Belanche, L.; Vázquez, J.; Vázquez, M. Distance-based kernels for real-valued data. A: Annual Conference of the Gesellschaft für Klassifikation. "Data analysis, machine learning and applications: proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität: Freiburg, March 7-9, 2007". Berlín: Springer, 2007, p. 3-10.
The final authenticated version is available online at https://doi.org/10.1007/978-3-540-78246-9_1

Distance-based Kernels for Real-valued Data

Lluís Belanche[1], Jean Luis Vázquez[2] and Miguel Vázquez[3]

[1] Dept. de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
08034 Barcelona, Spain
belanche@lsi.upc.edu
[2] Departamento de Matemáticas
Universidad Autónoma de Madrid.
28049 Madrid, Spain
juanluis.vazquez@uam.es
[3] Dept. Sistemas Informáticos y Programación
Universidad Complutense de Madrid
28040 Madrid, Spain
mivazque@fdi.ucm.es

Abstract. We consider distance-based similarity measures for real-valued vectors of interest in kernel-based machine learning algorithms. In particular, a *truncated Euclidean* similarity measure and a *self-normalized* similarity measure related to the Canberra distance. It is proved that they are positive semi-definite (p.s.d.), thus facilitating their use in kernel-based methods, like the Support Vector Machine, a very popular machine learning tool. These kernels may be better suited than standard kernels (like the RBF) in certain situations, that are described in the paper. Some rather general results concerning positivity properties are presented in detail as well as some interesting ways of proving the p.s.d. property.

1 Introduction

One of the latest machine learning methods to be introduced is the Support Vector Machine (SVM). It has become very widespread due to its firm grounds in statistical learning theory (Vapnik (1998)) and its generally good practical results. Central to SVMs is the notion of *kernel function*, a mapping of variables from its original space to a higher-dimensional Hilbert space in which the problem is expected to be easier. Intuitively, the kernel represents the *similarity* between two data observations. In the SVM literature there are basically two common-place kernels for real vectors, one of which (popularly known as the RBF kernel) is based on the Euclidean distance between the two collections of values for the variables (seen as vectors).

Obviously not all two-place functions can act as kernel functions. The conditions for being a kernel function are very precise and related to the so-called *kernel matrix*

being positive semi-definite (p.s.d.). The question remains, how should the similarity between two vectors of (positive) real numbers be computed? Which of these similarity measures are valid kernels? There are many interesting possibilities that come from well-established distances that may share the property of being p.s.d. There has been little work on this subject, probably due to the widespread use of the initially proposed kernel and the difficulty of proving the p.s.d. property to obtain additional kernels.

In this paper we tackle this matter by examining two alternative distance-based similarity measures on vectors of real numbers and show the corresponding kernel matrices to be p.s.d. These two distance-based kernels could better fit some applications than the normal Euclidean distance and derived kernels (like the RBF kernel). The first one is a truncated version of the standard Euclidean metric in $\boldsymbol{R}$, which additionally extends some of Gower's work in Gower (1971). This similarity yields more sparse matrices than the standard metric. The second one is inversely related to the Canberra distance, well-known in data analysis (Chandon and Pinson (1971)). The motivation for using this similarity instead of the traditional Euclidean-based distance is twofold: (a) it is self-normalised, and (b) it scales in a log fashion, so that similarity is smaller if the numbers are small than if the numbers are big.

The paper is organized as follows. In Section 2 we review work in kernels and similarities defined on real numbers. The intuitive semantics of the two new kernels is discussed in Section 3. As main results, we intend to show some interesting ways of proving the p.s.d. property. We present them in full in Sections 4 and 5 in the hope that they may be found useful by anyone dealing with the difficult task of proving this property. In Section 6 we establish results for positive vectors which lead to kernels created as a combination of different one-dimensional distance-based kernels, thereby extending the RBF kernel.

2 Kernels and similarities defined on real numbers

We consider kernels that are similarities in the classical sense: strongly reflexive, symmetric, non-negative and bounded (Chandon and Pinson (1971)). More specifically, kernels k for positive vectors of the general form:

$$k(\mathbf{x},\mathbf{y}) = f\left(\sum_{j=1}^{n} g_j(d_j(x_j,y_j))\right), \tag{1}$$

where x_j, y_j belong to some subset of $\boldsymbol{R}$, $\{d_j\}_{j=1}^n$ are metric distances and $\{f, g_j\}_{j=1}^n$ are appropriate continuous and monotonic functions in $\boldsymbol{R}^+ \cup \{0\}$ making the resulting k a valid p.s.d. kernel. In order to behave as a similarity, a natural choice for the kernels k is to be distance-based. Almost invariably, the choice for distance-based real number comparison is based on the standard metric in $\boldsymbol{R}$. The aggregation of a number n of such distance comparisons with the usual 2-norm leads to Euclidean distance in $\boldsymbol{R}^n$. It is known that there exist inverse transformations

of this quantity (that can thus be seen as similarity measures) that are valid kernels. An example of this is the kernel:

$$k(\mathbf{x},\mathbf{y}) = \exp\{-\frac{||\mathbf{x}-\mathbf{y}||^2}{2\sigma^2}\},\ \mathbf{x},\mathbf{y} \in \boldsymbol{R}^n, \sigma \not\equiv 0 \in \boldsymbol{R}, \tag{2}$$

popularly known as the RBF (or Gaussian) kernel. This particular kernel is obtained by taking $d(x_j,y_j) = |x_j - y_j|, g_j(z) = z^2/(2\sigma_j^2)$ for non-zero σ_j^2 and $f(z) = e^{-z}$. Note that nothing prevents the use of different scaling parameters σ_j for every component. The decomposition need not be unique and is not necessarily the most useful for proving the p.s.d. property of the kernel.

In this work we concentrate on upper-bounded metric distances, in which case the partial kernels $g_j(d_j(x_j,y_j))$ are lower-bounded, though this is not a necessary condition in general. We list some choices for partial distances:

$$d_{TrE}(x_i,y_i) = \min\{1,|x_i-y_i|\} \qquad \text{(Truncated Euclidean)} \tag{3}$$

$$d_{Can}(x_i,y_i) = \frac{|x_i-y_i|}{x_i+y_i} \qquad \text{(Canberra)} \tag{4}$$

$$d(x_i,y_i) = \frac{|x_i-y_i|}{\max(x_i,y_i)} \qquad \text{(Maximum)} \tag{5}$$

$$d(x_i,y_i) = \frac{(x_i-y_i)^2}{x_i+y_i} \qquad \text{(squared } \chi^2\text{)} \tag{6}$$

Note the first choice is valid in $\boldsymbol{R}$, while the others are valid in $\boldsymbol{R}^+$. There is some related work worth mentioning, since other choices have been considered elsewhere: with the choice $g_j(z) = 1 - z$, a kernel formed as in (1) for the distance (5) appears as p.s.d. in Shawe-Taylor and Cristianini (2004). Also with this choice for g_j, and taking $f(z) = e^{z/\sigma}, \sigma > 0$ the distance (6), leads to a kernel that has been proved p.s.d. in Fowlkes et al. (2004).

3 Semantics and applicability

The distance in (3) is a truncated version of the standard metric in $\boldsymbol{R}$, which can be useful when differences greater than a specified threshold have to be ignored. In similarity terms, it models situations wherein data examples can become more and more similar until they are suddenly indistinguishable. Otherwise, it behaves like the standard metric in $\boldsymbol{R}$. Notice that this similarity may lead to more sparse matrices than those obtainable with the standard metric. The distance in (4) is called the Canberra distance (for one component). It is self-normalised to the real interval $[0,1)$, and is multiplicative rather than additive, being specially sensitive to small changes near zero. Its behaviour can be best seen by a simple example: let a variable stand for the number of children, then the distance between 7 and 9 is not the same

"psychological" distance than that between 1 and 3 (which is triple); however, $|7-9| = |1-3|$. If we would like the distance between 1 and 3 be much greater than that between 7 and 9, then this effect is captured. More specifically, letting $z = x/y$, then $d_{Can}(x,y) = g(z)$, where $g(z) = |z-1|/(z+1)$ and thus $g(z) = g(1/z)$. The Canberra distance has been used with great success in content-based image retrieval tasks in Kokare et al. (2003).

4 Truncated Euclidean similarity

Let x_i be an arbitrary finite collection of n different real points $x_i \in \boldsymbol{R}$, $i = 1,\ldots,n$. We are interested in the $n \times n$ similarity matrix $\mathbf{A} = (a_{ij})$ with

$$a_{ij} = 1 - d_{ij}, \quad d_{ij} = \min\{1, |x_i - x_j|\}, \tag{7}$$

where the usual Euclidean distances have been replaced by *truncated Euclidean distances*. We can also write $a_{ij} = (1-d_{ij})_+ = \max\{0, 1-|x_i - x_j|\}$.

Theorem 1. *The matrix* $\mathbf{A}$ *is positive definite (p.s.d.).*

PROOF. We define the bounded functions $X_i(x)$ for $x \in \boldsymbol{R}$ with value 1 if $|x - x_i| \leq 1/2$, zero otherwise. We calculate the interaction integrals

$$l_{ij} = \int_{\boldsymbol{R}} X_i(x) X_j(x) dx\,.$$

The value is the length of the interval $[x_i - 1/2, x_i + 1/2] \cap [x_j - 1/2, x_j + 1/2]$. It is easy to see that $l_{ij} = 1 - d_{ij}$ if $d_{ij} < 1$, and zero if $|x_i - x_j| \geq 1$ (i.e., when there is no overlapping of supports). Therefore, $l_{ij} = a_{ij} \quad$ if $i \neq j$. Moreover, for $i = j$ we have

$$\int_{\boldsymbol{R}} X_i(x) X_j(x)\, dx = \int X_i^2(x)\, dx = 1.$$

We conclude that the matrix $\mathbf{A}$ is obtained as the interaction matrix for the system of functions $\{X_i\}_{i=1}^N$. These interactions are actually the dot products of the functions in the functional space $L^2(\boldsymbol{R})$. Since a_{ij} is the dot product of the inputs cast into some Hilbert space it forms, by definition, a p.s.d. matrix.

Notice that rescaling of the inputs would allow us to substitute the two "1" (one) in equation (7) by any arbitrary positive number. In other words, the kernel with matrix

$$a_{ij} = (s - d_{ij})_+ = \max\{0, s - |x_i - x_j|\} \tag{8}$$

with $s > 0$ is p.s.d. The classical result for general Euclidean similarity in Gower (1971) is a consequence of this Theorem when $|x_i - x_j| \leq 1$ for all i, j.

5 Canberra distance-based similarity

We define the Canberra similarity between two points as follows

$$S_{Can}(x_i,x_j) = 1 - d_{Can}(x_i,x_j), \quad d_{Can}(x_i,x_j) = \frac{|\, x_i - x_j \,|}{x_i + x_j}, \tag{9}$$

where $d_{Can}(x_i,x_j)$ is called the *Canberra distance*, as in (4). We establish next the p.s.d. property for Canberra distance matrices, for $x_i,x_j \in \boldsymbol{R}^+$.

Theorem 2. *The matrix $A = (a_{ij})$ with $a_{ij} = S_{Can}(x_i,x_j)$ is p.s.d.*

PROOF. *First step.* Examination of equation (9) easily shows that for any $x_i,x_j \in \boldsymbol{R}^+$ (not including 0) the value of $s_{Can}(x_i,x_j)$ is the same for every pair of points x_i,x_j that have the same quotient x_i/x_j. This gives us the idea of taking logarithms on the input and finding an equivalent kernel for the translated inputs. From now on, define $x \equiv x_i, z \equiv x_j$, for clarity. We use the following straightforward result:

Lemma 1. *Let K' be a p.s.d. kernel defined in the region $B \times B$, let Φ be map from a region A into B, and let K be defined on $A \times A$ as $K(x,z) = K'(\Phi(x),\Phi(z))$. Then the kernel K is p.s.d.*

PROOF. Clearly Φ is a restriction of B, and K' is p.s.d in all $B \times B$.

Here, we take $K = S_{Can}$, $A = \boldsymbol{R}^+$, $\Phi(x) = \log(x)$, so that B is $\boldsymbol{R}$. We now find what K' would be by defining $x' = \log(x), \quad z' = \log(z)$, so that distance d_{Can} can be rewritten as

$$d_{Can}(x,z) = \frac{|\, x - z \,|}{x + z} = \frac{|\, e^{x'} - e^{z'} \,|}{e^{x'} + e^{z'}}.$$

As we noted above, $d_{Can}(x,z)$ is equivalent for any pair of points $x,z \in \boldsymbol{R}^+$ with the same quotients x/z or z/x. Assuming that $x > z$ without loss of generality, we write this as a *translation invariant* kernel by introducing the increment in logarithmic coordinates $h = |\, x' - z' \,| = x' - z' = \log(x/z)$:

$$d_{Can}(x,z) = \frac{e^{z'}e^h - e^{z'}}{e^{z'}e^h + e^{z'}} = \frac{e^h - 1}{e^h + 1}.$$

Substitution on $K = S_{Can}$ gives

$$S_{Can}(x,z) = 1 - \frac{e^h - 1}{e^h + 1} = \frac{2}{e^h + 1}$$

Therefore, for $x',z' \in \boldsymbol{R}$, $x' = z' + h$, we have

$$K'(x',z') = K'(x' - z') = \frac{2}{e^h + 1} = F(h). \tag{10}$$

Note that F is a convex function of $h \in [0,\infty)$ with $F(0) = 1$, $F(\infty) = 0$.

Second step. To prove our theorem we now only have to prove the p.s.d. property for kernel K' satisfying equation (10).

A direct proof uses an integral representation of convex functions that proceeds as follows. Given a twice continuously differentiable function F of the real variable $s \geq 0$, integrating by parts we find the formula

$$F(x) = -\int_x^\infty F'(s)\,ds = \int_x^\infty F''(s)(s-x)\,ds,$$

valid for all $x > 0$ on the condition that $F(s)$ and $sF'(s) \to 0$ as $s \to \infty$. The formula can be written as

$$F(x) = \int_0^\infty F''(s)(s-x)_+\,ds,$$

which implies that whenever $F'' > 0$, we have expressed $F(x)$ as an integral combination with positive coefficients of functions of the form $(s-x)_+$. This is a non-trivial, but commonly used, result in convex theory.

Third step. The functions of the form $(s-x)_+$ are the building blocks of the Truncated Euclidean Similarity kernels (7). Our kernel K' is represented as an integral combination of these functions with positive coefficients. In the previous Section we have proved that functions of the form (8) are p.s.d. We know that the sum of p.s.d. terms is also p.s.d., and the limit of p.s.d. kernels is also p.s.d. Since our expression for K' is, like all integrals, a limit of positive combinations of functions of the form $(s-x)_+$, the previous argument proves that equation (10) is p.s.d., and by Lemma 1 our theorem is proved. More precisely, what we say is that, as a convex function, F can be arbitrarily approximated by sums of functions of the type

$$f_n(x) = \max\{0, a_n - \frac{x}{r_n}\} \tag{11}$$

for $n \in [0,...,N]$, and the r_n equally spaced in the range of the input (so that the bigger the N the closer we get to (10)). Therefore, we can write

$$\frac{2}{e^h + 1} = \lim_{n\to\infty} \sum_{i=0}^{n} (a_i - \frac{h}{r_i})_+, \tag{12}$$

where each term in the succession (12) is of the form (11), equivalent to (8).

6 Kernels defined on real vectors

We establish now a result for positive vectors that leads to kernels analogous to the Gaussian RBF kernel. The reader can find useful additional material on positive and negative definite functions in Berg et al. 1984 (esp. Ch. 3).

Definition 1 (Hadamard function). *If $A = [a_{ij}]$ is a $n \times n$ matrix, the function $f : A \to f(A) = [f(a_{ij})]$ is called a Hadamard function (actually, this is the simplest type of Hadamard function).*

Theorem 3. *Let a p.s.d. matrix $A = [a_{ij}]$ and a Hadamard function f be given. If f is an analytic function with positive radius of convergence $R > |a_{ij}|$ and all the coefficients in its power series expansion are non-negative, then the matrix $f(A)$ is p.s.d. as proved in Horn and Johnson (1991).*

Definition 2 (p.s.d. function). *A real symmetric function $f(x,y)$ of real variables will be called p.s.d. if for any finite collection of n real numbers $x_1,...,x_n$, the $n \times n$ matrix A with entries $a_{ij} = f(x_i,x_j)$ is p.s.d.*

Lemma 2. *Let $b > 1 \in \boldsymbol{R}, c \in \boldsymbol{R}$ and let $c - f(x,y)$ be a p.s.d. function. Then $b^{-f(x,y)}$ is a p.s.d. function.*

PROOF. The function $x \to b^x$ is analytic with infinite radius of convergence and all the coefficients in its power series expansion are non-negative in case $b > 1$. By theorem (3) the function $b^{c-f(x,y)}$ is p.s.d.; then so is $b^c b^{-f(x,y)}$ and consequently $b^{-f(x,y)}$ is p.s.d. (since b^c is a positive constant).

Theorem 4. *The following function*

$$k(\mathbf{x},\mathbf{y}) = exp\left(-\sum\nolimits_{i=1}^{n} \frac{d(x_i,y_i)}{\sigma_i}\right), \qquad x_i, y_i, \sigma_i \in \boldsymbol{R}^+$$

where d is any of (3), (4), (5), is a valid p.s.d. kernel.

PROOF. For simplicity, make $d_i \equiv d(x_i,y_i)$. We know $1-d_i$ is a p.s.d. function, for the choices of d_i defined in (3), (4), (5). Therefore, $(1-d_i)/\sigma_i$ for $\sigma_i > 0 \in R$ is also p.s.d. Making $c \equiv \sum_{i=1}^{n} 1/\sigma_i$ and $f \equiv d_i/\sigma_i$, by lemma (2), the function $exp(-d_i/\sigma_i)$ is p.s.d. The product of p.s.d. functions is p.s.d., and thus $\prod_{i=1}^{n} exp(-d_i/\sigma_i) = exp\left(-\sum_{i=1}^{n} \frac{d_i}{\sigma_i}\right)$ is p.s.d.

This result is useful since it establishes new kernels analogous to the Gaussian RBF kernel but based on alternative metrics. Computational considerations should not be overlooked: the use of the exponential function considerably increases the cost of evaluating the kernel. Hence, kernels not involving this function are specially welcome.

Proposition 1. *Let $d(x_i,x_j) = \frac{|x_i - x_j|}{x_i + x_j}$ be the Canberra distance. Then $k(x_i,x_j) = 1 - d(x_i,x_j)/\sigma$ is a valid p.s.d. kernel if and only if $\sigma \geq 1$.*

PROOF. Let $d_{ij} \equiv d(x_i,x_j)$. We know $\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j (1-d_{ij}) \geq 0$ for all $c_i, c_j \in \boldsymbol{R}$. We have to show that $\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j (1 - \frac{d_{ij}}{\sigma}) \geq 0$. This can be expressed as $\sigma(\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j) \geq \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j d_{ij}$.

This result is a generalization of Theorem (2), valid for $\sigma = 1$. It is immediate that the following function (the *Canberra kernel*) is a valid kernel:

$$k(\mathbf{x},\mathbf{y}) = 1 - \frac{1}{n}\sum_{i=1}^{n} \frac{d_i(x_i,y_i)}{\sigma_i},\ \sigma_i \geq 1$$

The inclusion of the σ_i (acting as *learning parameters*) has the purpose of adding flexibility to the models. Concerning the truncated Euclidean distance, a corresponding kernel can be obtained in a similar way. Let $d(x_i,x_j) = min\{1,|x_i - x_j|\}$ and denote for a real number a, $a_+ \equiv 1 - min(1,a) = max(0,1-a)$. Then $\sigma - min\{\sigma,|x_i - x_j|\}$ is p.s.d. by Theorem (1) and so is $max\{0,1-\frac{|x_i-x_j|}{\sigma}\}$. In consequence, it is immediate to affirm that the following function (the *Truncated Euclidean kernel*) is again a valid kernel:

$$k(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{d_i(x_i,y_i)}{\sigma_i}\right)_+,\ \sigma_i > 0$$

7 Conclusions

We have considered distance-based similarity measures for real-valued vectors of interest in kernel-based methods, like the Support Vector Machine. The first is a truncated Euclidean similarity and the second a self-normalized similarity. Derived real kernels analogous to the RBF kernel have been proposed, so the kernel toolbox is widened. These can be considered as suitable alternatives for a proper modeling of data affected by multiplicative noise, skewed data and/or containing outliers. In addition, some rather general results concerning positivity properties have been presented in detail.

Acknowledgments

Supported by the Spanish project CICyT CGL2004-04702-C02-02.

References

BERG, C. CHRISTENSEN, J.P.R. and RESSEL, P. (1984): *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions,* Springer.

CHANDON, J.L. and PINSON, S. (1981): *Analyse Typologique. Théorie et Applications,* Masson, Paris.

FOWLKES, C., BELONGIE, S., CHUNG, F., and MALIK. J. (2004): Spectral Grouping Using the Nyström Method. *IEEE Trans. on PAMI*, 26(2), 214–225.

GOWER. J.C. (1971): A general coefficient of similarity and some of its properties, *Biometrics* 27, 857–871.

HORN, R.A. and JOHNSON, C.R. (1991): *Topics in Matrix Analysis*, Cambridge University Press.

KOKARE, M., CHATTERJI, B.N. and BISWAS, P.K. (2003): Comparison of similarity metrics for texture image retrieval. In: *IEEE Conf. on Convergent Technologies for Asia-Pacific Region*, 571–575.

SHAWE-TAYLOR, J. and CRISTIANINI, N. (2004): *Kernel Methods for Pattern Analysis*, Cambridge University Press.

VAPNIK. V. (1998): *The Nature of Statistical Learning Theory*. Springer-Verlag.