The equivalence between the one-class and paired support vector machines for nonseparable data

Clayton Scott

September 18, 2012

In their original paper on the one-class support vector machine (SVM), Schölkopf et al. (2001) establish that for separable data, the one-class SVM applied to patterns $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is equivalent to the corresponding "paired SVM", that is, the standard (two-class) SVM applied to the paired data $(\mathbf{x}_1, 1), \ldots, (\mathbf{x}_n, 1), (-\mathbf{x}_1, -1), \ldots, (-\mathbf{x}_n, -1)$. In this context, $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are said to be separable if the paired data are linearly separable. The authors also state, without proof, that the equivalence holds in the nonseparable case provided some hard-to-classify data points are removed. This note establishes a general equivalence for nonseparable data that does not require modification of the data.

1 The One-Class SVM

The one-class SVM, as introduced by Schölkopf et al. (2001), takes as input unlabeled data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and a parameter $0 \leq \nu \leq 1$, and returns parameters $(\mathbf{w}, \rho)$ solving

$$\begin{aligned} \min_{\mathbf{w}, \boldsymbol{\xi}, \rho > 0} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n} \xi_i - \rho \qquad (1) \\ \text{s.t.} \quad & \langle \mathbf{w}, \mathbf{x}_i \rangle \geq \rho - \xi_i, \quad \xi_i \geq 0 \quad \text{for } i = 1, 2, \ldots, n \end{aligned}$$

The resulting classifier is given by $\mathbf{x} \mapsto \operatorname{sgn}\{\langle \mathbf{w}, \mathbf{x} \rangle - \rho\}$. Here $\langle \cdot, \cdot \rangle$ denotes the standard dot product. For the purpose of comparison with the two-class and paired SVMs, it is convenient to express the one-class SVM as the solution of an alternative quadratic program, namely,

$$\begin{aligned} \min_{\mathbf{w}, \boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \qquad (2) \\ \text{s.t.} \quad & \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \text{for } i = 1, 2, \ldots, n \end{aligned}$$

The corresponding classifier is $\mathbf{x} \mapsto \operatorname{sgn}\{\langle \mathbf{w}, \mathbf{x} \rangle - 1/C\}$. The equivalence between (1) and (2) is given by the following result, which was established by Lee and Scott (2007).

Proposition 1. *If* (1) *results in* $\rho > 0$, *then* (2) *with* $C = \frac{1}{\nu n \rho}$ *leads to the same classifier.*

2 The Paired SVM

The paired SVM is a special case of the standard (two-class) SVM. The standard SVM takes as input a parameter $C > 0$ and labeled training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, where $\mathbf{x}_i$ are feature vectors and $y_i = \pm 1$ are labels, and returns $(\mathbf{w}, b)$ solving

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \tag{3}$$
$$\text{s.t.} \quad y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \text{for } i = 1, 2, \ldots, n$$

These parameters define the linear classifier $\mathbf{x} \mapsto \operatorname{sgn}\{\langle \mathbf{w}, \mathbf{x}\rangle + b\}$.

In the paired SVM, there are unlabeled feature vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$. These are used to form labeled data $(\mathbf{x}_1, 1), \ldots, (\mathbf{x}_n, 1), (-\mathbf{x}_1, -1), \ldots, (-\mathbf{x}_n, -1)$, which are then given to the standard SVM as input. By substitution, the paired SVM hyperplane solves

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_{i+n}) \tag{4}$$
$$\text{s.t.} \quad \langle \mathbf{w}, \mathbf{x}_i\rangle + b \geq 1 - \xi_i, \quad \text{for } i = 1, 2, \ldots, n \tag{5}$$
$$\langle \mathbf{w}, \mathbf{x}_i\rangle - b \geq 1 - \xi_{i+n}, \quad \text{for } i = 1, 2, \ldots, n \tag{6}$$
$$\xi_i \geq 0, \quad \xi_{i+n} \geq 0, \quad \text{for } i = 1, 2, \ldots, n$$

Because of symmetry, the above quadratic program can be simplified somewhat. There is always a solution of (4) with $b = 0$.

Proposition 2. *If* $(\mathbf{w}, b)$ *is a solution of* (4)*, then so is* $(\mathbf{w}, 0)$.

Proof. Suppose $(\mathbf{w}, b, \xi)$ is optimal and $b \neq 0$. Without loss of generality, assume $b > 0$. Since $b > 0$, it must be true that $\xi_i \leq \xi_{i+n}$ for all i. Consider the following cases at the given optimum for each i: (I) (5) and (6) are both strict inequalities; (II) (5) and (6) are both equalities; (III) (5) is a strict inequality and (6) is an equality; (IV) (6) is a strict inequality and (5) is an equality.

For i satisfying (I), by the KKT conditions, $\xi_i = \xi_{i+n} = 0$ and therefore $\langle \mathbf{w}, \mathbf{x}_i\rangle > 1 + b > 1$. Hence, for all $b' \in [0, b)$, $(\mathbf{w}, b', \xi)$ still satisfies the constraints for $\mathbf{x}_i$. The conclusion follows by taking $b' = 0$. For i satisfying (II), we have $\xi_{i+n} = \xi_i + 2b$, from which we deduce $\xi_{i+n} > \xi_i$ and $\xi_{i+n} \geq 2b$. Replacing b, ξ_i, and ξ_{i+n} by b', $\xi_i + b - b'$, and $\xi_{i+n} - b + b'$, for any $b' \in [0, b)$, the constraints on $\mathbf{x}_i$ are still satisfied, and the corresponding term in the objective function remains unchanged. The conclusion follows by taking $b' = 0$.

For case (III), we consider two sub-cases: (IIIa) $\xi_{i+n} = 0$, (IIIb) $\xi_{i+n} > 0$. For i satisfying case (IIIa), $\langle \mathbf{w}, \mathbf{x}_i\rangle = 1 + b > 1$, and therefore (5) and (6) remain valid if we replace b by any $b' \in [0, b)$. The conclusion follows by taking $b' = 0$. Case (IIIb) cannot occur. To see this, suppose it does occur for some i. Assume for the moment that (IIIb) occurs for only one i. By the KKT conditions, $\xi_i = 0$. Also, subtracting (6) from (5) we obtain $\xi_{i+n} < 2b$. We can obtain a feasible point with a smaller objective function value by replacing b with any $b' \in (\max\{0, 1 - \langle \mathbf{w}, \mathbf{x}_i\rangle, b - \xi_{i+n}\}, b)$ and ξ_{i+n} with $\xi'_{i+n} = \xi_{i+n} - b + b'$. By the previous cases, changing b in this manner does not affect the validity of the other constraints. If (IIIb) holds for more that one i, the above argument still applies, where now the lower bound on b' is maximized over these indicies.

Case (IV) cannot occur. To see this, suppose it does occur. The KKT conditions applied to (6) imply $\xi_{i+n} = 0$ and $\langle \mathbf{w}, \mathbf{x}_i \rangle > 1 + b$. Then (5) implies $1 + 2b < \langle \mathbf{w}, \mathbf{x}_i \rangle + b = 1 - \xi_i \leq 1$, contradicting $b > 0$.

□

By this result, it suffices to consider the following quadratic program:

$$\begin{aligned} \min_{\mathbf{w}, \boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_{i+n}) \\ \text{s.t.} \quad & \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \text{for } i = 1, 2, \ldots, n \\ & \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1 - \xi_{i+n}, \quad \xi_{i+n} \geq 0 \quad \text{for } i = 1, 2, \ldots, n \end{aligned} \tag{7}$$

This amounts to the so-called SVM *without offset*, applied to the paired data.

3 The Connection

The equivalence between the one-class SVM and the paired SVM is now evident.

Proposition 3. $\mathbf{w}$ *is optimal for* (2) *with parameter* C *if and only if* $\mathbf{w}$ *is optimal for* (7) *with parameter* $C/2$*.*

Proof. The proof follows easily from the observation that, at the optimum of (7), the slack variables ξ_i and ξ_{i+n} must be equal. □

References

G. Lee and C. Scott. The one class support vector machine solution path. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing — ICASSP 2007*, volume 2, pages II–521–II–524, Honolulu, USA, 2007.

B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1472, 2001.