Statistical Science
2006, Vol. 21, No. 3, 337–340
DOI: 10.1214/088342306000000484
Main article DOI: 10.1214/088342306000000493

Comment

Olivier Bousquet and Bernhard Schölkopf

Our contribution will be short, but we will try to compensate by being particularly opinionated. The field of support vector machines (SVMs) and related kernel methods has produced an impressive range of theoretical results, algorithms and success stories in real-world applications. While it originated in machine learning, it is also concerned with core problems of statistics and it is thus timely to publish a comprehensive article that discusses these methods from a statistician's point of view. We shall use this opportunity to make a few general comments, largely about the field rather than about the present paper.

Many papers about SVMs start off saying something like "SVMs are great because they are based on statistical learning theory" (this probably includes some of our own writings). Moguerza and Muñoz are more careful and only say that SVMs appeared in the context of statistical learning theory. What actually is the connection between SVMs and statistical learning theory?

Historically, SVMs and their precursors were (co-) developed by Vladimir Vapnik, one of the fathers of statistical learning theory. Statistical learning theory includes an analysis of machine learning which is independent of the distribution underlying the data. However, this analysis cannot provide any a priori guarantee that SVMs (or any other algorithm) will work well on a real-world problem. So what is special about SVMs, if anything?

In our view, what is special about SVMs is the combination of the following ingredients: first and foremost, the use of positive definite kernels; then regularization via the norm in the associated reproducing kernel Hilbert space; finally, the use of a convex loss function which is minimized by a classifier and not a regressor.

The magic of kernels. Positive definite kernels and their feature space interpretation do provide a very nice way to look at a whole class of algorithms; however, it is important to stress that they do not bring any *statistical* guarantee by themselves. The statistical guarantees available stem from the regularization (or learning theory) point of view. We shall return to this point below.

The main advantages of positive definite kernels are the following:

1. They allow easy construction of a nonlinear algorithm from a linear one, often without incurring additional computational cost.
2. They provide generality via the fact that they can be defined on nonvectorial data and do not, in general, require an explicit mapping to a reproducing kernel Hilbert space.

Historically, the first point was initially considered one of the major advantages of kernels and it triggered a significant number of kernel algorithms other than SVMs, starting with kernel principal component analysis (PCA). More recently, the second point has arguably taken over the role of the key selling point for kernel methods. The application of learning algorithms to nonvectorial data has become the field where nowadays a lot of the action is happening in the machine learning world, in particular concerning applications on structured data (e.g., in biology or natural language processing). We are curious to see whether the field of statistics will also embrace these possibilities.

A sober look at the geometric interpretation. The geometric point of view is an original way to look at SVMs and quite possibly the right way to come up with an algorithm like the SVM in the first place. However, it does not yield comprehensive statistical understanding. More precisely, there is no way to prove that large margin separating hyperplanes perform better than other types of hyperplanes independently of the distribution of the data.

Sure enough, the geometric point of view does provide intuition and motivates a large number of related algorithms, but one should not be fooled by geometric intution or two-dimensional illustrations. The fact that data that are not linearly separable in input space suddenly becomes linearly separable in the so-called feature space (as depicted on Figure 1 of the main paper)

Olivier Bousquet is Director of Research, Pertinence, F-75002 Paris, France (e-mail: o.bousquet@pertinence.com). Bernhard Schölkopf is Professor and Director, Max Planck Institute for Biological Cybernetics, D-72076 Tübingen, Germany (e-mail: bs@tuebingen.mpg.de).