Computational Statistics and Data Analysis 68 (2013) 1–16

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Model selection in kernel ridge regression

Peter Exterkate [*]

CREATES, Department of Economics and Business, Aarhus University, Fuglesangs Allé 4, 8210 Aarhus V, Denmark

ARTICLE INFO

Article history:
Received 28 February 2012
Received in revised form 4 June 2013
Accepted 5 June 2013
Available online 14 June 2013

Keywords:
Nonlinear forecasting
Shrinkage estimation
Kernel methods
High dimensionality

ABSTRACT

Kernel ridge regression is a technique to perform ridge regression with a potentially infinite number of nonlinear transformations of the independent variables as regressors. This method is gaining popularity as a data-rich nonlinear forecasting tool, which is applicable in many different contexts. The influence of the choice of kernel and the setting of tuning parameters on forecast accuracy is investigated. Several popular kernels are reviewed, including polynomial kernels, the Gaussian kernel, and the Sinc kernel. The latter two kernels are interpreted in terms of their smoothing properties, and the tuning parameters associated to all these kernels are related to smoothness measures of the prediction function and to the signal-to-noise ratio. Based on these interpretations, guidelines are provided for selecting the tuning parameters from small grids using cross-validation. A Monte Carlo study confirms the practical usefulness of these rules of thumb. Finally, the flexible and smooth functional forms provided by the Gaussian and Sinc kernels make them widely applicable. Therefore, their use is recommended instead of the popular polynomial kernels in general settings, where no information on the data-generating process is available.

1. Introduction

In many areas of application, forecasters face a trade-off between model complexity and forecast accuracy. Due to the uncertainty associated with model choice and parameter estimation, a complex nonlinear predictive model is often found to produce less accurate forecasts than a simpler, e.g. linear, model. Thus, a forecaster wishing to estimate a nonlinear relation generally restricts the search space drastically, for example to polynomials of low degree, or to regime-switching models (Teräsvirta, 2006) or neural networks (White, 2006; Castillo et al., 2008). A recent comprehensive overview was given by Kock and Teräsvirta (2011). The improvement of such models upon the predictive accuracy of linear models is often found to be limited; see for example Stock and Watson (1999), Teräsvirta et al. (2005), and Medeiros et al. (2006). Moreover, these techniques tend not to be very suitable for large numbers of predictors.

Another manifestation of this complexity–accuracy trade-off is that, while a very large number of potentially relevant predictors may be available, the *curse of dimensionality* implies that better forecasts can be obtained if a large proportion of the predictors is discarded. This situation arises, for example, in economic applications. Hundreds of predictors are often available, and economic theory does not usually provide guidelines concerning which variables should or should not be included in a model. A reduction in the number of predictors can be achieved by selecting a small subset of representative variables, using techniques such as the Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), or sparse canonical correlation analysis (An et al., 2013). However, it may be undesirable to leave many potential predictors out of the model completely. Another popular way to proceed is to summarize the predictors by a small number of principal components. This approach has found successful forecasting applications in macroeconomics (e.g. Stock and Watson, 2002) and in finance (e.g. Ludvigson and Ng, 2007, 2009).

[*] Tel.: +45 8716 5548.
E-mail address: exterkate@creates.au.dk.

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.06.006

These techniques estimate either nonlinear models with few predictors, or linear models with many predictors. This paper discusses *kernel ridge regression*, a forecasting technique that overcomes both problems simultaneously, making it suitable for estimating nonlinear models with many predictors. While kernel methods are not widely known in the fields of economics and finance, they have found ample applications in machine learning; a review can be found in Hofmann et al. (2008). Typical applications are in classification rather than regression, such as the optical recognition of handwritten characters (Schölkopf et al., 1998). Recently, Exterkate et al. (2013) use this technique in a macroeconomic forecasting application and they report an increase in forecast accuracy, compared to traditional methods.

The central idea in kernel ridge regression is to employ a flexible set of nonlinear prediction functions and to prevent overfitting by penalization. This is done by mapping the set of predictors into a high-dimensional (or even infinite-dimensional) space of nonlinear functions of the predictors. A linear forecast equation is then estimated in this high-dimensional space, using a penalty (or shrinkage, or ridge) term to avoid overfitting. Computational tractability is achieved by choosing the mapping in a convenient way, so that calculations in the high-dimensional space are actually prevented.

Kernel ridge regression provides a large amount of flexibility in model building, but it also leaves the researcher with a number of nontrivial decisions to make. One such decision concerns which kernel to use. Although any choice of kernel leads to restrictions on the functional form of the forecast equation, little attention is generally being paid to such implications. Additionally, kernel ridge regression involves tuning parameters, and their practical interpretation is not always clear. This feature makes it difficult to select "reasonable" values for these parameters, resulting in time-consuming grid searches or in suboptimal forecasting performance.

To give a clear interpretation of the kernel functions and their associated tuning parameters, we review the kernel methodology from two different points of view, namely, function approximation and Bayesian statistics. This combination of perspectives enables us to relate one of the two tuning parameters that are found in most applications of kernel ridge regression to the signal-to-noise ratio in the data, and the other to the smoothness of the prediction function. We give explicit rules of thumb for selecting their values by using cross-validation over small grids. Cross-validation may also be used to select among different types of kernel. However, we provide empirical evidence against including the popular polynomial kernels in the cross-validation exercise.

In Section 2 we describe the kernel methodology, from the perspective of function approximation and from Bayesian statistics. We discuss several popular kernels and the functional forms of their associated forecast equations, and we interpret their tuning parameters. Section 3 presents a Monte Carlo simulation to show the effects of different methods for choosing the kernel and its tuning parameters. Selecting the tuning parameters using cross-validation affects the forecast quality only marginally, compared to using the true values. The kernel can also be chosen by cross-validation; however, using a polynomial kernel when the data-generating process is non-polynomial has a substantial impact on the forecast accuracy. We also present simulations in which all kernels estimate misspecified models, and the "smooth" Gaussian and Sinc kernels perform best in this case. We provide conclusions in Section 4.

2. Methodology

Kernel ridge regression can be understood as a function approximation tool, but it can also be given a Bayesian interpretation. We review the method from both viewpoints in Sections 2.1 and 2.2, respectively. We present some popular kernel functions in Section 2.3. In Section 2.4 we give an interpretation to the associated tuning parameters, and we derive "reasonable" values for these parameters.

2.1. Kernel ridge regression for function approximation

We first introduce some notation. We have T observations $(y_1, x_1), (y_2, x_2), \ldots, (y_T, x_T)$, with $y_t \in \mathbb{R}$ and $x_t \in \mathbb{R}^N$, and our goal is to find a function f so that $f(x_t)$ is a "good" approximation to y_t for all $t = 1, 2, \ldots, T$. Then, a new observation $x_* \in \mathbb{R}^N$ is observed and we wish to predict the corresponding y_*. We denote this prediction by $\hat{y}_* = f(x_*)$. By selecting f from a large and flexible class of functions while preventing overfitting, we hope to achieve that this prediction is accurate.

To describe the class of functions from which we select f, we first choose a mapping $\varphi : \mathbb{R}^N \to \mathbb{R}^M$. The regression function f will be restricted to a certain set of linear combinations of the form $\varphi(x)' \gamma$, with coefficient vector $\gamma \in \mathbb{R}^M$. The number of regressors M is either a finite integer with $M \geq N$, or $M = \mathbb{N}$, representing a countably infinite number of regressors. Examples of mappings of both types are presented in Section 2.3.

If a flexible functional form is desired, the number of regressors M needs to be large. Therefore we wish to avoid M-dimensional computations, and it turns out that we can do so by requiring only that the inner product $\kappa(x_s, x_t) = \varphi(x_s)' \varphi(x_t)$ can be found using only N-dimensional computations, for any $x_s, x_t \in \mathbb{R}^N$. In the machine learning literature this idea is known as the *kernel trick* (Boser et al., 1992), and the function $\kappa : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ is commonly called the kernel function. Conversely, functions κ for which a corresponding φ exists can be characterized by a set of conditions due to Mercer (1909). All kernel functions discussed in this study satisfy these conditions; a thorough justification can be found in Hofmann et al. (2008).

Finally, define a space of functions $\mathcal{H}_0$ consisting of the functions $f : \mathbb{R}^N \to \mathbb{R}$ of the form $f(x) = \sum_{s=1}^{S} \alpha_s^f \kappa\left(x, x_s^f\right)$, for a finite set $x_1^f, x_2^f, \ldots, x_S^f \in \mathbb{R}^N$ and real numbers $\alpha_1^f, \alpha_2^f, \ldots, \alpha_S^f$. Every such $f(x)$ is a linear combination of the elements