol{\Omega}|\mathbf{D}) &\propto \log[p(\mathbf{D}|\boldsymbol{\Omega})p(\boldsymbol{\Omega})] \\ &= T c\log d + T\log\frac{\Gamma(c+1/2)}{\Gamma(c)} - \frac{\mathbf{W}+T}{2}\log 2\pi \\ &\quad -\left(c+\frac{1}{2}\right)\sum_{t=1}^{T}\log\left[d+\frac{(y_t-f(\mathbf{x}_t,\boldsymbol{\omega}))^2}{2}\right] \\ &\quad +\sum_{m=1}^{M}\left(\frac{\mathbf{W}_m}{2}\log\alpha_m\right) - \sum_{m=1}^{M}\left(\frac{\alpha_m}{2}\sum_{\omega\in\mathcal{W}_m}\omega^2\right). \end{aligned}$$

3.1. Variational approximation

The Student-t distribution of each observation y_t can be considered as a mixture of an infinite number of zero-mean Gaussians with inverse variance β_t. Let $\boldsymbol{\beta}=\{\beta_1,\beta_2,\ldots,\beta_T\}$; then

$$p(\mathbf{D}|\boldsymbol{\Omega}) = \int_0^\infty p(\mathbf{D},\boldsymbol{\beta}|\boldsymbol{\Omega})\,d\boldsymbol{\beta} = \int_0^\infty p(\mathbf{D}|\boldsymbol{\beta},\boldsymbol{\Omega})p(\boldsymbol{\beta}|\boldsymbol{\Omega})\,d\boldsymbol{\beta}, \tag{4}$$