Fixed and Variable-Width Gaussian Networks

Kůrková, Věra
2012

Institute of Computer Science
Academy of Sciences of the Czech Republic

Fixed and Variable-Width Gaussian Networks

Věra Kůrková and Paul C. Kainen

Technical report No. 1174

2012

Pod Vodárenskou věží 2, 182 07 Prague 8, phone: +420 266 051 111, fax: +420 286 585 789,
e-mail:ics@cs.cas.cz

Institute of Computer Science
Academy of Sciences of the Czech Republic

Fixed and Variable-Width Gaussian Networks

Věra Kůrková and Paul C. Kainen

Technical report No. 1174

2012

Abstract:

The role of widths of Gaussians in computational models which they generate is investigated. It is shown that networks with Gaussian kernel units are functionally equivalent merely when they are generated by kernels with the same widths and have the same number of units which differ merely by a permutation. Suitability of Gaussian kernel models with fixed widths for regression is proven in terms of their universal approximation capability.

Keywords:
Gaussian radial networks, Gaussian kernel networks, functionally equivalent networks, approximation of functions by neural networks

1 Introduction

Originally, artificial neural networks were built from biologically inspired computational units. These units, called perceptrons, compute functions in the form of plane waves and thus they are highly nonlocal. As an alternative, localized computational units were proposed merely due to their good mathematical properties. Broomhead and Lowe [1] introduced radial-basis-functions (RBF) and Girosi and Poggio [6] proposed more general kernel units. In particular, support vector machine (SVM) built from units defined by symmetric positive semidefinite kernels became very popular [2]. Among localized computational units, a prominent position is occupied by units induced by the Gaussian function. RBF units with the Gaussian radial function are the most common type of RBF's and Gaussians with fixed widths are typical symmetric positive definite kernels. Both these computational models, the one with Gaussian units having variable widths (RBF) and the one with Gaussian units having fixed widths (symmetric positive definite kernels), have their advantages. Gaussian RBF networks are known to be universal approximators [11]. In addition to their capability to approximate arbitrarily well all reasonable real-valued functions, model complexity of Gaussian RBF networks is often lower than complexity of traditional linear approximators, in particular in high-dimensional tasks (see, e.g., [9, 8, 7] for some estimates). On the other hand, Gaussian kernel models with fixed widths benefit from geometrical properties of Hilbert spaces which they generate. These properties allow an extension of maximal margin classification algorithm to data which are not linearly separable [2], generate suitable stabilizers for modeling of generalization in terms of regularization [5], and lead to mathematical description of theoretically optimal solutions of learning tasks [3, 12, 10].

In this paper, we investigate the role of widths of Gaussians in computational models which they generate. First, we show that two networks with Gaussian kernel units are functionally equivalent merely when they are generated by kernels with the same widths and have the same number of units which differ merely by a permutation. Thus possibilities of compressions of parameter spaces are limited to equivalences induced by permutations. Then we show that besides of well-known classification capabilities of Gaussian kernel models, they are also suitable for regression as even with a fixed width, they are large enough to approximate all continuous or square integrable multivariable functions.

The paper is organized as follows. In section 2, notations and basic concepts on one-hidden-layer networks are introduced. In section 3, it is shown that for two different widths, Gaussian kernel networks are not functionally equivalent. Section 4 shows that Gaussian kernel networks with fixed width are universal approximators.

2 Dictionaries of Computational Units

The most widespread computational model used in neurocomputing is a *one-hidden-layer network with one linear output unit*. Such networks compute linear combinations of functions computable by a given type of computational units. The coefficients oflinear combinations are called output weights. Networks with n units from a dictionary G compute functions from the set

$$\operatorname{span}_n G := \left\{ \sum_{i=1}^{n} w_i g_i \,|\, w_i \in \mathbb{R}, g_i \in G \right\}.$$

One says that n is the *number of hidden units*. Hence, the set of input-output functions of all such networks, with an arbitrary number of hidden units, is

$$\operatorname{span} G := \bigcup \{\operatorname{span}_n G,\ n \in \mathbb{N}_+ \} = \left\{ \sum_{i=1}^{n} w_i g_i \,|\, w_i \in \mathbb{R},\ g_i \in G, n \in \mathbb{N}_+ \right\}.$$

Typically, dictionaries are parameterized families of functions. Let

$$G_K(X,Y) := \{K(.,y) : X \to \mathbb{R} \,|\, y \in Y\},$$

where $K : X \times Y \to \mathbb{R}$ is a function of two variables, an input vector $x \in X \subseteq \mathbb{R}^d$ and a parameter $y \in Y \subseteq \mathbb{R}^s$. In mathematics, various functions of two variables are called kernels (from the German

term "kern", introduced by Hilbert in the context of theory of integral operators). In neurocomputing and learning theory, the term kernel is often reserved for symmetric positive semidefinite functions.

In this paper, we focus on dictionaries defined in terms of the Gaussian function. The first one G_{F_d} is induced by the function

$$F_d(x,(a,c)) := e^{-\|a(x-c)\|^2} : \mathbb{R}^d \times \mathbb{R}^{d+1} \to \mathbb{R}.$$

$$F_d : \mathbb{R}^d \times \mathbb{R}^{d+1} \to \mathbb{R} \;\; (x,(a,c)) \mapsto e^{-\|a(x-c)\|^2}.$$

So

$$G_{F_d}(X) := \{F_d(.,(a,c)) : X \to \mathbb{R} \,|\, a > 0,\, c \in \mathbb{R}^d\}$$

consists of functions on X computable by Gaussian RBF units with varying centers c and varying widths $\frac{1}{a}$.

The second dictionary $G_{K_d^a}$ is induced by the Gaussian of a fixed width $\frac{1}{a}$,

$$K_d^a(x,c) := e^{-\|a(x-c)\|^2} : \mathbb{R}^d \times \mathbb{R}^{d+1} \to \mathbb{R}.$$

So

$$G_{K_d^a}(X) := \{K_d^a(.,c) : X \to \mathbb{R} \,|\, c \in \mathbb{R}^d\}$$

consists of functions on X computable by Gaussian RBF units with varying centers c and fixed width $\frac{1}{a}$. Thus we have $G_{F_d}(X) := \bigcup_{a \in \mathbb{R}_+} G_{K_d^a}(X)$.

3 Functionally Equivalent Gaussian RBFs

In this section, we investigate functional equivalence of Gaussian kernel networks with different numbers of hidden units and different widths. We show that two Gaussian RBF networks compute the same input-output function merely when when they have the same numbers of hidden units with the same centers, widths, and output weights which differ merely by a permutation.

Two neural networks are called *functionally equivalent* if they compute the same input-output function. Functional equivalences of neural networks can be studied in terms of linear dependencies of dictionaries. Recall that a set of functions F is *linearly independent* if for any finite subset of its elements $\{f_1,\ldots,f_m\}$ and real numbers $w_1,\ldots,w_m$, $\sum_{i=1}^m w_i f_i = 0$ implies $w_i = 0$ for all $i = 1,\ldots,m$. A point $x \in X$ is a limit point of X if every neighborhood of x contains at least one point of X different from x itself. If a dictionary is linearly independent, then two networks are functionally equivalent only when they have the same number of units with the same parameters which can only differ by permutation. We show that the dictionary of Gaussian RBFs on any open subset of $\mathbb{R}^d$ is functionally equivalent.

Theorem 3.1 *For every positive integer d, the dictionary*

$$G_{F_d}(\mathbb{R}^d) = \{\exp(-a^2\|. - c\|^2) : \mathbb{R}^d \to \mathbb{R} \,|\, a \in \mathbb{R}_+, c \in \mathbb{R}^d\}$$

is linearly independent.

Proof. We show that no nontrivial linear combination of elements of $G_{F_d}(\mathbb{R}^d)$ is equal to zero. Let m be a positive integer, $w_1,\ldots,w_m$ be non zero real numbers, and $\{(a_j,c_j) \in \mathbb{R}_+ \times \mathbb{R}^d \,|\, j = 1,\ldots,m\}$ be a set of distinct pairs. To prove the statement by contradiction, assume that for all $x \in \mathbb{R}^d$

$$\sum_{j=1}^m w_j e^{-a_j^2\|x-c_j\|^2} = 0. \tag{3.1}$$

Without loss of generality we can suppose that $1 = \max\{a_j \,|\, j = 1,\ldots,m\}$ (otherwise we change scale) and that $1 = a_1 = \ldots = a_k > a_{k+1} \geq ... \geq a_m > 0$. We may also assume that $c_1 = 0$ (otherwise we change system of coordinates). In addition we can assume that $\|c_2\| \geq \|c_i\|$ for all $j = 3,\ldots,k$. As the pairs $(1,c_1),\ldots,(1,c_k)$ are distinct, so are $c_2,\ldots,c_k$.

Thus we have for all $j = 3, \ldots, k$,

$$c_2 \cdot (c_2 - c_j) > 0. \tag{3.2}$$

Indeed, $c_2 \cdot c_j = \|c_2\| \, \|c_j\| \cos(\alpha(c_2, c_j))$, where $\alpha(c_2, c_j)$ denotes the angle between the vectors c_2 and c_j. For those j for which $\|c_j\| < \|c_2\|$, we have $\|c_j\| \cos(\alpha(c_2, c_j)) < \|c_2\|$ and so $c_2 \cdot (c_2 - c_j) > 0$. For those j for which $\|c_j\| = \|c_2\|$, we have $\cos(\alpha(c_2, c_j)) < 1$ because $c_j \neq c_2$ and so $c_2 \cdot (c_2 - c_j) > 0$.

Multiplying both sides of the equation (3.1) by $e^{\|x\|^2}$ we get for all $x \in \mathbb{R}^d$,

$$w_1 + \sum_{j=2}^{k} \bar{w_j} e^{2c_j \cdot x} + \sum_{j=k+1}^{m} \bar{w_j} e^{\|x\|^2 (1 - a_j^2) + 2a_j^2 c_j \cdot x} = 0,$$

where $\bar{w_j} = w_j e^{-c_j^2}$ for $j = 2, \ldots, m$. For all $j = k+1, \ldots, m$, $1 - a_j^2 < 0$ and so $w_1 + \lim_{\|x\| \to \infty} \sum_{j=2}^{k} \bar{w_j} e^{2c_j \cdot x} = 0$.

If $k = 1$, we have a contradiction with the assumption that $w_1 \neq 0$. If $k > 1$, we set $x = tc_2$ and so we obtain $w_1 + \lim_{t \to \infty} \sum_{j=2}^{k} \bar{w_j} e^{2tc_j \cdot c_2} = 0$. Thus we get
$\lim_{t \to \infty} w_1 e^{-2t\|c_2\|^2} + \bar{w_2} + \lim_{t \to \infty} \sum_{j=3}^{k} \bar{w_j} e^{-2tc_2 \cdot (c_2 - c_j)} = 0$. As $c_2 \cdot (c_2 - c_j) > 0$ for all $j = 3, \ldots, k$, both limits in this equation are equal to zero and thus we get a contradiction with $w_2 \neq 0$. □

Theorem 3.1 implies that functionally equivalent Gaussian kernel networks must have the same width, the same number of hidden units and can differ merely by a permutation of hidden units. It also shows that the only reduction of parameter spaces of Gaussian RBF networks based on their functional equivalence is induced by permutations of hidden units. Search in such reduced parameter spaces might be implementable for genetic algorithms which operate with strings of vectors of parameters.

4 Universal Approximation Property

In this section, we show that although Gaussian kernel units with fixed widths have much less free parameters than Gaussian radial units with varying widths, they still generate classes of input-output functions large enough to be universal approximators. Recall that a class of one-hidden-layer networks with units from a dictionary G is said to have the *universal approximation property in a normed linear space* $(\mathcal{X}, \|.\|_{\mathcal{X}})$ if it is *dense* in this space, i.e., $\mathrm{cl}_{\mathcal{X}} \operatorname{span} G = \mathcal{X}$, where $\mathrm{cl}_{\mathcal{X}}$ denotes the closure with respect to the topology induced by the norm $\|.\|_{\mathcal{X}}$. Function spaces where the universal approximation has been of interest are spaces $(\mathcal{C}(X), \|.\|_{\sup})$ of continuous functions on subsets X of $\mathbb{R}^d$ (typically compact) with the supremum norm and the space $(\mathcal{L}^2(\mathbb{R}^d), \|.\|_{\mathcal{L}^2})$ of square integrable functions on $\mathbb{R}^d$ with the norm $\|f\|_{\mathcal{L}^2} = \left(\int_{\mathbb{R}^d} f(y)^2 dy\right)^{1/2}$.

Theorem 4.1 *Let d be a positive integer and $a > 0$, then*
(i) for $X \subseteq \mathbb{R}^d$ Lebesgue measurable, $\operatorname{span} G_{K_d^a}(X)$ *is dense in* $(\mathcal{L}^2(X), \|.\|_{\mathcal{L}^2})$*;*
(ii) for $X \subset \mathbb{R}^d$ compact, $\operatorname{span} G_{K_d^a}(X)$ *is dense in* $(\mathcal{C}(X), \|.\|_{\sup})$*.*

Proof. First, assume that $X = \mathbb{R}^d$. Suppose that $\mathrm{cl}_{\mathcal{L}^2} \operatorname{span} G_{K_d^a}(\mathbb{R}^d) \neq \mathcal{L}^2(\mathbb{R}^d)$. Then by Hahn-Banach Theorem [13, p. 60] there exists a linear functional l on $\mathcal{L}^2(\mathbb{R}^d)$ such that for all $f \in \mathrm{cl}_{\mathcal{L}^2} \operatorname{span} G_{K_d^a}(\mathbb{R}^d)$, $l(f) = 0$ and for some $f_0 \in \mathcal{L}^2(\mathbb{R}^d) \setminus \mathrm{cl}_{\mathcal{L}^2} \operatorname{span} G_K(\mathbb{R}^d)$, $l(f_0) = 1$. By Riesz Representation Theorem [4], there exists $h \in \mathcal{L}^2(\mathbb{R}^d)$, such that for all $g \in \mathcal{L}^2(\mathbb{R}^d)$, $l(g) = \int_{\mathbb{R}^d} g(y) h(y) dy$. Thus for all $f \in \mathrm{cl}_{\mathcal{L}^2} \operatorname{span} G_{K_d^a}(\mathbb{R}^d)$, $\int_{\mathbb{R}^d} f(y) h(y) dy = 0$. Denoting $k^a(x) = e^{-a^2 \|x\|^2}$, we get for all $x \in \mathbb{R}^d$, $\int_{\mathbb{R}^d} h(y) k^a(x - y) dy = h * k^a(x) = 0$. Thus by Plancherel Theorem [13, p.188], $\|\widehat{h * k^a}\|_{\mathcal{L}^2} = 0$. As $\widehat{h * k^a} = \frac{1}{(2\pi)^{d/2}} \hat{h} \, \hat{k^a}$ [13, p.183], we have $\|\hat{h} \, \hat{k^a}\|_{\mathcal{L}^2} = 0$. As $\widehat{e^{-a^2 \|.\|^2}} = (\sqrt{2}a)^{-d} e^{-(1/a^2)\|.\|^2}$ [13, p.186], we obtain $\|\hat{h}\|_{\mathcal{L}^2} = 0$. So by Plancherel Theorem, $\|h\|_{\mathcal{L}^2} = 0$. Hence we get $1 = l(f_0) = \int_{\mathbb{R}^d} f_0(y) \, h(y) dy \leq \|f_0\|_{\mathcal{L}^2} \, \|h\|_{\mathcal{L}^2} = 0$, which is a contradiction.

Now, let $X \subset \mathbb{R}^d$, be an arbitrary Lebesgue measurable set. We obtain (i) by extending functions from $\mathcal{L}^2(X)$ to $\mathcal{L}^2(\mathbb{R}^d)$ by setting their values equal to zero outside of X and restricting their approximations from $\operatorname{span} G_K(\mathbb{R}^d)$ to X. For X compact, $\mathcal{C}(X) \subset \mathcal{L}^2(X)$ and so the statement follows directly from (i). □

Acknowledgments.

This work was partially supported by MŠMT grant INTELLI OC10047 and the Institutional Research Plan RVO 67985807.

Bibliography

[1] D. S. Broomhead and D. Lowe. Error bounds for approximation with neural networks. *Complex Systems*, 2:321–355, 1988.

[2] C. Cortes and V. N. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

[3] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of AMS*, 39:1–49, 2002.

[4] A. Friedman. *Modern Analysis*. Dover, New York, 1982.

[5] F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10:1455–1480, 1998 (AI memo 1606).

[6] F. Girosi and T. Poggio. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247(4945):978–982, 1990.

[7] G. Gnecco, V. Kůrková, and M. Sanguineti. Can dictionary-based computational models outperform the best linear ones? *Neural Networks*, 24(8):881–887, 2011.

[8] G. Gnecco, V. Kůrková, and M. Sanguineti. Some comparisons of complexity in dictionary-based and linear computational models. *Neural Networks*, 24(1):171–182, 2011.

[9] P. C. Kainen, V. Kůrková, and M. Sanguineti. Complexity of Gaussian radial basis networks approximating smooth functions. *J. of Complexity*, 25:63–74, 2009.

[10] V. Kůrková. Inverse problems in learning from data. In E. Kaslik and S. Sivasundaram, editors, *Recent advances in dynamics and control of neural networks*, page to appear. Cambridge Scientific Publishers, 2012.

[11] J. Park and I. Sandberg. Universal approximation using radial–basis–function networks. *Neural Computation*, 3:246–257, 1991.

[12] T. Poggio and S. Smale. The mathematics of learning: dealing with data. *Notices of AMS*, 50:537–544, 2003.

[13] W. Rudin. *Functional Analysis*. Mc Graw-Hill, 1991.