Open problem: kernel methods on manifolds and metric spaces

Feragen, Aasa; Hauberg, Søren

Published in:
29th Annual Conference on Learning Theory

Publication date:
2016

Document version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Feragen, A., & Hauberg, S. (2016). Open problem: kernel methods on manifolds and metric spaces: what is the probability of a positive definite geodesic exponential kernel? In V. Feldman, A. Rakhlin, & O. Shamir (Eds.), *29th Annual Conference on Learning Theory* (pp. 1647–1650). JMLR: Workshop and Conference Proceedings, Vol.. 49

JMLR: Workshop and Conference Proceedings vol 49:1–4, 2016

Open Problem: Kernel methods on manifolds and metric spaces

What is the probability of a positive definite geodesic exponential kernel?

Aasa Feragen AASA@DI.KU.DK
University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark

Søren Hauberg SOHAU@DTU.DK
Technical University of Denmark, Richard Petersens Plads, Bld. 321, 2800 Kgs. Lyngby, Denmark

Abstract

Radial kernels are well-suited for machine learning over general geodesic metric spaces, where pairwise distances are often the only computable quantity available. We have recently shown that geodesic exponential kernels are only positive definite for all bandwidths when the input space has strong linear properties. This negative result hints that radial kernel are perhaps not suitable over geodesic metric spaces after all. Here, however, we present evidence that large intervals of bandwidths exist where geodesic exponential kernels have high probability of being positive definite over finite datasets, while still having significant predictive power. From this we formulate conjectures on the probability of a positive definite kernel matrix for a finite random sample, depending on the geometry of the data space and the spread of the sample.

Keywords: Kernel methods, geodesic metric spaces, geodesic exponential kernel, positive definiteness, curvature, bandwidth selection.

1. Introduction

In a number of applications, learning can be improved by incorporating domain-specific knowledge that constrains the data to reside on a nonlinear subspace such as a Riemannian manifold. Examples include *diffeomorphism groups* in computational anatomy (Grenander and Miller, 1998) and probability distributions under the Fisher information metric (Amari and Nagaoka, 2000). In such nonlinear spaces, the geodesic distance is often the only computable quantity and methods that solely depend on pairwise distances are therefore practical. Manifold statistics (Fletcher and Joshi, 2004) are popular for such data, but in practice often suffer from poor numerical precision and do not even scale to medium-size data sets. Kernel methods (Schölkopf and Smola, 2002) provide an attractive alternative, where the most common kernel is the *geodesic exponential kernel*:

$$k(x, x') = \exp\left(-\lambda d^q(x, x')\right), \quad q \in \mathbb{R}_+, \tag{1}$$

where d is the *geodesic* distance metric defined by shortest path length in X. These kernels can be defined on any geodesic data space, not just manifolds.

We have shown (Feragen et al., 2015) that Gaussian kernels ($q = 2$) are only PD for all λ if the metric space is flat in the sense of Alexandrov (Bridson and Haefliger, 1999). Similar restrictions are shown for other q. This **suggests that geodesic exponential kernels are not generally useful in nonlinear spaces**. Below, however, we show that for finite datasets, there exist intervals of bandwidths λ for Eq. 1 for which kernel matrices are PD. Moreover, in our simulations, these intervals contain bandwidths with good predictive power. This raises a series of open questions.

2. State-of-the-art

There is a vast literature on generalizing kernels to non-Euclidean data spaces, roughly falling into three approaches:

- The nonlinear data space X is linearized and a kernel is designed over the linear approximation space (Courty et al., 2012; Jaakkola and Haussler, 1998). This discards the nonlinear structure of X and thereby the domain specific knowledge it encodes.
- X is embedded in a higher-dimensional Euclidean space, over which the kernel is designed (Harandi et al., 2014; Jayasumana et al., 2014). Since the distance measure in the embedding space can arbitrarily depart from the geodesic distance, this approach also discards the domain specific knowledge that X was designed to capture.
- A kernel is designed directly on X through the geodesic distance, e.g. using Eq. 1 (Chapelle et al., 1999; Jayasumana et al., 2015, 2013; Harandi and Salzmann, 2015). The fact that such kernels are not PD (Feragen et al., 2015) for all bandwidths is ignored, and the analysis proceeds as if the kernel were, in fact, PD. This strategy has recently attracted wide attention in computer vision.

The first two approaches discard the domain specific knowledge that the data space was supposed to encode, while the third approach violates the fundamental assumption made by kernel methods. As a consequence, the following statistical analysis is not guaranteed to be well-defined.

3. Open Problems

Losing the "for all $\lambda > 0$" condition appears detrimental, because we lose the ability to freely train the bandwidth parameter λ. In practice, however, we find that there often exist large intervals of λ parameters that give PD kernel matrices for concrete finite datasets. Fig. 1 shows two simulation studies of Gaussian kernels ($q = 2$ in Eq. 1) on the unit sphere. On the left, we test PD'ness by plotting the minimum eigenvalue of the kernel matrix $K = [k(x_i, x_j)]_{ij}$, which is positive if and only if K is PD (Schölkopf and Smola, 2002). We observe a "PD interval" of λ parameters. This can be explained theoretically: When $\lambda \to \infty$, the kernel matrix approaches the identity matrix, whose minimum eigenvalue is 1. As the "minimum eigenvalue" function is continuous, there must be some λ' such that the kernel matrix K is PD for $\lambda > \lambda'$. On the right panel of Fig. 1, the blue curve estimates how reliable this interval is by sampling 200 points from two distributions on the unit sphere, computing their Gaussian kernel matrix, and checking for PD'ness for a range of λ parameters. We repeat this 300 times and plot the percentage of PD matrices for each λ value. Note that the percentage quickly approaches 100.

These simulations suggest that our current theory must be refined, and we conjecture:

Conjecture 1 *There exist conditions on the geometry of the data space X, the spread of the data, the exponent $q \leq 2$, the PD range of λ parameters and the sample size N such that for a random sample $\{x_1, \ldots, x_N\} \subset X$, and a fixed $\varepsilon > 0$, the kernel matrix $[\exp(-\lambda d^q(x_i, x_j))]_{ij}$ is PD with probability* 1-ε.

Conjecture 1 would be very powerful as it allows optimizing λ for specific tasks, and gives insight into the expected generalizability of a PD interval to an unseen test set. Moreover, we conjecture:

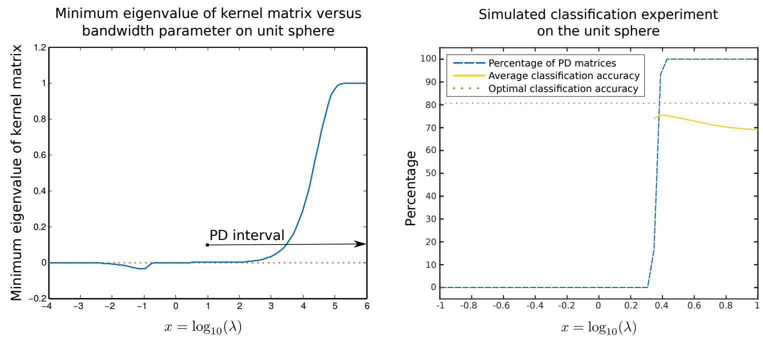

Fig. 1 *We show two simulation experiments on the unit sphere.* **Left:** *The minimal eigenvalue of the geodesic Gaussian kernel matrix for* 100 *uniformly sampled data points.* **Right:** *We compute the kernel matrix for a dataset with* 100 *points drawn from each of the Gaussian distributions* $\mathcal{N}([1/2, 1/2, 1/2], I)$ *and* $\mathcal{N}(-[1/2, 1/2, 1/2], I)$ *projected onto the unit sphere. This experiment is repeated* 300 *times for each of a range of bandwidths* λ*, and we report the percentage of PD geodesic Gaussian kernel matrices (blue), and the average cross-validated classification accuracy with a support vector machine among those kernel matrices that are PD (yellow). For reference, the expected classification accuracy for the optimal separating hyperplane of the generating distributions are shown in red.*

Conjecture 2 *The PD range of* λ *parameters depends on the geometry of the data space* X*:*

a) The PD range depends on the curvature of X*, defined in the* $CAT(\kappa)$ *sense (Bridson and Haefliger, 1999).*

b) The PD range depends on the distortion of the metric of the nonlinear data space under a low-distortion embedding into Euclidean spaces.

Conjecture 2, along with simple tests for curvature and recent results on embeddability into Euclidean spaces (Sidiropoulos and Wang, 2015; Matousek and Sidiropoulos, 2010), would lead to simple, testable conditions for positive definiteness, or for the size of the positive definite range of λ parameters. In particular, Conjecture 2b may lead to simple tests for whether lacking PD'ness is a general property of the data, or caused by a select few outliers.

If the bounding λ' is too large, then all PD kernel matrices are very close to I and therefore non-informative. However, our second simulation shown by the yellow curve in Fig. 1 shows that there are bandwidth parameters that simultaneously lead to a high probability of a PD kernel matrix, *and* a close to optimal classification accuracy.

Conjecture 3 *The PD range of* λ *parameters is* useful*, and its usefulness depends on the power* q*.*

Clearly there is a dependence on q because, as when $q \to 0$, the kernel matrix approaches a matrix of ones, which is non-informative. Conjecture 3 seeks to quantify this dependence.

What sets these conjectures apart from the current mindset is the formulation in terms of *probability* of positive definiteness in place of a deterministic worst-case analysis. We believe this probabilistic analysis will be intimately connected with the geometry of the data space.

Acknowledgments

We thank Oswin Krause for insightful discussions. S.H. is funded by the Danish Council for Independent Research, Natural Sciences.

References

S. Amari and H. Nagaoka. *Methods of information geometry*. Translations of mathematical monographs; v. 191. American Mathematical Society, 2000.

M.R. Bridson and A. Haefliger. *Metric spaces of non-positive curvature*. Springer, 1999.

O. Chapelle, P. Haffner, and V.N. Vapnik. Support vector machines for histogram-based image classification. *Neural Networks, IEEE Transactions on*, 10(5):1055–1064, 1999.

N. Courty, T. Burger, and P.-F. Marteau. Geodesic analysis on the Gaussian RKHS hypersphere. In *ECML PKDD*, volume 7523 of *LNCS*, pages 299–313. 2012.

A. Feragen, F. Lauze, and S. Hauberg. Geodesic exponential kernels: When curvature and linearity conflict. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2015.

P.T. Fletcher and S.C. Joshi. Principal geodesic analysis on symmetric spaces: Statistics of diffusion tensors. In *Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis, Revised Selected Papers*, pages 87–98, 2004.

U. Grenander and M.I. Miller. Computational anatomy: An emerging discipline. *Q. Appl. Math.*, LVI(4):617–694, December 1998.

M. Harandi and M. Salzmann. Riemannian coding and dictionary learning: Kernels to the rescue. In *CVPR*, pages 3926–3935, Boston, USA, jun 2015.

M. Harandi, M. Salzmann, S. Jayasumana, R. Hartley, and H. Li. Expanding the family of grassmannian kernels: An embedding perspective. In *European Conference on Computer Vision (ECCV)*, pages 408–423, 2014.

T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems (NIPS)*, pages 487–493, 1998.

S. Jayasumana, M. Salzmann, H. Li, and M. Harandi. A Framework for Shape Analysis via Hilbert Space Embedding. In *International Conference on Computer Vision (ICCV)*, 2013.

S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi. Optimizing over radial kernels on compact manifolds. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3802–3809, 2014.

S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi. Kernel Methods on Riemannian Manifolds with Gaussian RBF Kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015.

J. Matousek and A. Sidiropoulos. Inapproximability for metric embeddings into $\mathbb{R}^d$. *IEEE Symposium on Foundations of Computer Science (FOCS 2008); Transactions of the AMS*, 2010.

B. Schölkopf and A.J. Smola. *Learning with kernels : support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. MIT Press, 2002.

A. Sidiropoulos and Y. Wang. Metric embeddings with outliers. *ArXiV preprint*, `http://arxiv.org/abs/1508.03600`, 2015.