A Unifying View of Support Measure Machines, Support Vector Machines, and Parzen Window Classifiers

Krikamol Muandet
Empirical Inference Department
Max Planck Institute for Intelligent Systems

Bernhard Schölkopf
Empirical Inference Department
Max Planck Institute for Intelligent Systems

Abstract: This paper presents a unifying view of two well-known kernel-based classifiers, namely support vector machines (SVMs) and Parzen window classifiers. In particular, given the training data, both learning algorithms can be viewed as a solution to a regularization problem on probability distributions, depending on how the distributions are constructed from the training data. This simple insight may shed light on the unification of various kernel-based learning algorithms.

Keywords: support measure machines (SMMs), support vector machines (SVMs), parzen window classifiers (PWCs), kernel methods, regularization, reproducing kernel Hilbert space (RKHS).

1 Introduction

The learning framework for probability measures was recently introduced in [1]. Based on a simple distributional view of the traditional kernel-based learning framework, the authors generalize kernel-based discriminative learning to general probability measures. This generalization not only allows for useful and straightforward theoretical analyses, but also shed light on a wide range of application domains such as multiple-instance learning, learning on data with substantial and heterogenous uncertainty, missing data problems, domain adaptation/generalization, group anomaly detection, large-scale machine learning, etc.

Although having been introduced independently, the proposed framework has an intrinsic connection to two well-known existing learning algorithms, namely a support vector machine (SVM) and Parzen window classifier (PWC). We show in this extended abstract precisely this connection and discuss some important questions regarding this connection.

2 Regularization on Distributions

A regularization problem on probability distributions can be formulated as follow. Given training examples $(\mathbb{P}_i, y_i) \in \mathscr{P} \times \mathbb{R}$, $i = 1, \ldots, m$, a strictly monotonically increasing function $\Omega : [0, +\infty) \to \mathbb{R}$, and a loss function $\ell : (\mathscr{P} \times \mathbb{R}^2)^m \to \mathbb{R} \cup \{+\infty\}$, we find $f \in \mathcal{H}$ such that the regularization functional

$$\ell\left(\mathbb{P}_1, y_1, \mathbb{E}_{\mathbb{P}_1}[f], \ldots, \mathbb{P}_m, y_m, \mathbb{E}_{\mathbb{P}_m}[f]\right) + \Omega\left(\|f\|_{\mathcal{H}}\right) \quad (1)$$

is minimized where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) with a reproducing kernel k. By representer theorem on distributions [1], any solution f admits a representation of the form

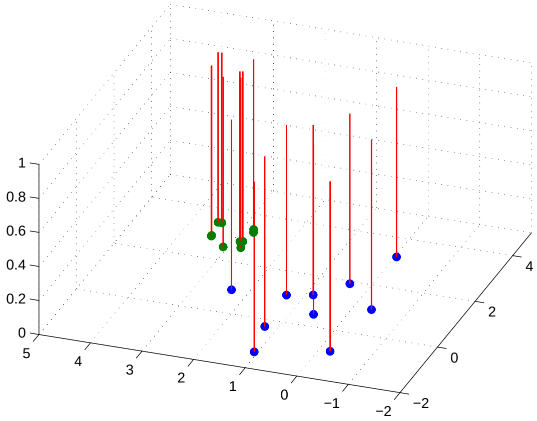

Fig. 1: The classification boundary of the SVM and the Parzen window classifier on toy dataset.

$$f = \sum_{i=1}^{m} \alpha_i \mathbb{E}_{x\sim\mathbb{P}_i}[k(x,\cdot)] \triangleq \sum_{i=1}^{m} \alpha_i \mu[\mathbb{P}_i] \quad (2)$$

for some $\boldsymbol{\alpha} \in \mathbb{R}^m$. The term $\mu[\mathbb{P}_i] = \mathbb{E}_{x\sim\mathbb{P}_i}[k(x,\cdot)]$ is known as a kernel mean embedding of the distribution $\mathbb{P}_i$ and has been intensively studied. In other words, any solution f can be written as a linear combination of the kernel mean embeddings of the training distributions.

In this paper we focus on a hinge loss $\ell_H(\hat{y}) = \max(0, 1 - \hat{y} \cdot y)$ where $\hat{y} = f(x)$. In which case we call an algorithm that minimizes (1) a *linear* support measure machine (SMM)[1].

[1] See [1] for a nonlinear extension of SMM.

3 Connecting the Dots

Given training data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \in \mathcal{X} \times \{+1, -1\}$, we will show that both SVM and PWC can be recovered as a solution to the regularization functional (1).

3.1 Support Vector Machines

The connection between SVM and SMM is quite straightforward. In this case, we replace each training sample x_i by a Dirac measure δ_{x_i} centered at that sample. Note that this reparameterization does not alter the problem as the map $x_i \mapsto k(x_i, \cdot)$ and $\delta_{x_i} \mapsto \int k(x, \cdot)\delta_{x_i}(x)$ are equivalent. Replacing $\mathbb{P}_i$ in (1) by δ_{x_i} yields

$$\ell_H\left(x_1, y_1, f(x_1), \ldots, x_m, y_m, f(x_m)\right) + \Omega\left(\|f\|_{\mathcal{H}}\right)$$

which corresponds to the SVM [2]. In summary, the SVM treats every samples as a probability distribution that captures the *local* information of the sample.

3.2 Parzen Window Classifiers

As opposed to the SVM, the PWC approaches the problem from a completely opposite direction. In general, it begins by estimating the class conditional distribution

$$\widehat{P}(x|y) = \frac{1}{|\{i|y_i = y\}|} \sum_{i, y_i = y} k(x, x_i)$$

and by Bayes rule the posterior can be computed by

$$\widehat{P}(y|x) = \frac{\sum_{i, y_i = y} k(x, x_i)}{\sum_{i=1}^{n} k(x, x_i)}. \tag{3}$$

Consequently, for binary classification problem, the estimated class conditional $\widehat{P}(\cdot|y = +1)$ and $\widehat{P}(\cdot|y = -1)$ can be seen as class-specific mean functions in $\mathcal{H}$, i.e.,

$$\begin{aligned} M^+ &= \frac{1}{|\{i|y_i = +1\}|} \sum_{y_i = +1} k(x_i, \cdot) \\ M^- &= \frac{1}{|\{i|y_i = -1\}|} \sum_{y_i = -1} k(x_i, \cdot) \end{aligned}$$

where $M^+ = \widehat{P}(\cdot|y = +1)$ and $M^- = \widehat{P}(\cdot|y = -1)$. Hence, the classification function based on the posterior (3) can be equivalently written as

$$f(x) = \operatorname{sign}(\langle x, M^+\rangle_{\mathcal{H}} - \langle x, M^-\rangle_{\mathcal{H}}) \tag{4}$$

It is not difficult to see that for a positive semi-definite kernel k the mean functions M^+ and M^- are two distinct functions in the RKHS $\mathcal{H}$. As a result, the classification function (4) can be obtained by solving a classification problem on training examples $(M^+, +1)$ and $(M^-, -1)$, which is exactly equivalent to minimizing

$$\ell_H(\widehat{\mathbb{P}}_+, +1, \mathbb{E}_{\widehat{\mathbb{P}}_+}[f], \widehat{\mathbb{P}}_-, -1, \mathbb{E}_{\widehat{\mathbb{P}}_-}[f]) + \Omega\left(\|f\|_{\mathcal{H}}\right)$$

As opposed to the SVM, the PWC treats class conditional distributions as training samples, emphasizing more on the *global* information of the training data.

4 Discussions

Intuitively, we may consider both SVM and PWC as the extreme ends of the spectrum of learning algorithms: one that look locally at the training data and another one that consider its global properties. Consequently, It is natural to ask what constitute various learning algorithms along this spectrum?

From the distributional point of view, these two learning algorithms employ different learning strategy. Roughly speaking, the SVM performs a learning at the most fine-grained level for the best accuracy at the expense of training time. On the other hand, the Parzen window classifier trade-off the accuracy with the training time (no training time) in order to obtain the solution very quickly. These two learning approaches are also different in term of estimation time. A good learning strategy therefore should trade-off these quantities.

To conclude, we presents a unifying view of two well-known kernel-based classifiers, namely support vector machines (SVMs) and Parzen window classifiers (PWCs). In particular, both of them can be recovered as a solution to a regularization problem on probability distributions. This simple insight may shed light on the unification of various kernel-based learning algorithms.

Acknowledgments

Krikamol Muandet would like to thank Samory Kpotufe for a fruitful discussion on the insight into the trade-off of the estimation time.

References

[1] Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. Learning from distributions via support measure machines. In *NIPS*, pages 10–18. MIT Press, 2012.

[2] Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, MA, USA, 2001.