

The Aitchison and Aitken Kernel Function Revisited

Hamse Y. Mussa¹

¹ Unilever Centre for Molecular Science Informatics, Department of Chemistry, Lensfield Road, Cambridge, United Kingdom

Correspondence: H. Y. Mussa, Unilever Centre for Molecular Science Informatics, Department of Chemistry, Lensfield Road, Cambridge, CB2 1EW, United Kingdom. Tel: 44-1223-763-854. E-mail: hym21@cam.ac.uk

Received: September 26, 2012 Accepted: January 30, 2013 Online Published: February 20, 2013

doi:10.5539/jmr.v5n1p22 URL: <http://dx.doi.org/10.5539/jmr.v5n1p22>

Abstract

Over three decades ago Aitchison and Aitken proposed a novel kernel function for estimating the density functions of underlying distributions in discrete input spaces. To the best of our knowledge, it has not been shown whether this kernel function is positive definite (*i.e.*, a reproducing kernel function) on these spaces. Its positive definiteness would have enriched and enlarged its applicability domain: a positive definite kernel function has an associated Reproducing Kernel Hilbert Space, a framework on which a variety of powerful statistical and machine learning schemes can be developed.

This paper aims to demonstrate that Aitchison and Aitken's kernel function is indeed positive definite on discrete metric spaces. We also touch on possible applications of the proposed theorem.

Keywords: positive definite, kernel function, reproducing kernel Hilbert spaces, discrete input spaces

1. Introduction

In a seminal paper Aitchison and Aitken (1976) proposed a kernel function defined on discrete descriptor/input spaces. These authors introduced this kernel function, which is henceforth referred to as the AA-kernel, for estimating density functions in binary input spaces (Aitchison & Aitken, 1976). Its simple non-parametric nature together with its consistency properties have made the AA-kernel a useful tool for generating discriminant functions. For example, classifiers based on the AA-kernel have recently been widely employed in cheminformatics classification problems where the molecules are represented by a zero-one (*i.e.*, binary) variables (Harper et al., 2001; Hert et al., 2004; Wilton et al., 2006; Lowe et al., 2011). Furthermore, in the past few years R-packages featuring the AA-kernel have started to appear in the literature.

According to Aronszajn (1950), to every positive definite kernel function (PDKF)-in the sense defined below-on $X \times X$ there corresponds a unique Reproducing Kernel Hilbert Space (RKHS) on X , where X can be any non-empty set (Wahba, 1998). RKHS provides a general framework on which a diverse set of powerful data analysis tools (the so-called Reproducing Kernel Hilbert Space Methods) can be developed, whereby density function estimations, the widely popular Support Vector Machines (Vapnik, 1995) and function approximations from finite data, to name but a few, can be viewed as special cases (Poggio & Girosi, 1997; Wahba, 1990; Hofmann et al., 2008).

The AA-kernel, which is defined on a discrete metric space, is not a Gaussian function in this space, but it can be viewed as the counterpart of a Gaussian kernel function defined on an Euclidean ("standard") metric space (Aitchison & Aitken, 1976). It is well documented that a Gaussian kernel function is PDKF on standard metric spaces (Berg, 1998), but the same cannot be said for the AA-kernel on its discrete metric space. To the best of the author's knowledge, it has not been demonstrated whether (or not) the AA-kernel is positive definite-a "positive result" would have significantly enlarged the applicability domain of the AA-kernel: As stated in the preceding paragraph, for a positive definite kernel function there is a corresponding unique Reproducing Kernel Hilbert Space. This means that if one proves the AA-kernel to be positive definite, then general probabilistic or deterministic data analysis models based on Reproducing Kernel Hilbert Space defined on discrete metric input spaces can be devised.

The following section gives the main definition and several important properties of PDKFs, which are relevant to the topic addressed in this paper. Also in this section the AA-kernel is defined. In Section 3, we demonstrate that the AA-kernel is positive definite. The final section gives our concluding remarks citing possible applications of

the theorem proposed in this paper.

2. The AA-Kernel and Positive Definite Kernel Function (PDKF)

In the context of the work presented in this paper, a kernel function is a two-input symmetrical function (Shawe-Taylor & Cristianini, 2004, Chapter 3).

The AA-kernel is a two-input symmetrical function given as (Aitchison & Aitken, 1976)

$$K(\mathbf{x}_i, \mathbf{x}_j; \lambda) = (\lambda)^{n-d(\mathbf{x}_i, \mathbf{x}_j)} \left(\frac{1-\lambda}{c-1} \right)^{d(\mathbf{x}_i, \mathbf{x}_j)} \quad (1)$$

where \mathbf{x}_i and \mathbf{x}_j are binary variables $\in \mathcal{X} = \mathcal{B}^n$ with $\mathcal{B} = \{0, 1, \dots, c-1\}$; $0.5 \leq \lambda \leq 1$, and n and c (≥ 2) refer to the number of discrete entries that each variable has and the categories that an entry can assume, respectively; and $d(\mathbf{x}_i, \mathbf{x}_j)$ is a discrete metric defined on \mathcal{B}^n , i.e., it denotes the number of disagreements in corresponding elements of \mathbf{x}_i and \mathbf{x}_j .

In this work, for clarity and without loss of generality, we set c to 2, i.e., $\mathcal{B} = \{0, 1\}$. The two λ values, $\lambda = 1$ and $\lambda = 0.5$, lead to two extreme forms of density estimations: uniform distribution and relative frequencies of appearance of the given discrete data, respectively. Thus these two values of λ are ignored—that is, $0.5 \leq \lambda \leq 1$ becomes $0.5 < \lambda < 1$ (see Aitchison & Aitken, 1976).

Expressing $d(\mathbf{x}_i, \mathbf{x}_j)$ as

$$d(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j) \quad (2)$$

proffers a simple way to compute the value of $d(\mathbf{x}_i, \mathbf{x}_j)$. It certainly does not imply that \mathcal{B}^n is a normed space. Instead, here $(\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j)$ merely represents a convenient scheme to calculate $d(\mathbf{x}_i, \mathbf{x}_j)$.

Having defined and described the AA-kernel, for completeness we now briefly discuss what a positive definite Kernel Function (PDKF) is. We also cite a number of useful properties of PDKFs, which we deem most relevant for the purpose of this paper.

Definition (Wahba, 1998) *A two-argument symmetric function $K(\mathbf{x}_i, \mathbf{x}_j; \lambda)$ is said to be positive definite kernel function on $\mathcal{X} \times \mathcal{X}$ if for any N and any N (data) points $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathcal{X}$ the $N \times N$ matrix with elements $K(\mathbf{x}_i, \mathbf{x}_j; \lambda)$, $\sum_{i,j}^N \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j; \lambda) \geq 0$.*

Where $\alpha_i, \alpha_j \in \mathcal{R}$, λ is a real-valued tunable smoothing parameter and \mathcal{X} being any non-empty set.

Note that in the case of the AA-kernel \mathcal{X} is \mathcal{B}^n , i.e., \mathbf{x}_i and \mathbf{x}_j can be considered as n -dimensional vectors.

Before proceeding further to show that the AA-kernel is a PDKF, which constitutes the core objective of this paper, a highly useful proposition is provided. The proposition encapsulates several important closure properties of PDKFs, which are relevant for the purpose of this paper. The proof of this proposition can be found in Shawe-Taylor and Cristianini's book (2004).

Proposition *If g_1, g_2 , and h are PDKFs over $\mathcal{X} \times \mathcal{X}$, $a, \gamma \in \mathcal{R}^+$, $f(\cdot)$ is a real-valued function on \mathcal{X} , and \mathbf{x}_i and $\mathbf{x}_j \in \mathcal{X}$, then so are the following PDKFs in the sense defined above:*

A1 $K(\mathbf{x}_i, \mathbf{x}_j) = g_1(\mathbf{x}_i, \mathbf{x}_j) \times g_2(\mathbf{x}_i, \mathbf{x}_j)$.

A2 $K(\mathbf{x}_i, \mathbf{x}_j) = a + \gamma h(\mathbf{x}_i, \mathbf{x}_j)$.

A3 $K(\mathbf{x}_i, \mathbf{x}_j) = a f(\mathbf{x}_i) f(\mathbf{x}_j)$.

A4 $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$ is a linear kernel function.

3. AA-Kernel Function Is Positive Definite

This section constitutes the nub of the paper. First we formulate a theorem stating that the AA-kernel is positive definite. We then provide the full proof of the theorem.

Theorem 1 *If $0.5 < \lambda < 1$, and \mathbf{x}_i and $\mathbf{x}_j \in \mathcal{B}^n$ with $\mathcal{B} = \{0, 1\}$, then $K(\mathbf{x}_i, \mathbf{x}_j; \lambda) = (\lambda)^{n-d(\mathbf{x}_i, \mathbf{x}_j)} (1-\lambda)^{d(\mathbf{x}_i, \mathbf{x}_j)}$ is a positive definite kernel function on $\mathcal{B}^n \times \mathcal{B}^n$.*

Proof. Given

$$K(\mathbf{x}_i, \mathbf{x}_j; \lambda) = (\lambda)^{n-d(\mathbf{x}_i, \mathbf{x}_j)} (1-\lambda)^{d(\mathbf{x}_i, \mathbf{x}_j)} \quad (3)$$

which-using Equation 2-can be rewritten as

$$K(\mathbf{x}_i, \mathbf{x}_j; \lambda) = \lambda^{n-(\mathbf{x}_i-\mathbf{x}_j)^T (\mathbf{x}_i-\mathbf{x}_j)} (1-\lambda)^{(\mathbf{x}_i-\mathbf{x}_j)^T (\mathbf{x}_i-\mathbf{x}_j)} \quad (4)$$

then after some simple algebraic manipulations, Equation 4 becomes

$$K(\mathbf{x}_i, \mathbf{x}_j; \lambda) = \lambda^n \left(\frac{1-\lambda}{\lambda}\right)^{\mathbf{x}_i^T \mathbf{x}_i} \left[\left(\frac{1-\lambda}{\lambda}\right)^{-\mathbf{x}_i^T \mathbf{x}_j}\right] \left[\left(\frac{1-\lambda}{\lambda}\right)^{-\mathbf{x}_j^T \mathbf{x}_i}\right] \left(\frac{1-\lambda}{\lambda}\right)^{\mathbf{x}_j^T \mathbf{x}_j} \quad (5)$$

Let $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ denote $\left(\frac{1-\lambda}{\lambda}\right)^{\mathbf{x}_i^T \mathbf{x}_i}$ and $\left(\frac{1-\lambda}{\lambda}\right)^{\mathbf{x}_j^T \mathbf{x}_j}$, respectively. This gives

$$\begin{aligned} K(\mathbf{x}_i, \mathbf{x}_j; \lambda) &= \lambda^n f(\mathbf{x}_i) \left[\left(\frac{1-\lambda}{\lambda}\right)^{-\mathbf{x}_i^T \mathbf{x}_j} \left(\frac{1-\lambda}{\lambda}\right)^{-\mathbf{x}_j^T \mathbf{x}_i}\right] f(\mathbf{x}_j) \\ &= \lambda^n f(\mathbf{x}_i) \left(1 - \frac{2\lambda-1}{\lambda}\right)^{-\mathbf{x}_i^T \mathbf{x}_j} \left(1 - \frac{2\lambda-1}{\lambda}\right)^{-\mathbf{x}_j^T \mathbf{x}_i} f(\mathbf{x}_j) \\ &= \lambda^n f(\mathbf{x}_i) f(\mathbf{x}_j) \left[\left(1 - \frac{2\lambda-1}{\lambda}\right)^{-\mathbf{x}_i^T \mathbf{x}_j} \left(1 - \frac{2\lambda-1}{\lambda}\right)^{-\mathbf{x}_j^T \mathbf{x}_i}\right] \end{aligned} \quad (6)$$

where $\left(1 - \frac{2\lambda-1}{\lambda}\right)^{-\mathbf{x}_i^T \mathbf{x}_j}$ and $\left(1 - \frac{2\lambda-1}{\lambda}\right)^{-\mathbf{x}_j^T \mathbf{x}_i}$ are $\left(\frac{1-\lambda}{\lambda}\right)^{-\mathbf{x}_i^T \mathbf{x}_j}$ and $\left(\frac{1-\lambda}{\lambda}\right)^{-\mathbf{x}_j^T \mathbf{x}_i}$, respectively.

Based on A3., $\lambda^n f(\mathbf{x}_i) f(\mathbf{x}_j)$ is PDKF. This means $K(\mathbf{x}_i, \mathbf{x}_j; \lambda)$ is PDKF if and only if $\left(1 - \frac{2\lambda-1}{\lambda}\right)^{-\mathbf{x}_i^T \mathbf{x}_j}$ and $\left(1 - \frac{2\lambda-1}{\lambda}\right)^{-\mathbf{x}_j^T \mathbf{x}_i}$ are PDKFs.

One only requires to demonstrate that $\left(1 - \frac{2\lambda-1}{\lambda}\right)^{-\mathbf{x}_i^T \mathbf{x}_j}$ is PDKF and then use the same argument for $\left(1 - \frac{2\lambda-1}{\lambda}\right)^{-\mathbf{x}_j^T \mathbf{x}_i}$.

By definition $0.5 < \lambda < 1$; hence $0 < \frac{2\lambda-1}{\lambda} < 1$. Then by invoking the binomial theorem, one can express $\left(1 - \frac{2\lambda-1}{\lambda}\right)^{-\mathbf{x}_i^T \mathbf{x}_j}$ as

$$\begin{aligned} \left(1 - \frac{2\lambda-1}{\lambda}\right)^{-\mathbf{x}_i^T \mathbf{x}_j} &= \left(1 - \frac{2\lambda-1}{\lambda}\right)^{-q} \\ &= 1 + (-q)(-\beta) + \frac{(-q)((-q)-1)}{2!}(-\beta)^2 + \frac{(-q)((-q)-1)((-q)-2)}{3!}(-\beta)^3 + \dots \\ &= 1 + q\beta + \frac{q(q+1)}{2!}\beta^2 + \frac{q(q+1)(q+2)}{3!}\beta^3 + \dots \\ &= 1 + \gamma_1 q + \gamma_2 q(q+1) + \gamma_3 q(q+1)(q+2) + \dots \end{aligned} \quad (7)$$

where β , γ_m and q denote $\frac{2\lambda-1}{\lambda}$, $\frac{\beta^m}{m!}$ and $q = q(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$ (a linear kernel function), respectively; with m being a positive integer.

On the RHS of Equation 7, in the final expression, all coefficients $\gamma_m \in \mathcal{R}^+$. This means, based on A1, A2 and A4, that $1 + \gamma_1 q + \gamma_2 q(q+1) + \gamma_3 q(q+1)(q+2) + \dots$ is PDKF. In other words, $\left(1 - \frac{2\lambda-1}{\lambda}\right)^{-\mathbf{x}_i^T \mathbf{x}_j}$ is a positive definite kernel function. Hence the AA-kernel, $K(\mathbf{x}_i, \mathbf{x}_j; \lambda)$, that we started with is positive definite. This completes the proof of Theorem 1.

4. Summary

Over three decades ago Aitchison and Aitken proposed a novel kernel function for estimating the density functions of underlying distributions in discrete metric spaces. To the best of our knowledge, it has not been shown whether this kernel function is positive definite on discrete metric spaces. The positive definiteness of this kernel function would have enriched and enlarged its applicability domain, because a PDKF has an associated Reproducing Kernel Hilbert Space (RKHS). A RKHS provides an excellent framework on which a variety of powerful statistical and machine learning schemes can be developed as discussed at great length and detail in some of the references cited in Section 1.

We therefore anticipate that the proposed and proven theorem in this paper can be applied wherever (in statistics or machine learning) the application of models based on the RKHS concept deemed appropriate for the analysis of discrete datasets.

Acknowledgements

The author would like to thank Unilever for financial support.

References

- Aitchison, J., & Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, *63*, 413-420. <http://dx.doi.org/10.1093/biomet/63.3.413>
- Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, *68*, 337-404. Retrieved from <http://www.ams.org/journals/tran/1950-068-03/S0002-9947-1950-0051437-7/>
- Berg, C., Christensen, J. P. R., & Ressel, P. J. (1984). *A Classical Introduction to Modern Number Theory (Graduate Texts in Mathematics)* (1st ed.). New York, NY: Springer-Verlag.
- Harper, G., Bradshaw, J., Gittins, J. C., Green, D. V. S., & Leach, A. R. (2001). Prediction of biological activity for high-throughput screening using binary kernel discrimination. *J. Chem. Inf. Comput. Sci.*, *41*, 1295-1300. <http://dx.doi.org/10.1021/ci000397q>
- Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., & Schuenhauer, A. (2004). A comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.*, *2*, 3256-3266. <http://dx.doi.org/10.1039/B409865J>
- Hofmann, T., Scholkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *Annals of Statistics*, *36*, 1171-1220. <http://dx.doi.org/10.1214/009053607000000677>
- Lowe, R., Mussa, H. Y., Mitchell, J. B. O., & Glen, R. C. (2011). Classifying molecules using a sparse probabilistic kernel binary classifier. *J. Chem. Inf. Model.*, *51*, 1539-1544. <http://dx.doi.org/10.1021/ci200128w>
- Poggio, T., & Girosi, F. (1988). A sparse representation for function approximation. *Neural Computation*, *10*, 1445-1454. <http://dx.doi.org/10.1162/089976698300017250>
- Shawe-Taylor, J., & Cristianini N. (2004). *Kernel Methods for Pattern Analysis* (1st ed., pp. 75-76). Cambridge, UK: Cambridge University Press.
- Vapnik, V. (1995). *The Nature of Statistical Learning Theory* (1st ed.). New York, NY: Springer-Verlag.
- Wahba, G. (1990). *Spline Models for Observational Data* (1st ed.). SIAM.
- Wahba, G. (1998). Support vector machines, reproducing kernel Hilbert spaces and randomized gacv. *Tech. rep.* University of Wisconsin, Wisconsin.
- Wilton, D. J., Harrison, R. F., Willett, P., Delaney, J., Lawson, K., & Mullier, G. (2006). Virtual screening using binary kernel discrimination: Analysis of pesticide data. *J. Chem. Inf. Model.*, *46*, 471-477. <http://dx.doi.org/10.1021/ci050397w>