NESTED SUPPORT VECTOR MACHINES

Gyemin Lee and Clayton Scott

Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, Michigan, USA
E-mail: {gyemin, cscott}@eecs.umich.edu

ABSTRACT

The one-class and cost-sensitive support vector machines (SVMs) are state-of-the-art machine learning methods for estimating density level sets and solving weighted classification problems, respectively. However, the solutions of these SVMs do not necessarily produce set estimates that are nested as the parameters controlling the density level or cost-asymmetry are continuously varied. Such a nesting constraint is desirable for applications requiring the simultaneous estimation of multiple sets, including clustering, anomaly detection, and ranking problems. We propose new quadratic programs whose solutions give rise to nested extensions of the one-class and cost-sensitive SVMs. Furthermore, like conventional SVMs, the solution paths in our construction are piecewise linear in the control parameters. We also describe a decomposition algorithm to solve the quadratic programs. The results of these methods are demonstrated on synthetic data sets.

Index Terms— pattern classification, one class support vector machine, cost sensitive support vector machine, nested set estimation, solution paths

1. INTRODUCTION

Many statistical learning problems may be characterized as problems of *set estimation*. In these problems, the input takes the form of a random sample of points in a feature space, while the desired output is a subset G of the feature space. For example, in density level set estimation, a random sample from a density is given and G is an estimate of a density level set. In binary classification, labeled training data is available, and G is the set of all feature vectors predicted to belong to one of the classes.

In other statistical learning problems, the desired output is a *family* of sets G_θ with the index θ taking values in a continuum. For example, estimating density level sets at multiple levels is an important task for many problems including clustering [1], outlier ranking, minimum volume set estimation [2], and anomaly detection [3]. Estimating cost-sensitive classifiers at a range of different cost asymmetries is important for ranking, Neyman-Pearson classification [4], transductive anomaly detection [5], and ROC studies.

Support vector machines (SVMs) are powerful nonparametric approaches to set estimation [6]. However, both the one-class SVM for level set estimation and the standard two-class SVM for classification do not produce set estimates that are *nested* as the parameter controlling the density level or, respectively, misclassification cost is varied. Since the true sets being estimated are in fact nested in these two applications, estimators that enforce the nesting constraint will not only avoid nonsensical solutions, but should also be more accurate and less variable.

In this paper, we develop nested variants of the one-class and two-class SVMs by incorporating nesting constraints into the dual quadratic programs defining these methods. Decomposition algorithms for solving the modified duals are also presented. Like the solution paths for the conventional unnested SVMs [7, 8, 9], the nested SVM solution paths are also piecewise linear in the control parameters. We compare our nested paths to the unnested paths on synthetic data sets.

2. SUPPORT VECTOR MACHINES

Suppose that we have a random sample $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ where $\mathbf{x}_i \in \mathbb{R}^d$ is a d-dimensional feature vector and $y_i \in \{-1, +1\}$ is its class; in one-class classification, all the y_i's are the same. The support vector machine (SVM) finds a hyperplane that separates data points in a high dimensional space $\mathcal{H}$ based on the maximum margin principle. $\mathcal{H}$ is the reproducing kernel Hilbert space generated by a positive semidefinite kernel k. The kernel function k corresponds to an inner product in $\mathcal{H}$ through $k(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle$ where $\Phi : \mathbb{R}^d \to \mathcal{H}$ is a function that maps each data point $\mathbf{x}_i$ into $\mathcal{H}$. More detailed discussion regarding SVMs can be found in [6].

2.1. One-Class SVMs

The OC-SVM was proposed in [10, 11] to estimate a level set of an underlying probability density given a data sample from the density. In particular, the OC-SVM introduced in [10] solves the following optimization problem:

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

The optimal $\mathbf{w} \in \mathcal{H}$ is the normal vector of the hyperplane and $\frac{\rho}{\|\mathbf{w}\|}$ is the margin between the origin and the hyperplane.

In practice the optimization problem is solved via its dual, which depends only on a set of Lagrange multipliers (one for each $\mathbf{x}_i$). These Lagrange multipliers define a decision function that determines whether a point is an outlier or not. Generally, only a fraction of the Lagrange multipliers take non-zero values; associated $\mathbf{x}_i$ are called support vectors. The control parameter $\nu \in [0, 1]$ can be shown to be an upper bound on the fraction of outliers and a lower bound on the fraction of support vectors [10]. Thus ν implicitly defines the corresponding density level .

Our nested SVM uses an alternative formulation of the OC-SVM involving a different parameter λ, which traces out the same class of

decision functions, and asymptotically has a one-to-one correspondence to the original parameter ν [12]. The dual problem is

$$\min_{\boldsymbol{\alpha}} \frac{1}{2\lambda} \sum_{i,j} \alpha_i \alpha_j k_{ij} - \sum_i \alpha_i \tag{2}$$
$$\text{s.t.} \quad 0 \le \alpha_i \le \frac{1}{N} \quad \text{for } \forall i$$

where $k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_N)$. Let $\boldsymbol{\alpha}^*(\lambda) = (\alpha_1^*(\lambda), \alpha_2^*(\lambda), \cdots, \alpha_N^*(\lambda))$ denote the optimal solution when the regularization parameter is λ. An advantage of the λ parameterization is that the $\alpha_i^*(\lambda)$ are piecewise linear in λ [9]. From this result, level sets at a range of density levels can be estimated with a computational cost comparable to solving (2) for a single λ using the algorithm in [7].

The set estimate conventionally associated with the OC-SVM is given by the following formula with $\eta = 1$:

$$\widehat{G}_\lambda = \{\mathbf{x} : \sum_i \alpha_i^*(\eta\lambda) k(\mathbf{x}_i, \mathbf{x}) > \lambda\}. \tag{3}$$

However, Vert and Vert [13] show that for any $\eta > 1$, the estimate in (3) is a consistent estimate of the true level set at level λ as $N \to \infty$, suggesting that the $\eta = 1$ estimate may be inconsistent.

Thus η is a design parameter of the OC-SVM, and we now argue that it relates to the nesting properties of the solution path. Note that if all the Lagrange multipliers satisfy the condition $\alpha_i^*(\eta\lambda) = \frac{1}{N}$, then (3) becomes equivalent to set estimation based on kernel density estimation (KDE). Indeed from the optimality condition of (2), we can compute λ_0 such that $\lambda > \frac{\lambda_0}{\eta}$ implies $\alpha_i^*(\eta\lambda) = \frac{1}{N}$. Thus for larger value of η, more of the sets $\widehat{G}_\lambda$ correspond to thresholded KDEs. Since a KDE is a proper density function, the estimated level sets are naturally nested. However, we lose some sparsity and are no longer implementing the large margin principle. On the other hand, when η is close to 1, the obtained estimates are not guaranteed to be nested as can be seen in the experimental results in Section 4. Therefore, the choice of η affects the trade-off between the nestedness and the reliance on the large margin principle.

2.2. Cost-Sensitive SVMs

The original SVM penalizes errors in both classes equally. However, there are many applications where the numbers of data samples from each class are not balanced, or false positives and false negatives incur different costs. To handle this issue, the cost-sensitive SVM (CS-SVM) was introduced [14].

Here we consider the CS-SVM without bias. Let $I_+ = \{i : y_i = +1\}$ and $I_- = \{i : y_i = -1\}$. Then the dual optimization problem of the CS-SVM is

$$\min_{\boldsymbol{\alpha}} \frac{1}{2\lambda} \sum_{i,j} \alpha_i \alpha_j y_i y_j k_{ij} - \sum_i \alpha_i \tag{4}$$
$$\text{s.t.} \quad 0 \le \alpha_i \le \gamma \quad \text{for } i \in I_+$$
$$0 \le \alpha_i \le 1 - \gamma \quad \text{for } i \in I_-.$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_N)$. $\gamma \in [0, 1]$ controls the cost asymmetry between false positives and false negatives.

Once an optimal solution $\boldsymbol{\alpha}^*(\gamma) = (\alpha_1^*(\gamma), \alpha_2^*(\gamma), \cdots, \alpha_N^*(\gamma))$ is found, the sign of the decision function

$$f_\gamma(\mathbf{x}) = \frac{1}{\lambda} \sum_i \alpha_i^*(\gamma) y_i k(\mathbf{x}, \mathbf{x}_i) \tag{5}$$

determines the class of a data point $\mathbf{x}$.

Bach et al. [8] extended the method of Hastie et al. [7] to the CS-SVMs in order to derive the optimal solution path. The $\alpha_i^*(\gamma)$ are again piecewise linear in γ.

3. NESTED SUPPORT VECTOR MACHINES

In this section, we present nested extensions of the OC-SVM and the CS-SVM.

3.1. Nested One-Class SVMs

First, we select a finite number of density levels $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_M > 0$ a priori. Then we generate a finite family of nested level set estimates corresponding to the preselected levels. We present a decomposition algorithm as an efficient method to find these level set estimates. Finally, we linearly interpolate the solution coefficients of the nested finite collection to extend to a continuous nested family defined for all λ.

3.1.1. A Finite Family of Nested Sets

Our nested OC-SVM estimates level sets at levels $\lambda_1, \lambda_2, \cdots, \lambda_M$ simultaneously by minimizing the sum of duals (2) corresponding to different levels and by imposing a constraint that causes the resulting sets to be nested. In particular, for M different levels $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_M > 0$, the objective function solves the following optimization problem:

$$\min_{\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_M} \sum_{m=1}^{M} \left[\frac{1}{2\eta\lambda_m} \sum_{i,j} \alpha_{im} \alpha_{jm} k_{ij} - \sum_i \alpha_{im} \right] \tag{6}$$
$$\text{s.t.} \quad 0 \le \alpha_{im} \le \frac{1}{N} \quad \text{for } \forall i, m \tag{7}$$
$$\frac{\alpha_{i1}}{\lambda_1} \le \frac{\alpha_{i2}}{\lambda_2} \le \cdots \le \frac{\alpha_{iM}}{\lambda_M} \quad \text{for } \forall i \tag{8}$$

where $\boldsymbol{\alpha}_m = (\alpha_{1m}, \alpha_{2m}, \cdots, \alpha_{Nm})$ and α_{im} corresponds to data point $\mathbf{x}_i$ and level λ_m ($\alpha_{im} = \alpha_i(\eta\lambda_m)$). Then the optimal solution $\boldsymbol{\alpha}_m^* = (\alpha_{1m}^*, \alpha_{2m}^*, \cdots, \alpha_{Nm}^*)$ defines the decision boundary at level λ_m. The additional inequality constraint (8) makes $\frac{\alpha_{im}}{\lambda}$ monotone, which by (3) implies

$$\widehat{G}_{\lambda_1} \subset \widehat{G}_{\lambda_2} \subset \cdots \subset \widehat{G}_{\lambda_M}.$$

3.1.2. Interpolation

For an intermediate level λ between two levels, say λ_1 and λ_2 without loss of generality, we can write $\lambda = \epsilon\lambda_1 + (1 - \epsilon)\lambda_2$ for some $\epsilon \in [0, 1]$. Then we define the parameter $\alpha_i^*(\lambda)$ through linear interpolation:

$$\alpha_i^*(\lambda) = \epsilon\alpha_i^*(\lambda_1) + (1 - \epsilon)\alpha_i^*(\lambda_2).$$

This method is motivated from the fact that the Lagrange multipliers are piecewise linear in λ for the original SVM. A simple calculation shows that the new level set estimate at level λ

$$\widehat{G}_\lambda = \{\mathbf{x} : \sum_i \alpha_i^*(\eta\lambda) k(\mathbf{x}_i, \mathbf{x}) > \lambda\}$$

satisfies $\widehat{G}_{\lambda_1} \subset \widehat{G}_\lambda \subset \widehat{G}_{\lambda_2}$ and grows in a nested fashion.

3.1.3. *Decomposition Algorithm*

The objective function (6) requires optimization over NM parameters. Due to its large size, standard quadratic programming algorithms are not efficient to solve the nested SVM .

Instead, we use a decomposition algorithm that first divides the large optimization problem into subproblems and then iteratively optimizes the smaller problems. A similar idea was proposed for multiclass classification [15]. Below is the strategy the decomposition algorithm follows:

1. Choose a point $\mathbf{x}_i$ from the data set.
2. Solve the subproblem defined by the point.
3. Repeat 1 and 2 until the value of objective function changes by less than a predetermined tolerance.

Currently, we cycle through all the $\mathbf{x}_i$. In future work, we will investigate the use of optimality conditions to facilitate the choice of $\mathbf{x}_i$. As a starting point of the decomposition algorithm, the feasible point $\alpha_{im} = \frac{1}{N}$ for $\forall i, m$ can be used.

Each step of the algorithm optimizes a set of parameters corresponding to a data point. Without loss of generality, we can assume that the data point is $\mathbf{x}_1$ and $\{\alpha_{1m}\}_{m=1}^{M}$ will be optimized while fixing the other α_{im}. The objective function then can be written in terms of α_{1m}:

$$\begin{aligned}
&\sum_m \left[\frac{1}{2\eta\lambda_m}\sum_{i,j}\alpha_{im}\alpha_{jm}k_{ij} - \sum_i \alpha_{im}\right] \\
=&\sum_m \left[\frac{1}{2\eta\lambda_m}\alpha_{1m}^2 k_{11} + \alpha_{1m}\left(\frac{1}{\eta\lambda_m}\sum_{j\neq 1}\alpha_{jm}k_{1j} - 1\right)\right] + \mathrm{C} \\
=&\sum_m \left[\frac{1}{2\eta\lambda_m}\alpha_{1m}^2 k_{11} + \alpha_{1m}\left(f_{1m} - \frac{\alpha_{1m}^{\text{old}}}{\eta\lambda_m}k_{11} - 1\right)\right] + \mathrm{C} \\
=&\frac{k_{11}}{\eta}\sum_m \left[\frac{1}{2\lambda_m}\alpha_{1m}^2 - \frac{\alpha_{1m}}{\lambda_m}\left(\alpha_{1m}^{\text{old}} + \frac{\eta\lambda_m(1-f_{1m})}{k_{11}}\right)\right] + \mathrm{C}
\end{aligned}$$

where α_{1m}^{old} and $f_{1m} = \frac{1}{\eta\lambda_m}\left(\sum_{j\neq 1}\alpha_{jm}k_{1j} + \alpha_{1m}^{\text{old}}k_{11}\right)$ denote the parameter from the previous iteration step and the corresponding output, respectively. C is a constant that does not depend on α_{1m}.

Then the algorithm solves the new subproblem

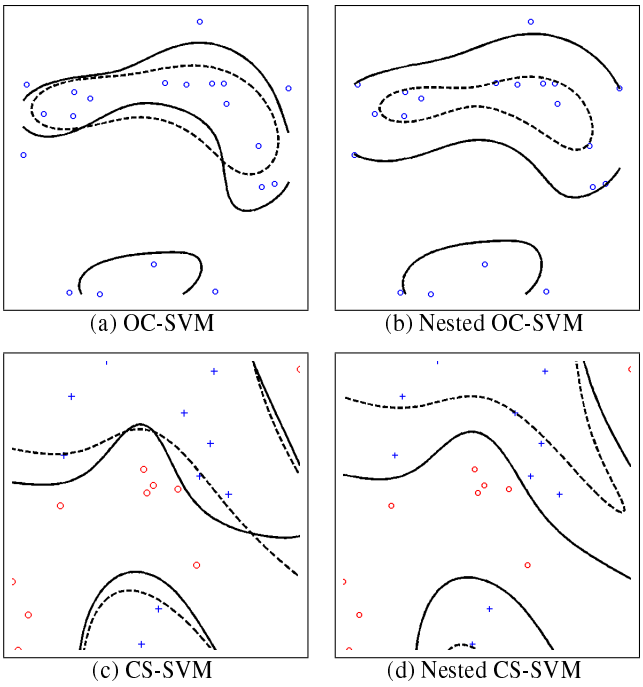
$$\min_{\alpha_{11},\alpha_{12},\cdots,\alpha_{1M}} \sum_m \left[\frac{1}{2\lambda_m}\alpha_{1m}^2 - \frac{\alpha_{1m}}{\lambda_m}\alpha_{1m}^{\text{new}}\right] \tag{9}$$

$$\begin{aligned}
\text{s.t.}\quad & 0 \le \alpha_{1m} \le \frac{1}{N} \quad \text{for } \forall m \\
& \frac{\alpha_{11}}{\lambda_1} \le \frac{\alpha_{12}}{\lambda_2} \le \cdots \le \frac{\alpha_{1M}}{\lambda_M}
\end{aligned} \tag{10}$$

where $\alpha_{1m}^{\text{new}} = \alpha_{1m}^{\text{old}} + \frac{\eta\lambda_m(1-f_{1m})}{k_{11}}$. Notice that α_{1m}^{new} becomes the solution if it is feasible. This subproblem is much smaller and can be solved via standard quadratic optimization techniques.

3.2. Nested Cost-Sensitive SVMs

Nested cost-sensitive SVMs aim to produce nested positive decision sets $\{\mathbf{x} : f_\gamma(\mathbf{x}) > 0\}$ as the cost asymmetry γ varies. Our approach here parallels the approach developed for the OC-SVM.



(a) OC-SVM (b) Nested OC-SVM (c) CS-SVM (d) Nested CS-SVM

Fig. 1. Comparison between SVMs and nested SVMs. Overlapping of decision boundaries are apparent in the left panels, while right panels show nested structures.

For a fixed λ and preselected cost asymmetries $0 \le \gamma_1 \le \gamma_2 \le \cdots \le \gamma_M \le 1$, the nested cost-sensitive SVM minimizes the objective function

$$\min_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_M} \sum_{m=1}^{M}\left[\frac{1}{2\lambda}\sum_{i,j}\alpha_{im}\alpha_{jm}y_i y_j k_{ij} - \sum_i \alpha_{im}\right] \tag{11}$$

$$\text{s.t.}\quad 0 \le \alpha_{im} \le \gamma_m \quad \text{for } i \in I_+, \forall m \tag{12}$$

$$0 \le \alpha_{im} \le 1-\gamma_m \quad \text{for } i \in I_-, \forall m \tag{13}$$

$$\alpha_{i1} \le \alpha_{i2} \le \cdots \le \alpha_{iM} \quad \text{for } i \in I_+ \tag{14}$$

$$\alpha_{i1} \ge \alpha_{i2} \ge \cdots \ge \alpha_{iM} \quad \text{for } i \in I_- \tag{15}$$

where α_{im} corresponds to data point $\mathbf{x}_i$ at cost asymmetry γ_m and $\boldsymbol{\alpha}_m = (\alpha_{1m}, \alpha_{2m}, \cdots, \alpha_{Nm})$. Here the constraints (14) and (15) impose the nesting of set estimates.

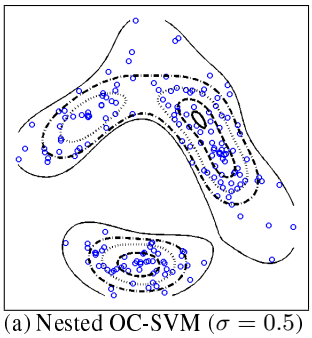
An analogous interpolation and decomposition algorithm can be developed for the nested CS-SVM.

4. EXPERIMENTS

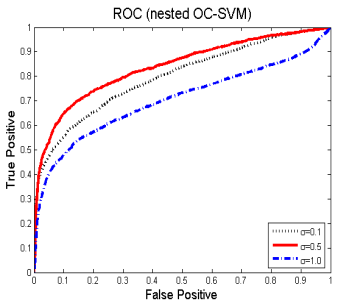
In our implementation, we used $M = 30, \eta = 1.05$ and `quadprog` provided in `Matlab` as a QP solver. Throughout the experiments, we used the normalized Gaussian kernel function

$$k(\mathbf{x}, \mathbf{x}') = \frac{1}{(\sqrt{2\pi}\sigma)^d}\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}\right).$$

Fig. 1 compares the results from the SVM path algorithms and nested SVMs. The left figures are obtained from the unnested solution path algorithms and the right figures are from the nested SVMs. As can be seen, the results from the OC-SVM and the CS-SVM are

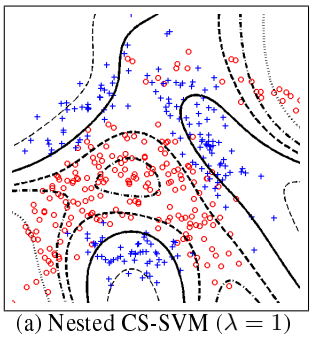


(a) Nested OC-SVM ($\sigma = 0.5$)

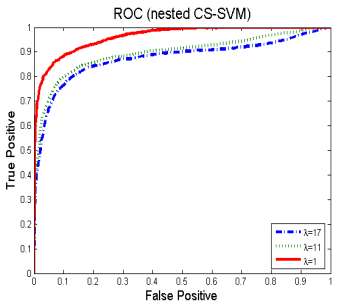


(b) ROC curve

Fig. 2. Nested OC-SVM on "banana" dataset. Small circles represent data points and five contours depict the decision boundaries.



(a) Nested CS-SVM ($\lambda = 1$)



(b) ROC curve

Fig. 3. Nested CS-SVM on "banana" dataset. Small circles and crosses are data points in each class. Decision boundaries for five different values of cost asymmetries are shown.

not nested. On the other hand, the nesting construction is obvious in the right figures.

Fig. 2 illustrates a result of the nested OC-SVM on the two dimensional "banana" dataset[1]. Here only the negative samples were used to train the SVM. In the left panel, five nested level set estimates are shown. The result of the nested CS-SVM on the same dataset is presented in Fig. 3. Five decision boundaries drawn in the figure show the nesting structure; that is, a positively classified data point at γ will be classified positive for any $\gamma' \geq \gamma$.

Correctly selecting the bandwidth of the Gaussian kernel is critical to its performance. A possible choice of bandwidth selection criteria is the area under the ROC curve (AUC), in which we choose σ that maximizes the AUC. For one class problems, we can generate an artifical second class according to a uniform distribution. Using this approach, the right panel in Fig. 2 presents the ROC curves for three different kernel bandwidth values. Small σ leads to overfitting while large σ leads to overly rigid shapes that fail to capture the detail of the underlying density.

The nested CS-SVM has one more control parameter: the regularization parameter λ. A smaller λ penalizes margin errors more and usually generates a more complex decision boundary. λ could also be selected by AUC maximization. Examples of the ROC curves are drawn in Fig. 3.

[1] `http://ida.first.fhg.de/projects/bench/`

5. CONCLUSIONS

In this paper, we introduce a way to impose nesting structures on families of sets produced by SVMs. The key step involves forming a new optimization problem with constraints ensuring the desired structure. A decomposition algorithm is presented to solve the proposed quadratic program. This algorithm generates a finite number of nested set estimates at a set of preselected control parameters, and linear interpolation extends these sets to a continuous nested family. Similar approaches apply to both the OC-SVM and the CS-SVM. The nested OC-SVM yields a family of nested density level set estimates indexed by density level, while the nested CS-SVM yields a family of nested classifiers indexed by cost asymmetry.

6. REFERENCES

[1] J. A. Hartigan, "Consistency of single linkage for high-density clusters," *Journal of the American Statistical Association*, vol. 76, pp. 388–394, 1981.

[2] C. Scott and R. Nowak, "Learning minimum volume sets," *Journal of Machine Learning Research*, vol. 7, pp. 665–704, 2006.

[3] C. Scott and E. D. Kolaczyk, "Annotated minimum volume sets for nonparametric anomaly discovery," in *IEEE Workshop on Statistical Signal Processing*, 2007, pp. 234–238.

[4] C. Scott and R. Nowak, "A Neyman-Pearson approach to statistical learning," *IEEE Transactions on Information Theory*, vol. 51, pp. 3806–3819, 2005.

[5] C. Scott and G. Blanchard, "Transductive anomaly detection," Tech. Rep., 2008, `http://www.eecs.umich.edu/~cscott`.

[6] B. Schölkopf and A.J. Smola, *Learning with Kernels*, MIT Press, Cambridge, MA, 2002.

[7] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu, "The entire regularization path for the support vector machine," *Journal of Machine Learning Research*, vol. 5, pp. 1391–1415, 2004.

[8] F. R. Bach, D. Heckerman, and E. Horvitz, "Considering cost asymmetry in learning classifiers," *Journal of Machine Learning Research*, vol. 7, pp. 1713–1741, 2006.

[9] G. Lee and C. D. Scott, "The one class support vector machine solution path," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2007, vol. 2, pp. II–521–II–524.

[10] B. Schölkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, and R.C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, pp. 1443–1472, 2001.

[11] D.M.J. Tax and R.P.W. Duin, "Support vector domain description," *Pattern Recognition Letters*, vol. 20, pp. 1191–1199, 1999.

[12] C.C. Chang and C.J. Lin, "Training ν-support vector classifiers: Theory and algorithm," *Neural Computation*, vol. 13, pp. 2119–2147, 2001.

[13] R. Vert and J. Vert, "Consistency and convergence rates of one-class SVMs and related algorithms," *Journal of Machine Learning Research*, vol. 7, pp. 817–854, 2006.

[14] E. Osuna, R. Freund, and F. Girosi, "Support vector machines: Training and applications," Tech. Rep. AIM-1602, MIT Artificial Intelligence Laboratory, Mar 1997.

[15] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2001.