Neural Networks 97 (2018) 74–91

Contents lists available at ScienceDirect

Neural Networks

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

Stretchy binary classification

Kar-Ann Toh [a,*], Zhiping Lin [b], Lei Sun [c], Zhengguo Li [d]

[a] School of Electrical and Electronic Engineering, Yonsei University, Seoul, 03722, Republic of Korea
[b] School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore
[c] School of Information and Electronics, Beijing Institute of Technology, 100081, PR China
[d] Institute for Infocomm Research, 1 Fusionopolis Way, #21-01, 138632, Singapore

HIGHLIGHTS

- Proposed a novel cost function for counting the samples that are misclassified.
- Conjectured an analytic solution to a constrained p-norm minimization problem.
- Linkage of the proposed formulation to two existing classifiers.
- Provided variance analysis for the proposed analytic solution.
- Extensive experiments with comparison to state-of-the-arts.

ARTICLE INFO

Article history:
Received 12 January 2017
Received in revised form 8 August 2017
Accepted 28 September 2017
Available online 10 October 2017

Keywords:
Pattern classification
Parameter learning
Sparse estimation

ABSTRACT

In this article, we introduce an analytic formulation for compressive binary classification. The formulation seeks to solve the least ℓ^p-norm of the parameter vector subject to a classification error constraint. An analytic and stretchable estimation is conjectured where the estimation can be viewed as an extension of the pseudoinverse with left and right constructions. Our variance analysis indicates that the estimation based on the left pseudoinverse is unbiased and the estimation based on the right pseudoinverse is biased. Sparseness can be obtained for the biased estimation under certain mild conditions. The proposed estimation is investigated numerically using both synthetic and real-world data.

1. Introduction

Pattern classification has a wide range of applications spanning fields in engineering, financial, social, medical and life sciences. The approaches to classifier design can generally be divided into those by *deterministic* means and those by *probabilistic* means, even though their underlying mechanisms can be linked theoretically (Duda, Hart, & Stork, 2001; Hastie, Tibshirani, & Friedman, 2001). Under the supervised learning paradigm, the *deterministic approach* (also known as non-probabilistic approach) formulates the classification problem as to minimize the misclassification rate (or to maximize the classification accuracy) based on a given set of training samples. The adopted mathematical model for data learning can be treated as the problem of input–output data mapping under the paradigm of neural networks (Faris, Aljarah, & Mirjalili, 2016; Funahashi, 1989; Hornik, Stinchcombe, & White, 1990; Huang & Du, 1999, 2008; Sprecher, 1993; Zhang, Zhang, Lok, & Lyu, 2007), which can be traced to neural and information decision sciences (Huang & Jiang, 2012; Nickerson, 1972; Proctor & Cho, 2006).

The *probabilistic approach* capitalizes on the *Bayesian decision theory* which quantifies the tradeoffs among various classification decisions using probability and the cost that accompany such decisions (Duda et al., 2001). Depending on the assumptions regarding the probability density of data and the design model prior, various methods have been developed to solve the decision formulation. The *maximum-likelihood estimation* views the parameters of the data density function as fixed but unknown quantities where the best estimation of their value is defined to be the one that maximizes the probability of obtaining the samples actually observed. The *Bayesian estimation* views the parameters as random variables of certain known type of distribution. Such information is converted to posterior density by the inference of parameter values from observed data samples. Due to the often limited availability of data and its limited representativeness for learning, *structural risk minimization* adopts an inductive principle for learning model selection. Generally, the method adopts a model of capacity control and seeks a trade-off between the hypothesis space complexity and

* Corresponding author.
E-mail addresses: katoh@ieee.org (K.-A. Toh), ezplin@ntu.edu.sg (Z. Lin), sunlei@bit.edu.cn (L. Sun), ezgli@i2r.a-star.edu.sg (Z. Li).

https://doi.org/10.1016/j.neunet.2017.09.015
0893-6080/

the quality of fitting the training data. From Vapnik (1998, 1999), the hypothesis space complexity is known as the VC dimension. The quality of data fitting is called the empirical error.

Apart from the classification accuracy concern, an important consideration for classifier design is the computational cost. This is in view of the often limited computing and memory facilities in contrast to the large amount of data for processing. Based on the assumption that the desired signal content is intrinsically sparse, compressed sensing (also known as sparse estimation) addresses such concern from reducing the size of estimation parameters while maintaining competitive accuracy. Grounded on the knowledge that the distance metric plays a critical role in shrinking the parameters, a typical parametric search is either based on a relevant distance metric or coupled with constraints that can produce the desired shrinkage behavior. This can be viewed as an adoption of a parametric prior with shrinking capability upon accumulation of evidence from the Bayesian perspective.

In this article, we address the problem of network-based classifier learning by adopting the deterministic approach. Different from existing methods, we conjecture an analytic and stretchable learning solution without needing an iterative search or likelihood maximization steps. Extensive numerical evaluations using synthetic and real-world data are presented to support the claim. The main contributions[1] include: (i) a novel classification error counting cost function for constraining the parametric search; (ii) a novel deterministic method for sparse network parameter estimation. This estimation is the first of its kind whereby classification network learning and compressed estimation are performed at the same time; (iii) an analysis of variance regarding the estimation bias; (iv) a linkage of the proposed formulation to two existing classifiers; and (v) an extensive numerical evaluation to illustrate the estimation mechanism and its validity. The conceivable impact of such estimation is evident from the vast application potentials in real-world classification problems particularly in this era of big data.

The organization of this article is as follows: Section 2 provides the relevant background material for subsequent development. In particular, learning based on the linear prediction network model is introduced and this is followed by a brief coverage of parameter regularization and shrinkage methods in the literature. Next, a novel error rate based classification cost function is proposed to pave the development of a classification network with stretchable parameters in Section 3. This is followed by a variance analysis in Section 4. In Section 5, a linkage of the proposed methodology to two well-known state-of-the-art classifiers is shown. A synthetic example is presented in Section 6 to illustrate the stretching behavior on representative scenarios. Subsequently, in Section 7, the proposed method is evaluated through experimentation on benchmark data sets from the literature. Finally, our concluding remarks are given in Section 8.

2. Preliminaries

2.1. Linear regression

The linear model for regression expresses its output as a linear combination of the input variables. By denoting an input vector using $\boldsymbol{x} \in \mathbb{R}^d$ and a corresponding weight vector using $\boldsymbol{\alpha} \in \mathbb{R}^d$, the linear regression model $g(\boldsymbol{\alpha}, \boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\alpha}$ can be used to learn a target output $y \in \mathbb{R}$ by minimizing the error of fit, $e = y - \boldsymbol{x}^T\boldsymbol{\alpha}$. In the generalized form (Duda et al., 2001), the input vector $\boldsymbol{x}$ can be mapped onto a transformed space $\boldsymbol{p}(\boldsymbol{x}) : \mathbb{R}^d \mapsto \mathbb{R}^D$ where the model output can be written as $g(\boldsymbol{\alpha}, \boldsymbol{x}) = \boldsymbol{p}(\boldsymbol{x})^T\boldsymbol{\alpha}$. This is also known as a linear network structure. Typically, the transformed dimension D is chosen to be larger than d such that a larger degree of freedom for weight vector (correspondingly $\boldsymbol{\alpha} \in \mathbb{R}^D$) estimation is obtained. The transformed feature vector $\boldsymbol{p}(\boldsymbol{x})$ can be considered as a basis expansion function with popular choice taking the form of *Sigmoid* (see e.g., Bishop, 1995; Guliyev & Ismailov, 2016), *Gaussian* (see e.g., Huang & Du, 2008; Poggio & Girosi, 1990; Rouhani & Javan, 2016), *Polynomial* (see e.g., Toh, Tran, & Srinivasan, 2004; Tong, 2016), and *Random Projection* (see e.g., Cao, Zhang, Luo, Yin, & Lai, 2016; Toh, 2008; Widrow, Greenblatt, Kim, & Park, 2013).

Consider a training set consisting of m samples, a popular cost function for predictor learning is the *sum of squared errors* (Duda et al., 2001; Hastie et al., 2001) given by

$$\text{SSE} = \sum_{i=1}^{m} e_i^2 = \sum_{i=1}^{m} (y_i - g(\boldsymbol{\alpha}, \boldsymbol{x}_i))^2 = (\mathbf{y} - \mathbf{P}\boldsymbol{\alpha})^T(\mathbf{y} - \mathbf{P}\boldsymbol{\alpha}), \tag{1}$$

where $\mathbf{P} = [\boldsymbol{p}(\boldsymbol{x}_1), \ldots, \boldsymbol{p}(\boldsymbol{x}_m)]^T \in \mathbb{R}^{m\times D}$ packs the transformed input vectors in matrix form. When $\mathbf{P}$ has a full rank, then minimization of (1) gives a solution in analytic form:

$$\boldsymbol{\alpha} = (\mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}^T\mathbf{y}. \tag{2}$$

This solution, which is particularly useful when the system is overdetermined (i.e., $m > D$), is often referred to as the *least squares error* (LSE) solution for *regression* applications.

2.2. Regularization and restricting the feasible set

There are two common ways to deal with a possible singularity of the covariance matrix $\mathbf{P}^T\mathbf{P}$ in (2). The first way, called regularization, is to constrain the parameter size (i.e., having $\|\boldsymbol{\alpha}\|_2^2 \leqslant t$, $t \in \mathbb{R}^+$ where $\|\boldsymbol{\alpha}\|_2^2 := \sum_{j=0}^{D-1}\alpha_j^2 = \boldsymbol{\alpha}^T\boldsymbol{\alpha}$) during minimization. Such minimization problem is often posed in an unconstrained form with inclusion of a regularization penalty factor $\lambda \in \mathbb{R}$ which controls the strength of the penalty:

$$\text{SSE}_{ridge}(\boldsymbol{\alpha}) = \frac{1}{2}\sum_{i=1}^{m}\left(y_i - \boldsymbol{p}(\boldsymbol{x}_i)^T\boldsymbol{\alpha}\right)^2 + \frac{\lambda}{2}\|\boldsymbol{\alpha}\|_2^2. \tag{3}$$

Solving (3) results in an analytic solution given by

$$\boldsymbol{\alpha} = (\mathbf{P}^T\mathbf{P} + \lambda\mathbf{I})^{-1}\mathbf{P}^T\mathbf{y}. \tag{4}$$

This solution comes with a scaled identity matrix $\lambda\mathbf{I}$ which stabilizes the inverse operation. This minimization is commonly known as the *ridge regression* in the literature (Hastie et al., 2001; Hoerl & Kennard, 1970; Tikhonov, 1963).

In the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996, 2011), an ℓ^1-norm is adopted as the metric for constraining the size of parameters, giving

$$\text{SSE}_{lasso}(\boldsymbol{\alpha}) = \sum_{i=1}^{m}\left(y_i - \boldsymbol{p}(\boldsymbol{x}_i)^T\boldsymbol{\alpha}\right)^2 + \lambda\|\boldsymbol{\alpha}\|_1, \tag{5}$$

where $\|\boldsymbol{\alpha}\|_1 := \sum_{j=0}^{D-1}|\alpha_j|$. By replacing the penalty metric with a p-norm given by

$$\ell^p : \qquad \|\boldsymbol{\alpha}\|_p := \left(\sum_{i=0}^{D-1}|\alpha_i|^p\right)^{1/p}, \tag{6}$$

the minimization is called the *bridge regression* (Frank & Friedman, 1993):

$$\text{SSE}_{bridge}(\boldsymbol{\alpha}) = \sum_{i=1}^{m}\left(y_i - \boldsymbol{p}(\boldsymbol{x}_i)^T\boldsymbol{\alpha}\right)^2 + \lambda\|\boldsymbol{\alpha}\|_p^p. \tag{7}$$

[1] Some of the preliminary ideas in this paper has been presented at the IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (Toh, 2015). The current manuscript extends beyond the preliminary work in both theoretical findings and experimental observations.