Averaging of kernel functions

Lluís A. Belanche and Alessandra Tosi
belanche@lsi.upc.edu, atosi@lsi.upc.edu

Soft Computing Research Group
Computer Science School
Technical University of Catalonia Barcelona, Spain

European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning

Bruges (Belgium), 25 - 27 April 2012

Motivation

Kernels generally (and informally) seen as similarity measures

1. Similarities and kernels are two-place symmetric functions ...

2. Are all kernels similarities? No (boundedness, transitivity, ...)

3. Are all similarities kernels? No (PSD)

We deal with averaging kernels as (if they were) similarities

The notion of similarity

1. Human beings use the notion of *similarity* for problem solving: inductive reasoning, analogical thinking...

2. Computer Science: Case Based Reasoning, Data Mining, Information Retrieval, Pattern Matching, Neural Networks, SVMs, ...

The notion of similarity

1. For atomic elements the exist many similarity measures

2. For vectors of elements, a way is needed to *combine* the *partial* similarities s_k for each variable k to get a meaningful value

3. The combination has an important semantic role and it is not a trivial choice.

4. Intuition says "combine by averaging"

Characterization of kernels

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:

Definition 1 *In the real case, the symmetric matrix $A_{n\times n}$ is positive semi-definite (PSD) if and only 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 every 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.*

The concept of an A-average

To capture the notion of *averaging*, we adopt the concept of an *A-average*, defined as:

Definition 2 *Let* $[a,b]$ *be a non-empty real interval. Call* $A(x_1,\dots,x_n)$ *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. $A(x,\dots,x) = x$.

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

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

The concept of an A-average

Some derived properties: $\text{mín}\, x_i \leq A(x_1, \ldots, x_n) \leq \text{máx}\, x_i$

Theorem 2 *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* $x_i \in [a,b]$*.*

The concept of an A-average

An important class of A-averages is formed by choosing $f(z) = z^q$:

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

These are usually called *generalized or quasi-linear means*:

- *arithmetic* mean for $q = 1$
- *geometric* mean for $q = 0$
- *harmonic* mean for $q = -1$
- *root mean square* or RMS mean for $q = 2$

A-averages as kernel aggregators

- The arithmetic average (function M_1) is a valid kernel aggregator.

- The *product* of kernels is also a kernel. However, the product is not an average.

- Is there any other generalized mean guaranteeing the kernel property?

A-averages as kernel aggregators

Notation

It is convenient to express the *aggregation* of m kernels in terms of their PSD matrices:

for $k = 1, \dots, m$, let $A_k = (a_{ij}^k)$ represent a $n \times n$ PSD real matrix.

Given $f : \mathbb{R}^m \to \mathbb{R}$, define the $n \times n$ real matrix $\bar{A} = (f(a_{ij}^1, \dots a_{ij}^m))$.

A-averages as kernel aggregators

FitzGerald, Micchelli and Pinkus (1995)

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:*

1. *f is a real entire function*

2. *f is of the form*

   $$f(\mathbf{x}) = \sum_{\alpha \in \mathbb{Z}_+^m} c_\alpha \mathbf{x}^\alpha, \ \mathbf{x} \in \mathbb{R}^m, \text{ where } c_\alpha \geq 0 \text{ for all } \alpha \in \mathbb{Z}_+^m.$$

Some implications and application examples

Generalized means The matrix $\bar{A}$ 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 $\mathbf{0} \in \mathbb{R}^n$ (except for $q=1$).

Hyperbolic sine mean A real entire A-average can be defined as:

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

However, its Taylor expansion has negative coefficients:

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

Generalized means as kernel generators

- A different perspective is obtained if we look at the generalized means as a way to *generate* new kernels.

- It turns out that the harmonic (M_{-1}), geometric (M_0) and inverse RMS (M_{-2}) means generate valid kernels within their domains.

- Remarkable, since this is *not* true for the arithmetic mean.

Generalized means as kernel generators

Theorem 4 *The following functions are PSD kernels.*

(i) $k_{\text{geom}} := M_0(x,y) = \sqrt{xy}$ *(the geometric kernel)*

(ii) $k_{\text{harm}} := M_{-1}(x,y) = \frac{2xy}{x+y}$ *(the harmonic kernel)*

(iii) $k_{\text{IRMS}} := M_{-2}(x,y) = \left(\frac{x^{-2}+y^{-2}}{2}\right)^{-\frac{1}{2}} = \frac{\sqrt{2}xy}{\sqrt{x^2+y^2}}$ *(the IRMS kernel)*

Conclusions

1. We have proven that the only feasible average for kernel learning is the arithmetic average.

2. Is this a negative result? Yes and no.

3. For the wide family M_q of generalized means, defining $Q = \{q \in \mathbb{R} \,/\, M_q$ 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.