ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Averaging of kernel functions

Lluís A. Belanche and Alessandra Tosi

Faculty of Computer Science - Dept. of Software
Technical University of Catalonia, Barcelona, Spain

Abstract. In kernel-based machines, the integration of several kernels to build more flexible learning methods is a promising avenue for research. In particular, in Multiple Kernel Learning a compound kernel is build by learning a kernel that is the weighted mean of several sources. We show in this paper that the only feasible average for kernel learning is precisely the *arithmetic* average. We also show that three familiar means (the geometric, inverse root mean square and harmonic means) for positive real values actually *generate* valid kernels.

1 Introduction

Kernel methods have won great popularity as a tool for the identification of nonlinear systems. Support Vector Machines (SVMs) are basic kernel-based methods that are used for tasks such as classification and regression, among others [1]. Perhaps the biggest limitation of kernel-based methods lies in the choice of a proper kernel for a given problem.

The kernel function is a very flexible container under which to express knowledge about the problem. One way of tayloring kernels is developing *partial* kernels (e.g., one for every descriptive variable) and building a final kernel as the composition or *aggregation* of these partial kernels, an idea that can be traced back to Vapnik [2]. Once these are built, the final obtained kernel is the one used in the SVM. How can this aggregation process be defined? A natural idea is to use a (possibly weighted) *average* of the partial kernels. We show that, for a wide family of averages (including the familiar means), the only average guaranteeing that the aggregated function is always a valid kernel is the *arithmetic* average, and give an easy-to-ckeck necessary condition for even more general averages. In addition, we show that three of the familiar means (the geometric, inverse root mean square and harmonic means) are *generators* of valid kernels.

2 Preliminaries

Probably the simplest characterization for a symmetric function $K : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ being a kernel is via the matrix it generates on finite subsets [1].

Definition 1 *In the real case, the symmetric matrix $A_{n\times n}$ is positive semi-definite (PSD) if, for all vectors $z \in \mathbb{R}^n$, $z'Az \geq 0$.*

Theorem 1 *The function $K : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is a kernel in $\mathcal{H}$ if and only if for any positive $p \in \mathbb{N}$ and choice of finite subsets $\{x_1, x_2, ..., x_p\} \subset \mathcal{H}$, the associated matrix $K_{p\times p} = (k_{ij})$, where $k_{ij} = K(x_i, x_j)$ is a symmetric PSD matrix.*

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Definition 2 *Let $[a, b]$ be a non-empty real interval. Call $A(x_1, \ldots, x_n)$ the A-average of $x_1, ..., x_n \in [a, b]$ to every n-place real function A fulfilling:*

Axiom A1. *A is continuous, symmetric and strictly increasing in each x_i.*

Axiom A2. $A(x, \ldots, x) = x$.

Axiom A3. *For any $k \leq n$:* $A(x_1, \ldots, x_n) = A(\underbrace{y_k, \ldots, y_k}_{k \text{ times}}, x_{i_{k+1}}, \ldots, x_{i_n})$

where $y_k = A(x_{i_1}, , \ldots, x_{i_k})$ and $(i_1, ..., i_n)$ is a permutation of $(1, \ldots, n)$.

The means defined by these axioms fulfill Cauchy's property of means, namely, that $\min x_i \leq A(x_1, \ldots, x_n) \leq \max x_i$ (the proof is straightforward using axioms A1 and A2). A further interesting property of these means is that we can add averaged elements to an A-average without changing the overall result. Formally, $A(x_1, ..., x_n) = A(z_1, ..., z_m, x_1, ..., x_n)$ if and only if $A(z_1, ..., z_m) = A(x_1, ..., x_n)$. As a consequence, if $y = A(x_1, ..., x_n)$, then $A(x_1, ..., x_n, A(x_1, ..., x_n)) = y$.

Theorem 2 *([3]) Let $f : [a, b] \longrightarrow \mathbb{R}$ be a continuous, strictly monotone mapping. Let g be the inverse function of f. Then,*

$$A(x_1, ..., x_n) \equiv g\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i)\right)$$

is a well-defined A-average for all $n \in \mathbb{N}$ and $x_i \in [a, b]$.

An important class of A-averages is formed by choosing $f(z) = z^q, q \in \mathbb{R}$:

$$M_q(x_1, ..., x_n) = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i)^q\right)^{\frac{1}{q}}$$

Several well-known means (usually called *generalized means*) are derived by choosing particular values of q. Specifically, the *arithmetic* mean for $q = 1$, the *geometric* mean for $q = 0$, the *root mean square* or RMS mean for $q = 2$ and the *harmonic* mean for $q = -1$. A property of the means M_q is that, for $q \neq q'$, $M_q(x_1, ..., x_n) \geq M_{q'}(x_1, ..., x_n)$ if and only if $q > q'$, with equality only if $x_1 = x_2 = \ldots = x_n$. A substantial part of the classic book by Hardy *et al* [4] is devoted to the study of these functions and their applications.

3 A-averages as kernel aggregators

It is a well-known fact that the *sum* of $m > 0$ kernels is again a kernel. Therefore the arithmetic average (function M_1 in the notation of this paper) is a valid kernel aggregator. The *product* of $m > 0$ kernels is also a kernel. However, the product is not an average; it requires the application of the m-root to the result to form the geometric mean (function M_0). Is there any other generalized mean guaranteeing this kernel property? As we shall see in the remaining of this

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

section, the answer is 'no'. Given that this is a rather broad class of averaging functions, the result can be considered as a negative one.

It is convenient to express the *aggregation* in terms of a collection of a number m of PSD matrices: for $k = 1, \ldots, m$, let $A_k = (a_{ij}^k)$ represent $n \times n$ PSD real matrices. Given $f : \mathbb{R}^m \to \mathbb{R}$, define the $n \times n$ real matrix $\bar{A} = (f(a_{ij}^1, \ldots a_{ij}^m))$.

Definition 3 *A function $h : \mathbb{C}^m \to \mathbb{C}$ in the form*

$$h(x_1+iy_1, \ldots, x_m+iy_m) = P(x_1, \ldots, x_m, y_1, \ldots, y_m)+iQ(x_1, \ldots, x_m, y_1, \ldots, y_m)$$

with $x_i, y_i \in \mathbb{R}$ is said to be entire *if it is holomorphic on all $\mathbb{C}^m$, i.e., it is continuous and differentiable in each complex variable on all $\mathbb{C}^m$, and satisfies the Cauchy-Riemann equations in each complex variable:*

$$\frac{\partial P}{\partial x_i}(x_1^0, \ldots, x_m^0, y_1^0, \ldots, y_m^0) = +\frac{\partial Q}{\partial y_i}(x_1^0, \ldots, x_m^0, y_1^0, \ldots, y_m^0)$$
$$\frac{\partial P}{\partial y_i}(x_1^0, \ldots, x_m^0, y_1^0, \ldots, y_m^0) = -\frac{\partial Q}{\partial x_i}(x_1^0, \ldots, x_m^0, y_1^0, \ldots, y_m^0)$$

for each $(x_1^0, \ldots, x_m^0, y_1^0, \ldots, y_m^0) \in \mathbb{C}^m$, $i = 1, \ldots, m$.

Definition 4 *A function $f : \mathbb{R}^m \to \mathbb{R}$ is said to be* real entire *if it is real on $\mathbb{R}^m$ and it is the restriction on $\mathbb{R}^m$ of an entire function h defined on $\mathbb{C}^m$:*

$$f(x_1, \ldots, x_m) \equiv Re\,(h(x_1, \ldots, x_m, 0, \ldots, 0))$$

Proposition 1 *Let $f : \mathbb{R}^m \longrightarrow \mathbb{R}$, the condition of differentiability on all $\mathbb{R}^m$ is necessary for the function f to be real entire.*

Proof. Let f be non-differentiable at least in a point $(\bar{x}_1, \ldots, \bar{x}_m) \in \mathbb{R}^m$ and suppose that f *is* real entire, i.e. there exists an entire function $h : \mathbb{C}^m \to \mathbb{C}$,

$$h(z_1, \ldots, z_m) = h(x_1 + iy_1, \ldots, x_m + iy_m) =$$
$$= P(x_1, \ldots, x_m, y_1, \ldots, y_m) + iQ(x_1, \ldots, x_m, y_1, \ldots, y_m)$$

such that $f(x_1, \ldots, x_m) \equiv Re\,(h(x_1, \ldots, x_m, 0, \ldots, 0))$. Since h is entire, all the partial derivatives $\frac{\partial}{\partial x_i} P(x_1^0, \ldots, x_m^0, y_1^0, \ldots, y_m^0)$ exist and are continuous in each $(x_1^0, \ldots, x_m^0, y_1^0, \ldots, y_m^0) \in \mathbb{C}^m$, and in particular in $(\bar{x}_1, \ldots, \bar{x}_m, 0, \ldots, 0)$. However, $\frac{\partial}{\partial x_i} P(\bar{x}_1, \ldots, \bar{x}_m, 0, \ldots, 0) = \frac{\partial}{\partial x_i}\,(Re\,(h(\bar{x}_1, \ldots, \bar{x}_m, 0, \ldots, 0)))$, which, being equal to $\frac{\partial}{\partial x_i} f(\bar{x}_1, \ldots, \bar{x}_m)$, leads to an absurd. □

Define now $\mathbb{Z}_+^m$ to mean all vectors in $\mathbb{Z}^m$ with nonnegative coordinates. For $\mathbf{x} \in \mathbb{R}^m$ and $\alpha \in \mathbb{Z}_+^m$, we use the standard notation $\mathbf{x}^\alpha = x_1^{\alpha_1} \cdots x_m^{\alpha_m}$. We now use the notion of a *real entire* function to characterize a general *A-average* function as a kernel aggregator thanks to the following theorem [5]:

Theorem 3 *Let $f : \mathbb{R}^m \longrightarrow \mathbb{R}$. Then a matrix $\bar{A}$ generated by f as above is PSD if and only if f is a real entire function of the form $f(\boldsymbol{x}) = \sum_{\alpha \in \mathbb{Z}_+^m} c_\alpha \boldsymbol{x}^\alpha$, $\boldsymbol{x} \in \mathbb{R}^m$, where $c_\alpha \geq 0$ for all $\alpha \in \mathbb{Z}_+^m$.*

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Application 1 *(Generalized means) We apply the previous results to the* A-average *functions of the type M_q, as defined above. With $q = 1$, it is plain to see that M_1 defines a matrix $\bar{A}$ that is a kernel, so we start with M_2 and $m = 2$. Let $A_1 = (a^1_{ij})$, $A_2 = (a^2_{ij})$ and consider the matrix $\bar{A} = (M_2(a^1_{ij}, a^2_{ij})) = (\left[\frac{1}{2}\left((a^1_{11})^2 + (a^2_{11})^2\right)\right]^{1/2})$. Using Theorem 3, f should be real entire; according to Proposition 1 we verify the differentiability of M_2 on $\mathbb{R}^2$:*

$$\frac{\partial M_2(x_1, x_2)}{\partial x_1} = \frac{x_1}{(x_1^2 + x_2^2)}; \qquad \frac{\partial M_2(x_1, x_2)}{\partial x_2} = \frac{x_2}{(x_1^2 + x_2^2)}$$

The partial derivatives of M_2 are not defined in $\mathbf{0} \in \mathbb{R}^2$, so M_2 is not real entire.

For general values of $q \neq 1$ and $m > 0$ the matrix $\bar{A}_{n\times n}$ is in general not PSD because M_q is not a real entire function. Indeed, the partial derivatives

$$\frac{\partial M_q(x_1, \ldots, x_m)}{\partial x_i} = (x_i)^{q-1}\left(\frac{1}{m}\sum_{j=1}^{m}(x_j)^q\right)^{\frac{1}{q}-1}, \quad i = 1, \ldots, m$$

are never defined in $\mathbf{0} \in \mathbb{R}^n$: for $q < 1$ the first *factor is a null denominator, whereas for $q > 1$ the* second *factor is a null denominator. For $q = 1$, the derivatives evaluate to 1 everywhere, showing that only M_1 is a valid choice.*

Application 2 *(Hyperbolic sine mean) Consider the real entire function $sinh(x)$, continuous and monotonic on $\mathbb{R}$, and not expressable as a member of the M_q family of means. Its inverse $arcsinh(x)$ is again a real entire function, because it is the restriction of the entire function $arcsinh(z) = \ln\left(z\sqrt{1+z^2}\right), z \in \mathbb{C}$. A valid average can be defined as:*

$$f_{\sinh}(x_1, x_2) := arcsinh\left(\frac{sinh(x_1) + sinh(x_2)}{2}\right)$$

This is a real entire function since it is the composition of real entire functions. However, its Taylor expansion has negative coefficients (the c_α in Theorem 3):

$$f_{\sinh}(x_1, x_2) = \frac{1}{2}x_1 + \frac{1}{2}x_2 + \frac{1}{16}x_1^3 - \frac{1}{16}x_1^2x_2 - \frac{1}{16}x_1x_2^2 + \frac{1}{16}x_2^3 + O(x_1, x_2)^4$$

Since an holomorphic function has a unique expansion as a power series [6], we conclude that this average is not a valid kernel aggregator.

4 Generalized means as kernel generators

A quite different perspective is obtained if we look at the generalized means as a way to *generate* new kernels from scratch. Under this light, it turns out that the harmonic (M_{-1}), geometric (M_0) and inverse RMS (M_{-2}) means generate valid kernels within their domains[1] (this is not true for the arithmetic mean).

[1]These domains are either $\mathbb{R}^+$ or $\mathbb{R}^-$ for M_0 and M_{-1} and $\mathbb{R}^+$ for M_{-2}.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Theorem 4 *The functions (i)* $k_{\text{geom}} := M_0(x,y) = \sqrt{xy}$*, (ii)* $k_{\text{harm}} := M_{-1}(x,y) = \frac{2xy}{x+y}$ *and (iii)* $k_{\text{IRMS}} := M_{-2}(x,y) = \left(\frac{x_i^{-2}+x_j^{-2}}{2}\right)^{-\frac{1}{2}} = \frac{\sqrt{2}x_ix_j}{\sqrt{x_i^2+x_j^2}}$ *are PSD kernels.*

Proof. Consider any $c \in \mathbb{R}^n$. We have:

$$(i) \sum_{i=1}^n \sum_{j=1}^n c_i c_j k_{\text{geom}}(x_i,x_j) = \sum_{i=1}^n \sum_{j=1}^n c_i c_j \sqrt{x_i x_j} = \left(\sum_{i=1}^n c_i \sqrt{x_i}\right)^2 \geq 0$$

$$(ii) \sum_{i=1}^n \sum_{j=1}^n c_i c_j \frac{2x_ix_j}{x_i+x_j} = 2\sum_{i=1}^n \sum_{j=1}^n c_i c_j x_i x_j \int_0^1 z^{x_i+x_j-1}\,dz =$$

$$2\sum_{i=1}^n \sum_{j=1}^n \int_0^1 (c_i x_i z^{x_i-1/2})(c_j x_j z^{x_j-1/2})\,dz = 2\int_0^1 \left(\sum_{i=1}^n c_i x_i z^{x_i-1/2}\right)^2 dz \geq 0$$

$$(iii) \sum_{i=1}^n \sum_{j=1}^n c_i c_j \frac{\sqrt{2}x_ix_j}{\sqrt{x_i^2+x_j^2}} = \frac{\sqrt{2}}{\sqrt{\pi}} \sum_{i=1}^n \sum_{j=1}^n c_i c_j x_i x_j \int_0^\infty t^{-1/2} e^{-t(x_i^2+x_j^2)}\,dt =$$

$$\sqrt{\frac{2}{\pi}} \int_0^\infty \sum_{i=1}^n \sum_{j=1}^n c_i c_j x_i x_j t^{-1/2} e^{-t(x_i^2+x_j^2)}\,dt = \sqrt{\frac{2}{\pi}} \int_0^\infty \left(\sum_{i=1}^n c_i x_i t^{-1/4} e^{-tx_i^2}\right)^2 dt \geq 0 \qquad \square$$

The interest in these measures lies in the particular *semantics* that they offer. For example, in the geometric mean, the three numbers $a, M_0(a,b), b$ form a geometric sequence, whereas the harmonic mean penalizes big differences between its arguments: consider $M_1(0.5, 11.5) = M_1(6,6) = 6$ but $M_{-1}(0.5, 11.5) \approx 0.96 < M_{-1}(6,6) = 6$; in all cases, the kernels are calculating "compromise" (i.e. average) values between their arguments.

A final point can be made about these three kernels, by forming their *normalized* counterparts. Given $k(x,y)$ a kernel, it is known that the function:

$$\hat{k}(x,y) = \frac{k(x,y)}{\sqrt{k(x,x)}\sqrt{k(y,y)}} \in [-1,1]$$

is again a valid kernel function [1]. Given that all these kernels are also averages, $k(x,x) = x$ always holds true, and therefore the normalization is always equal to $\sqrt{xy}$. We then find that (i) $\hat{k}_{\text{harm}}(x,y) = \frac{2\sqrt{xy}}{x+y}$, (ii) $\hat{k}_{\text{geom}}(x,y) = 1$ (not very useful) and (iii) $\hat{k}_{\text{IRMS}}(x,y) = \left(\sqrt{\frac{x}{y}+\frac{y}{x}}\right)^{-1}$ are valid kernels.

5 Application to Multiple kernel learning

The aggregation of kernels with the weighted arithmetic mean has at least two interpretations. First, we may define (partial) kernels K_i on a finite collection of n spaces $\mathcal{H}^i$; then an overall kernel in $\mathcal{H} = \mathcal{H}^1 \times \ldots \times \mathcal{H}^n$ is built as:

$$\mathcal{K}(x,y) = \sum_{i=1}^n \alpha_i K_i(x_i,y_i),\ x_i, y_i \in \mathcal{H}^i, \alpha_i \geq 0$$

where the α_i sum to 1 (a convex combination). In the simplest instance of this scheme every $\mathcal{H}_i$ represents the domain of a single feature, and the α_i

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

coefficients perform an explicit feature selection process; a more interesting case is found in selecting and combining appropriate *sets* of features.

Second, we can define kernels K_i directly on $\mathcal{H}$ and then a new kernel as:

$$\mathcal{K}(x,y) = \sum_{i=1}^{n} \alpha_i K_i(x,y),\ x,y \in \mathcal{H}, \alpha_i \geq 0$$

In this case, the interest is in departing from a set of kernels (may be a big one) and perform a selection on this set. This set of kernels can be *homogeneous* (different realizations of the same kernel using a different value for a parameter) or *heterogeneous* (different kernels) or both. Multiple Kernel Learning (MKL) seeks to address these situations by *learning* the kernel (i.e., fitting the best α_i from training data), achieving very good results on bioinformatics and computer vision applications [7]. MKL learns linear combinations of base kernels, corresponding to the concatenation of the base kernel feature spaces.

Richer representations can be achieved by combining kernels in other ways. Taking products of kernels corresponds to taking a tensor product of their feature spaces, leading to a much higher dimensional feature representation as compared to concatenation. This big product is a kernel, and one would be then tempted to take the n-root in good fairness, which would not lead to a valid kernel.

6 Conclusions

We have proven that the only feasible average for kernel learning is the arithmetic average. Are there other "reasonable" averages that guarantee this kernel property? If by reasonable we understand any function expressible as an A-average then the answer is not easy to characterize. However, the necessary differentiability condition is quite simple to check and may serve to rule out many averaging candidates. Moreover, for the wide family M_q of generalized means, defining $Q = \{q \in \mathbb{R} \,/\, M_q \text{ is a kernel}\}$, we have proven that $\{-2,-1,0\} \subset Q$ (and certainly $1 \notin Q$). What exactly Q is remains an open question.

Acknowledgements. This study has been partially funded by the Spanish Government project TIN2009-13895-C02-01.

References

[1] Shawe-Taylor, J., Cristianini, N. *Kernel Methods for Pattern Analysis*. Cambridge Univ. Press, 2004.

[2] Vapnik, V.N. *The nature of Statical Learning Theory*. Springer, New York, 1995.

[3] Hille, E. *Methods in Classical and Functional Analysis*. Addison-Wesley, 1971.

[4] Hardy, G., Littlewood, J.E., Pölya, G. *Inequalities*. Cambridge Univ. Press, 1952.

[5] FitzGerald, C., Micchelli, C., Pinkus, A. Functions that preserve families of positive semidefinite matrices. *Linear Algebra & its Applications* 221: 83-102, 1995.

[6] Krantz, S.G. *Function theory of several complex variables*. Wadsworth and Brooks/Cole Advanced Books and Software, Second edition, 1992.

[7] Bach, F. R., Lanckriet, G. R. G., Jordan, M. I. *Multiple kernel learning, conic duality, and the SMO algorithm*. International Conference on Machine Learning (pp. 6-13), 2004.