

UC Irvine

UC Irvine Previously Published Works

Title

Probabilistic Kernel Support Vector Machines

Permalink

<https://escholarship.org/uc/item/8kh7d8jp>

Authors

Chen, Yongxin
Georgiou, Tryphon T
Tannenbaum, Allen R

Publication Date

2019-04-14

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

PROBABILISTIC KERNEL SUPPORT VECTOR MACHINES

YONGXIN CHEN, TRYPHON T. GEORGIU, AND ALLEN R. TANNENBAUM

ABSTRACT. We propose a probabilistic enhancement of standard *kernel Support Vector Machines* for binary classification, in order to address the case when, along with given data sets, a description of uncertainty (e.g., error bounds) may be available on each datum. In the present paper, we specifically consider Gaussian distributions to model uncertainty. Thereby, our data consist of pairs (x_i, Σ_i) , $i \in \{1, \dots, N\}$, along with an indicator $y_i \in \{-1, 1\}$ to declare membership in one of two categories for each pair. These pairs may be viewed to represent the mean and covariance, respectively, of random vectors ξ_i taking values in a suitable linear space (typically \mathbb{R}^n). Thus, our setting may also be viewed as a modification of Support Vector Machines to classify distributions, albeit, at present, only Gaussian ones. We outline the formalism that allows computing suitable classifiers via a natural modification of the standard “kernel trick.” The main contribution of this work is to point out a suitable kernel function for applying Support Vector techniques to the setting of uncertain data for which a detailed uncertainty description is also available (herein, “Gaussian points”).

1. INTRODUCTION AND MOTIVATION

The Support Vector methodology is a nonlinear generalization of Linear Classification, typically utilized for binary classification. In this methodology, a training set of data becomes available where each datum belongs to one of two categories while an indicator function identifies the category for each. Imagine points on a plane labeled with two different colors. Solely based on their location and proximity to others of particular colors, one needs to classify new points and assign them to one of the two categories. When the points of the training set in the two categories can be separated by a line, then this line can be chosen to separate two respective regions, one corresponding to each of the two categories. On the other hand, when the “training” points of the two categories cannot be separated by a line, then a higher order curve needs to be chosen. The Support Vector methodology provides a disciplined and systematic way to do exactly that.

In the case where no simple boundary can be chosen to delineate regions for the points in the two categories, and a simple classification rule is still sought based on location, then a smooth boundary can be selected to define regions that contain the vast majority of points in the respective

Y. Chen is with the Department of Aerospace Engineering, Georgia Tech, Atlanta, GA; yongchen@gatech.edu.

T.T. Georgiou is with the Department of Mechanical Engineering, UCI, Irvine, CA; tryphon@uci.edu.

A.R. Tannenbaum is with Departments of Computer Science and Applied Mathematics & Statistics, Stony Brook University, NY arobertan@cs.stonybrook.edu.

categories. Thus, simplicity of the classification rule (Occam’s razor) may be desirable, as the domains corresponding to the two categories may not be completely separate and a misclassification error in that case must be deemed acceptable. This again can be handled by a suitable relaxation of SVM.

The purpose of this paper is to raise and in some ways address the issue that, in most engineering applications, data is uncertain. Measurements are often provided along with error bars that quantify expected margins. In other applications, data may yet be more complex. For instance, a datum may be an empirical probability distributions. Such distributions may be approximated by a Gaussian, a mixture of Gaussians, or other options. In such cases, a “datum” is now a point in a space more complex than \mathbb{R}^n . We seek to formulate a new paradigm, where the Support Vector methodology can be extended.

We postulate that measurements provide vectors $x \in \mathbb{R}^n$ together with covariance matrices Σ that quantify our uncertainty in the value x we recorded. For the purposes of binary classification, we record such pairs (x, Σ) along with the information on the category of origin of each, which is given by an indicator value $y = \pm 1$. A datum with a large variance naturally delineates a larger volume that should be associated to the corresponding category, or at least, impact proportionately the drawing of the separation between the two. Prior state-of-the-art does not do that. Thus, in the present note, we propose a new paradigm of treating measurement uncertainty as part of the data.

Below, in Section II, we highlight some background on Support Vector Machines. Section III gives our main result that identifies a kernel function for the case of data provided in the form of “Gaussian points.”

2. BACKGROUND

Support vector machines are well established techniques for (binary) classification and regression analysis [1]. The main idea is to imbed a given (training) data set into a high dimensional space \mathcal{H} , of dimension much higher than the native dimension of the base space \mathcal{X} and possibly infinite, so that for binary classification, the two classes can be separated with a hyperplane [2]. Effectively, this *separation hyperplane* projects down to the native base space \mathcal{X} , where the data points originally reside, as curves/surfaces/etc. that separate the two classes. The imbedding into \mathcal{H} is effected by a mapping

$$\varphi : x \in \mathcal{X} \mapsto \varphi(x) \in \mathcal{H},$$

where $\varphi(x)$ is referred to as the *feature vector*. The space \mathcal{H} has an inner product structure (Hilbert space) and, naturally, the construction of the separating hyperplane relies on the geometry of \mathcal{H} . However, and most importantly, the map φ does not need to be known explicitly and does not need to be applied to the data. All needed operations, inner product and projections that show up in the classifier and computations, can be effectively carried out in the base space using the so called “kernel trick.”

Indeed, to accomplish the task and construct the classification rule (via suitable curves/surfaces/etc.) it is sufficient to know the *kernel*

$$(1) \quad k(x, y) := \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}},$$

with $x, y \in \mathcal{X}$. It evaluates the inner product in the feature space as a function of the corresponding representatives x, y in the native base space. Thus, the kernel is a bivariate function which is completely characterized by the property of being positive, in the sense that for all $x_i \in \mathcal{X}$, $i \in \{1, \dots, N\}$, and any corresponding set of values $\alpha_i \in \mathbb{R}$ (as we are interested in real-valued kernels)

$$\sum_{i, j=1}^N \alpha_i \alpha_j k(x_i, x_j) \geq 0.$$

Necessity stems from the fact that the left hand side is the inner product

$$\left\langle \sum_{i=1}^N \alpha_i \varphi(x_i), \sum_{j=1}^N \alpha_j \varphi(x_j) \right\rangle_{\mathcal{H}}.$$

Sufficiency, in the existence of a feature map φ that realizes (1) is a celebrated theorem [3], see also [4, Theorem 7.5.2] and [1, page 30, Theorem 3.1 (Mercer)].

Classification relies on constructing a classifier that is built on a linear functional, when viewed in \mathcal{H} . It is of the form

$$(2) \quad x \rightarrow \text{sign}(\langle w, \varphi(x) \rangle_{\mathcal{H}} - b),$$

and the value ± 1 aims to differentiate between elements in two categories. The coefficients $w \in \mathcal{H}$ and $b \in \mathbb{R}$ are chosen so that

$$\{h \in \mathcal{H} \mid \langle w, h \rangle_{\mathcal{H}} - b = 0\}$$

is a separating hyperplane of the two subsets

$$\mathcal{S}_{\pm} = \{\varphi(x_i) \mid y_i = \pm 1\}$$

of the complete (training) data set. Once again, $\varphi(\cdot)$ does not need to be evaluated at any point in the construction, existence of such a map is enough, and it is guaranteed by the positivity of the kernel.

The construction of the classifier requires selection of the parameters $w = \sum_{i=1}^N c_i y_i \varphi(x_i) \in \mathcal{H}$ and $b \in \mathbb{R}$. These are chosen either (“hard margin”) to minimize

$$\langle w, w \rangle_{\mathcal{H}} \text{ subject to } y_i (\langle w, \varphi(x_i) \rangle_{\mathcal{H}} - b) \geq 1,$$

or, to (“soft margin”) minimize

$$(3) \quad \frac{1}{N} \sum_{i=1}^N \max\{0, 1 - y_i (\langle w, \varphi(x_i) \rangle_{\mathcal{H}} - b)\} + \lambda \langle w, w \rangle_{\mathcal{H}},$$

for all available points in the “training set.” The “hard margin” formulation coincides with the limit of the “soft” formulation as $\lambda \rightarrow 0$. The dual formulation of (3) becomes to maximize

$$\sum_{i=1}^N c_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^n y_i y_j c_i c_j k(x_i \cdot x_j),$$

subject to

$$\sum_{i=1}^N c_i y_i = 0, \text{ as well as } 0 \leq c_i \leq (2N\lambda)^{-1}$$

for all i . The coefficients c_i can now be obtained via quadratic programming, and b can be found as

$$b = \sum_{i=1}^N c_i y_i k(x_i, x_j) - y_j$$

with j corresponding to an index such that $0 < c_j < (2N\lambda)^{-1}$. The classification rule then becomes

$$(4) \quad x \rightarrow \text{sign}\left(\sum_{i=1}^N c_i y_i k(x_i, x) - b\right),$$

The above follows standard development, see e.g., [5] as well as the Wikipedia webpage on Support Vector Machines.

3. CLASSIFICATION OF GAUSSIAN POINTS

The problem we are addressing in the present note is the classification of *uncertain data points* into one of two categories, i.e., a binary classification as before. However, a salient feature of our setting is that data are only known with finite accuracy. Uncertainty is modeled in a probabilistic manner. For simplicity, in this paper, we consider only *Gaussian points*. These consist of pairs (x_i, Σ_i) representing the mean and variance of a normally distributed random vector ξ_i .

Alternatively, we may think of the data as points on a manifold of distributions (though, only Gaussian at present). These may represent approximations of empirical distributions that have been obtained at various times. An indicator y_i is provided as usual along with the information of the category that the current datum belongs to. If we regard it as representing a distribution, we postulate that it arose from experiment involving population y_i .

We follow the standard setting of *kernel Support Vector Machines* (kSVM) that was outlined in the background section, and we overlay a probabilistic component. This *Probabilistic kernel Support Vector Machine* (PkSVM) relies on a suitable modification of the kernel. To this end we consider the set of “data points”

$$\Omega := \{(x, \Sigma) \mid x \in \mathbb{R}^n, \Sigma \in S_{+,n}\}$$

with $S_{+,n}$ the cone of non-negative definite symmetric matrices in $\mathbb{R}^{n \times n}$. We also consider the family of normally distributed independent random vectors

$$\xi \sim N(x, \Sigma) \text{ for } (x, \Sigma) \in \Omega,$$

i.e., (ξ_i, ξ_j) are independent when $i \neq j$, and $(x_i, \Sigma_i) \in \Omega$. We utilize the popular exponential (RBF) kernel

$$(5) \quad k(x, y) = e^{-\frac{1}{2\sigma^2} \|x-y\|^2},$$

where the parameter $\sigma > 0$ affects the scale of the desired resolution. We readily compute

$$\begin{aligned} \mathbb{E}\{k(\xi_i, \xi_j)\} &= \mathbb{E}\{e^{-\frac{1}{2\sigma^2} \|\xi_i - \xi_j\|^2}\} \\ &= (2\pi)^{-n/2} |\Sigma_i + \Sigma_j|^{-1/2} \times \int e^{-\frac{1}{2} \|\xi_i - \xi_j - x_i + x_j\|_{(\Sigma_i + \Sigma_j)}^2 - \frac{1}{2\sigma^2} \|\xi_i - \xi_j\|^2} d\xi_i d\xi_j \\ (6) \quad &= |I + \sigma^{-2}(\Sigma_i + \Sigma_j)|^{1/2} \times e^{-\frac{1}{2} \|x_i - x_j\|_{(\sigma^2 I + (\Sigma_i + \Sigma_j))}^2} \end{aligned}$$

for $i \neq j$. When $i = j$, i.e., for any $\xi_i \in \Xi$,

$$(7) \quad \begin{aligned} \mathbb{E}\{k(\xi_i, \xi_i)\} &= 1 \\ &\leq |I + \sigma^{-2}(2\Sigma_i)|^{1/2}, \end{aligned}$$

where the inequality in (7) we get from (6) when substituting the same values for x_i, x_j and Σ_i, Σ_j , respectively. We now state our main result.

Theorem 1. *The function*

$$(8) \quad \begin{aligned} \kappa((x_i, \Sigma_i), (x_j, \Sigma_j)) &:= \\ &|I + \sigma^{-2}(\Sigma_i + \Sigma_j)|^{1/2} e^{-\frac{1}{2} \|x_i - x_j\|_{(\sigma^2 I + (\Sigma_i + \Sigma_j))}^2} \end{aligned}$$

defines a positive kernel on Ω .

Proof. Consider $k(x, y)$ in (5) for $x, y \in \mathbb{R}^n$. Then, for any collection $\alpha_i \in \mathbb{R}$ and any collection of $(x_i, \Sigma_i) \in \Omega$, for $i \in \{1, \dots, N\}$,

$$\begin{aligned} &\sum_{i,j=1}^N \alpha_i \alpha_j \kappa((x_i, \Sigma_i), (x_j, \Sigma_j)) \\ &\geq \mathbb{E}\left\{ \sum_{i,j=1}^N \alpha_i \alpha_j k(\xi_i, \xi_j) \right\} \\ &= \mathbb{E}\left\{ \left\langle \sum_{i=1}^N \alpha_i \varphi(\xi_i), \sum_{j=1}^N \alpha_j \varphi(\xi_j) \right\rangle_{\mathcal{H}} \right\} \geq 0, \end{aligned}$$

with $\varphi(\cdot)$ the map to the radial basis functions corresponding to the positive kernel (5). The inequality in line two, follows from (7). The claim in the theorem follows. \square

It can be seen that when applied to points in Ω having zero uncertainty, i.e., when the corresponding covariance matrices are identically 0, then the Probabilistic kernel Support Vector Machine model reduces to the standard one where points lie in \mathbb{R}^n . That is,

$$\kappa((x_i, 0), (x_j, 0)) = k(x_i, x_j).$$

Thereby, PkSVM is natural extension of standard SVM, able to account for error and uncertainty that is available and encoded in the data.

4. CONCLUSION

The formalism herein presents a new paradigm where data incorporate a quantification of own uncertainty. We concentrated on binary classification and the case of “Gaussian points” to present a proof of concept, in the form of a suitable kernel in Theorem 1. Numerical experiments and further insight will be provided and included in a forthcoming more detailed report. Moreover, the basic idea appears to be easily generalizable to more detailed and explicit descriptions of uncertainty, e.g., Gaussian Mixture models, with however, the caveat of added complexity in the resulting formulae. This we expect will be the starting point of future investigations.

ACKNOWLEDGEMENTS

This project was supported in part by NSF under grants 1665031, 1807664, 1839441 and 1901599, the AFOSR under Grants FA9550-17-1-0435, grants from National Institutes of Health (R01-AG048769, R01-CA198121), MSK Cancer Center Support Grant/Core Grant (P30 CA008748), and a grant from Breast Cancer Research Foundation (grant BCRF-17-193).

REFERENCES

- [1] B. Schölkopf, C. J. Burges, A. J. Smola *et al.*, *Advances in kernel methods: support vector learning*. MIT press, 1999.
- [2] T. M. Cover, “Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition,” *IEEE transactions on electronic computers*, no. 3, pp. 326–334, 1965.
- [3] N. Aronszajn, “Theory of reproducing kernels,” *Transactions of the American mathematical society*, vol. 68, no. 3, pp. 337–404, 1950.
- [4] D. Alpay, “An advanced complex analysis problem book,” *Topological vector spaces, functional analysis, and Hilbert spaces of analytic functions*. Birkäuser Basel, 2015.
- [5] B. Scholkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.