Neurocomputing 149 (2015) 836–846

Contents lists available at ScienceDirect

Neurocomputing

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

Kernel ridge regression for general noise model with its application

Shiguang Zhang [a,b], Qinghua Hu [c], Zongxia Xie [c], Jusheng Mi [a,d,*]

[a] College of Mathematics and Information Science, Hebei Normal University, Shijiazhuang, Hebei 050024, China
[b] College of Mathematics and Computer Science, Hengshui University, Hengshui, Hebei 053000, China
[c] School of Computer Science and Technology, Tianjin University, Tianjin 300072, China
[d] Hebei Key Laboratory of Computational Mathematics and Applications, Shijiazhuang, Hebei 050024, China

ARTICLE INFO

Article history:
Received 29 May 2013
Received in revised form
16 July 2014
Accepted 25 July 2014
Communicated by Christos Dimitrakakis
Available online 23 August 2014

Keywords:
Kernel ridge regression
Noise model
Loss function
Equality constraints
Short-term wind speed prediction

ABSTRACT

The classical ridge regression technique makes an assumption that the noise is Gaussian. However, it is reported that the noise models in some practical applications do not satisfy Gaussian distribution, such as wind speed prediction. In this case, the classical regression techniques are not optimal. So we derive an optimal loss function and construct a new framework of kernel ridge regression technique for general noise model (*N-KRR*). The Augmented Lagrangian Multiplier method is introduced to solve *N-KRR*. We test the proposed technique on artificial data and short-term wind speed prediction. Experimental results confirm the effectiveness of the proposed model.

1. Introduction

Regression techniques are widely used in stock price prediction, marketing analysis, power consuming prediction, wind speed forecasting, etc. Although these techniques have been successfully applied in various domains, new challenges and problems are still reported in some practical applications. This topic is attracting much attention from application and research areas these years [1–3].

We first introduce some notations. Given a set of training data $D_l=\{(x_1,y_1),(x_2,y_2),\dots,(x_l,y_l)\}$, where $x_i\in R^L, y_i\in R, i=1,2,\dots,l$. A multivariate linear model is $f(x)=\varpi^T\cdot x+b$, where $\varpi\in R^L$, $b\in R$. The task of Ridge regression is to learn the parameter vector ϖ and parameter b, by minimizing the objective function

$$g_{RR}=\frac{1}{2}\cdot\omega^T\cdot\omega+C\cdot\sum_{i=1}^{l}(y_i-\omega^T\cdot x-b)^2. \tag{1}$$

Ridge regression is a method of linear regression that implements a sum-of-squares error function together with regularization, thus controlling the bias variance trade-off [4,5]. It aims at discovering a linear structure hidden in the original data [6–9]. Meanwhile, nonlinear mappings may be estimated by kernel ridge regression [10–12], an extended version of linear ridge regression with kernel tricks. A nonlinear ridge regression model is constructed in a feature space H (the nonlinear kernel mapping $\Phi: R^L\longrightarrow H$, where H is the Hilbert space), induced by the nonlinear kernel function $K(x_i,x_j)=(\Phi(x_i)\cdot\Phi(x_j))$, $(\Phi(x_i)\cdot\Phi(x_j))$ is the inner product. The kernel mapping Φ may be any positive definite Mercer kernel. So the objective function minimized in kernel ridge regression can be written as

$$g_{KRR}=\frac{1}{2}\cdot\omega^T\cdot\omega+C\cdot\sum_{i=1}^{l}(y_i-\omega^T\cdot\Phi(x)-b)^2. \tag{2}$$

In recent years, kernel ridge regression (*KRR*) is gaining popularity as a data-rich nonlinear forecasting tool [3,6,8–11], which is applicable in many different contexts [12–14], such as economic field, machine learning, and especially optical character recognition.

In 2000, Suykens et al. [15–18] proposed kernel ridge regression model with Gaussian noise (*GN-KRR*, also known as least squares support vector regression, *LS-SVR*). We know that this technique is able to find the optimal model if the errors are Gaussian. However, it was reported that the noise in some real-world applications, just like wind power forecast and direction-of-arrival estimation problem, does not satisfy Gaussian distribution, but Beta distribution, Laplace distribution, or other models. In this case, the classical regression techniques are not optimal. Despite the fact that most works on wind power forecast assume a normal distribution function to represent the probability density function

* Corresponding author at: College of Mathematics and Information Science, Hebei Normal University, Shijiazhuang, Hebei 050024, China.
E-mail addresses: sg201106@163.com (S. Zhang), huqinghua@tju.edu.cn (Q. Hu), xiezongxia@hitsz.edu.cn (Z. Xie), mijsh@263.net (J. Mi).

http://dx.doi.org/10.1016/j.neucom.2014.07.051
0925-2312/

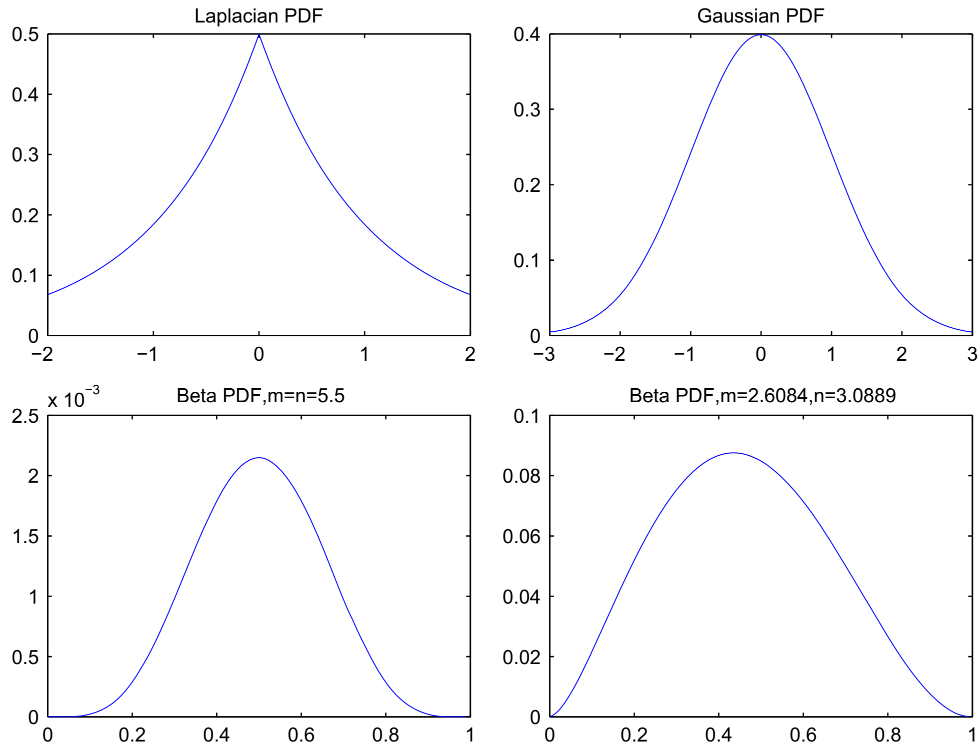

Fig. 1. Laplacian PDF, Gaussian PDF and Beta PDF of parameters.

(PDF) of the prediction error, in [19], a more comprehensive proposal modeling that PDF as a Beta function is justified. Fabbri et al. believed that Beta distribution function is more appropriate to fit the error than the standard normal distribution function of wind power forecast. Bludszuweit et al. [20] showed the advantages of using Beta PDF for approximating the error distribution on wind power forecasting instead of the Gaussian PDF. The error ϵ obeys the Beta distribution in the forecast of wind power, and the PDF of ϵ is $f(\epsilon)=\epsilon^{m-1}\cdot(1-\epsilon)^{n-1}\cdot h, \epsilon\in(0,1)$, where m and n are two parameters $(m>1, n>1)$, h is the normalization factor [21–23] (Laplacian distribution, Gaussian distribution and Beta distribution for different parameters are shown in Fig. 1). Zhang et al. [24] and Randazzo et al. [25] presented the estimation model under a Laplacian noise environment in the direction-of-arrival of coherent electromagnetic waves impinging estimation problem. Laplacian distribution is frequently encountered in various machine learning areas [26,27].

Based on the above analysis, we know that the error distributions do not satisfy Gaussian distribution in some application fields. We know that different loss functions should be derived for different noise models. Squared loss is fit for Gaussian distribution, and absolute loss is good for Laplacian distribution. We systematically discuss the optimal loss functions for different noise models.

It is not suitable to apply the kernel ridge regression with Gaussian noise (*GN-KRR*) to fit functions from data with non-Gaussian noise. In order to solve the above problems, we derive a general loss function and develop a new framework of kernel ridge regression models for the general noise (*N-KRR*). Finally, we design an algorithm to find the optimal solution to the corresponding regression tasks. While there are a large number of implementations of *KRR* algorithms in the past few years, we introduce the Augmented Lagrangian Multiplier method, presented in Section 3. If the task is non-differentiable or discontinuous, the sub-gradient descent method can be used [28].

The main contributions of our work are listed as follows: (1) we derive the optimal loss functions for different noise models; (2) we develop a new framework of kernel ridge regression technique for the general noise model; (3) the Augmented Lagrangian Multiplier method is applied to solve the proposed model, which guarantees the stability and validity of the solution; (4) we utilize the kernel ridge regression model for Beta noise (*BN-KRR*) to short-term wind speed prediction and show the effectiveness of the proposed model in practical applications.

This paper is structured as follows. In Section 2, we obtain the optimal loss function and construct a kernel ridge regression machine for the general noise model (*N-KRR*). Section 3 gives the solution and algorithm design of *N-KRR*. Numerical experiments are conducted on artificial data and short-term wind speed prediction in Section 4. Finally, the conclusions are drawn in Section 5.

2. Kernel ridge regression for general noise model

Kernel ridge regression can be understood as a function approximation technique. A key issue in developing a regression technique is to derive an optimal objective for the general noise model. Given a set of noisy training samples D_l, we require to estimate an unknown regression function $f(x)$. We assume that the noise is additive:

$$y_i=f(x_i)+e_i \quad (i=1,2\ldots,l), \tag{3}$$

where e_i is the error, and we assume that the observations are drawn in independent and identical distributed (i.i.d.) with $P(e_i)$ of standard deviation σ and mean μ. The objective is to estimate the function $f(x)$ with the data set $D_f\subseteq D_l$.

Following [29–31], the general approach is to minimize

$$H[f]=\sum_{i=1}^{l}c(e_i)+\lambda\cdot\Phi[f] \tag{4}$$

where $c(e_i)=c(y_i-f(x_i))$ is a loss function, λ is a positive number and $\Phi[f]$ is a smoothness functional.