Multilogistic Regression by Product Units

P. A. Gutiérrez
University of Córdoba, Spain

C. Hervás
University of Córdoba, Spain

F. J. Martínez-Estudillo
INSA – ETEA, Spain

M. Carbonero
INSA – ETEA, Spain

INTRODUCTION

Multi-class pattern recognition has a wide range of applications including handwritten digit recognition (Chiang, 1998), speech tagging and recognition (Athanaselis, Bakamidis, Dologlou, Cowie, Douglas-Cowie & Cox, 2005), bioinformatics (Mahony, Benos, Smith & Golden, 2006) and text categorization (Massey, 2003). This chapter presents a comprehensive and competitive study in multi-class neural learning which combines different elements, such as multilogistic regression, neural networks and evolutionary algorithms.

The Logistic Regression model (LR) has been widely used in statistics for many years and has recently been the object of extensive study in the machine learning community. Although logistic regression is a simple and useful procedure, it poses problems when is applied to a real-problem of classification, where frequently we cannot make the stringent assumption of additive and purely linear effects of the covariates. A technique to overcome these difficulties is to augment/replace the input vector with new variables, basis functions, which are transformations of the input variables, and then to use linear models in this new space of derived input features. Methods like sigmoidal feed-forward neural networks (Bishop, 1995), generalized additive models (Hastie & Tibshirani, 1990), and PolyMARS (Kooperberg, Bose & Stone, 1997), which is a hybrid of Multivariate Adaptive Regression Splines (MARS) (Friedman, 1991) specifically designed to handle classification problems, can all be seen as different non-linear basis function models. The major drawback of these approaches is stating the typology and the optimal number of the corresponding basis functions.

Logistic regression models are usually fit by maximum likelihood, where the Newton-Raphson algorithm is the traditional way to estimate the maximum likelihood a-posteriori parameters. Typically, the algorithm converges, since the log-likelihood is concave. It is important to point out that the computation of the Newton-Raphson algorithm becomes prohibitive when the number of variables is large.

Product Unit Neural Networks, PUNN, introduced by Durbin and Rumelhart (Durbin & Rumelhart, 1989), are an alternative to standard sigmoidal neural networks and are based on multiplicative nodes instead of additive ones.

BACKGROUND

In the classification problem, measurements x_i, $i = 1,2,...,k$, are taken on a single individual (or object), and the individuals are to be classified into one of J classes on the basis of these measurements. It is assumed that J is finite, and the measurements x_i are random observations from these classes. A training sample $D = \{(\mathbf{x}_n, \mathbf{y}_n); n = 1, 2,...,N\}$ is available, where $\mathbf{x}_n = (x_{1n},...,x_{kn})$ is the vector of measurements taking values in $\Omega \subset \mathbb{R}^k$, and $\mathbf{y}_n$ is the class level of the nth individual. In this chapter, we will adopt the common technique of representing the class levels using a "1-of-J" encoding vector $\mathbf{y} = (y^{(1)}, y^{(2)},...,y^{(J)})$, such as $y^{(l)} = 1$ if $\mathbf{x}$ corresponds to an example belonging to class l and $y^{(l)} = 0$ otherwise. Based on the training sample, we wish to find a decision function $C : \Omega \rightarrow \{1,2,...,J\}$ for classifying the individuals. In other words, C provides a partition, say $D_1,D_2,...,D_J$, of Ω, where D_l corresponds to the lth class, $l = 1,2,...,J$,

and measurements belonging to D_l will be classified as coming from the lth class. A misclassification occurs when a decision rule C assigns an individual (based on measurements vector) to a class j when it is actually coming from a class $l \neq j$.

To evaluate the performance of the classifiers we can define the Correctly Classified Rate by

$$CCR = \frac{1}{N}\sum_{n=1}^{N} I(C(\mathbf{x}_n) = \mathbf{y}_n),$$

where $I(.)$ is the zero-one loss function. A good classifier tries to achieve the highest possible CCR in a given problem.

Suppose that the conditional probability that $\mathbf{x}$ belongs to class l verifies:

$$p\left(y^{(l)} = 1 \middle| \mathbf{x}\right) > 0,\ l = 1,2,...,J,\ \mathbf{x} \in \Omega,$$

and set the function:

$$f_l(\mathbf{x}, \boldsymbol{\theta}_l) = \log \frac{p\left(y^{(l)} = 1 \middle| \mathbf{x}\right)}{p\left(y^{(J)} = 1 \middle| \mathbf{x}\right)},\ l = 1,2,...,J,\ \mathbf{x} \in \Omega$$

where $\boldsymbol{\theta}_l$ is the weight vector corresponding to class l and $f_J(\mathbf{x}, \boldsymbol{\theta}_J) \equiv 0$. Under a multinomial logistic regression, the probability that $\mathbf{x}$ belongs to class l is then given by

$$p\left(y^{(l)} = 1 \middle| \mathbf{x}, \boldsymbol{\theta}\right) = \frac{\exp f_l\left(\mathbf{x}, \boldsymbol{\theta}_l\right)}{\sum_{j=1}^{J} \exp f_j\left(\mathbf{x}, \boldsymbol{\theta}_j\right)},\ l = 1,2,...,J$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_{J-1})$.

The classification rule coincides with the optimal Bayes' rule. In other words, an individual should be assigned to the class which has the maximum probability, given the vector measurement $\mathbf{x}$:

$$C(\mathbf{x}) = \hat{l},$$

where

$$\hat{l} = \arg\max_l f_l(\mathbf{x}, \hat{\boldsymbol{\theta}}_l),\ \text{for } l = 1,...,J.$$

On the other hand, because of the normalization condition we have:

$$\sum_{l=1}^{J} p\left(y^{(l)} = 1 \middle| \mathbf{x}, \boldsymbol{\theta}\right) = 1,$$

and the probability for one of the classes (in the proposed case, the last) need not be estimated (observe that we have considered $f_J(\mathbf{x}, \boldsymbol{\theta}_J) \equiv 0$).

MULTILOGISTIC REGRESSION AND PRODUCT UNIT NEURAL NETWORKS

Multilogistic Regression by using Linear and Product-Unit models (MLRPU) overcomes the nonlinear effects of the covariates by proposing a multilogistic regression model based on the combination of linear and product-unit models, where the nonlinear basis functions of the model are given by the product of the inputs raised to arbitrary powers. These basis functions express the possible strong interactions between the covariates, where the exponents are not fixed and may even take real values. In fitting the proposed model, the non-linearity of the PUNN implies that the corresponding Hessian matrix is generally indefinite and the likelihood has more local maximum. This reason justifies the use of an alternative heuristic procedure to estimate the parameters of the model.

Non-Linear Model Proposed

The general expression of the proposed model is given by:

$$f_l(\mathbf{x}, \boldsymbol{\theta}_l) = \alpha_0^l + \sum_{i=1}^{k} \alpha_i^l x_i + \sum_{j=1}^{m} \beta_j^l \prod_{i=1}^{k} x_i^{w_{ji}},\quad l = 1,2,...,J-1$$

where

$\boldsymbol{\theta}_l = (\boldsymbol{\alpha}^l, \boldsymbol{\beta}^l, \mathbf{W})$,

$\alpha^l = (\alpha_0^l, \alpha_1^l, ..., \alpha_k^l)$,

$\beta^l = \left(\beta_1^l, ..., \beta_m^l\right)$ and

$\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_m)$,