NEW YORK UNIVERSITY

Support Vector Novelty Detection

Discussion of "Support Vector Method for Novelty Detection" (NIPS 2000) and "Estimating the Support of a High-Dimensional Distribution" (Neural Computation 13, 2001)
Bernhard Scholkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson

Eugene Weinstein
March 3rd, 2009
Machine Learning Seminar

One Does Not Belong

- Mexico's outgoing ruling party is threatening to boycott the inauguration of the new President elect Vicente Fox...
- Leaders from around the world are arriving in Mexico for Friday's inauguration of President elect Vicente Fox...
- The curtain has come down on the Summer Games in Sydney but not before the U. S. men's basketball team...
- Mexico begins a new era today when President elect Vicente Fox takes the oath of office...
- History was made in Mexico today when Vicente Fox was sworn in as President. This was no ordinary inauguration...

Novelty Detection

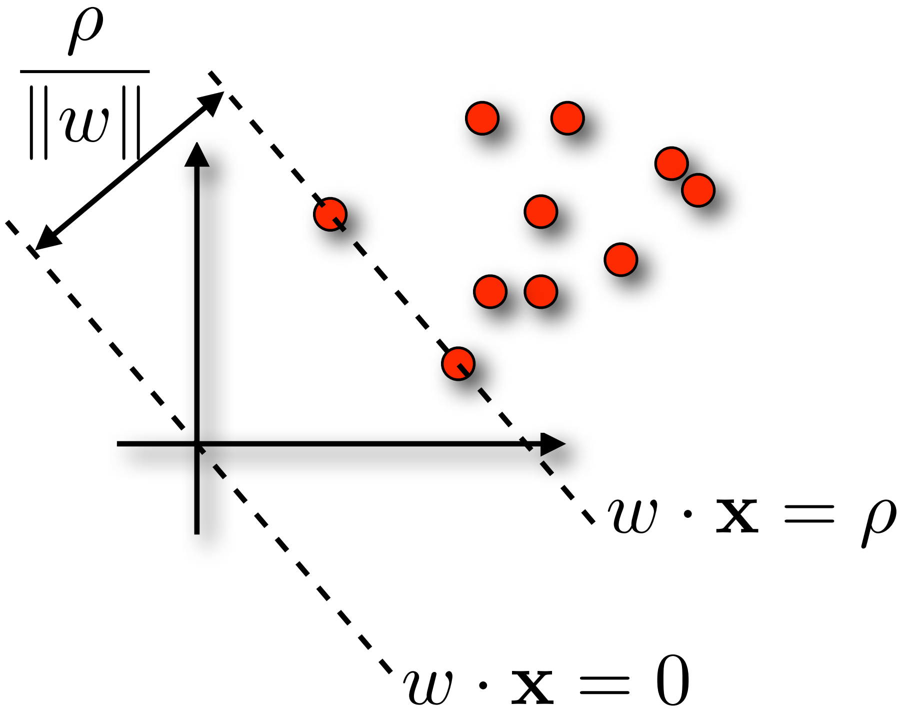
- Intuitive definition: find the outliers in a group or a stream of data points
- Some ideas in the (enormous) past literature
  - Fit a Gaussian mixture, train a HMM, etc.; outliers are points with low likelihood
  - Apply k-means clustering, k-nearest neighbors, etc.; outliers are points far from clusters/other points
- This work: ummm, Occam's Razor? Why solve all these hard problems?

Separating from Origin

- Want an algorithm that returns +1 in a "small" region enclosing most of the points, -1 outside this region
- Unsupervised setting: data points $\mathbf{x}_1, \ldots, \mathbf{x}_\ell \in \mathcal{X}$
- Linear classifier form

$$f(x) = \mathrm{sgn}(w \cdot \mathbf{x} - \rho)$$

- "separable case" here



Add Slack Variables, Kernel

- Allow points to violate margin constraints: slack

$$\min_{w\in F, \boldsymbol{\xi}\in\mathbb{R}^{\ell}, \rho\in\mathbb{R}} \quad \frac{1}{2}\|w\|^2 + \frac{1}{\nu\ell}\sum_i \xi_i - \rho$$

$$\text{subject to } (w\cdot\Phi(\mathbf{x}_i)) \geq \rho - \xi_i, \;\; \xi_i \geq 0.$$

- # points: ℓ, : learning parameter $\nu \in (0,1]$
- Kernel version uses mapping

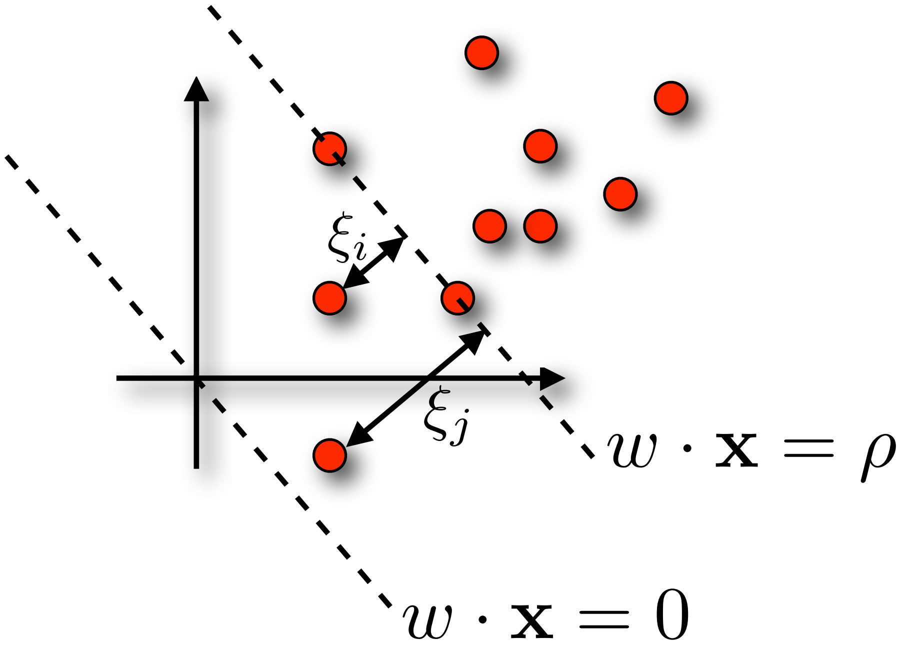
$$\Phi : \mathcal{X} \rightarrow F$$

- Add kernel

$$k(\mathbf{x},\mathbf{y}) = (\Phi(\mathbf{x})\cdot\Phi(\mathbf{y}))$$

- Kernelized classifier:

$$f(\mathbf{x}) = \operatorname{sgn}((w\cdot\Phi(\mathbf{x})) - \rho)$$



Solving the Optimization

- Lagrangian $L(w, \boldsymbol{\xi}, \rho, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|w\|^2 + \frac{1}{\nu\ell}\sum_i \xi_i - \rho$

  $(\alpha_i, \beta_i \geq 0)$ $\qquad - \sum_i \alpha_i \left((w \cdot \Phi(\mathbf{x}_i)) - \rho + \xi_i \right) - \sum_i \boldsymbol{\beta}_i \xi_i,$

- Set derivatives w.r.t. w, ξ, ρ equal to zero

$$w = \sum_i \alpha_i \Phi(\mathbf{x}_i),$$

$$\alpha_i = \frac{1}{\nu\ell} - \boldsymbol{\beta}_i \leq \frac{1}{\nu\ell}, \quad \sum_i \alpha_i = 1.$$

- Plug back in to get (kernelized) dual problem

$$\min_{\boldsymbol{\alpha}} \frac{1}{2} \sum_{ij} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \text{ subject to } 0 \leq \alpha_i \leq \frac{1}{\nu\ell}, \qquad \sum_i \alpha_i = 1.$$

- And the classifier form $f(\mathbf{x}) = \operatorname{sgn}\left(\sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) - \rho \right).$

Recovering the Threshold

- Two of the KKT conditions:

$$\forall_i, \;\; \alpha_i[w \cdot \Phi(\mathbf{x}_i) - \rho + \xi_i] = 0; \beta_i \xi_i = 0$$

- For any i such that $\alpha_i, \beta_i > 0$ (SVs), we have

$$\rho = (w \cdot \Phi(\mathbf{x}_i)) = \sum_j \alpha_j k(\mathbf{x}_j, \mathbf{x}_i).$$

- So you get the threshold back as part of your solution
  - i.e., no hacking/tuning/guessing the threshold!

Another View

Tax and Duin, '99

- We can also do novelty detection by finding the smallest-radius sphere/ball enclosing most of the points

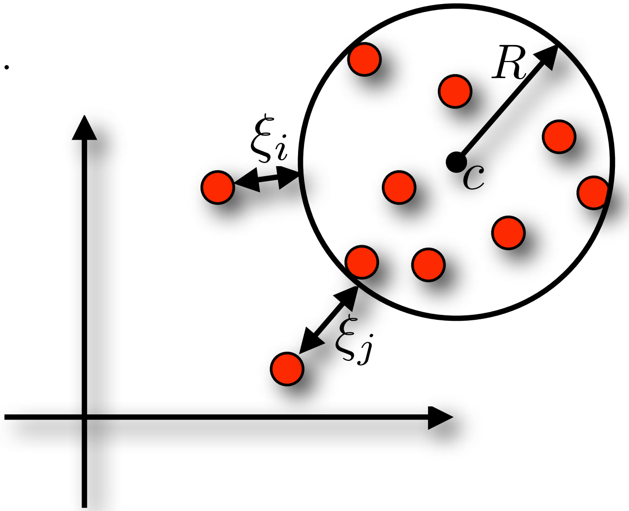
$$\min_{R\in\mathbb{R},\boldsymbol{\xi}\in\mathbb{R}^\ell,c\in F} \quad R^2 + \frac{1}{\nu\ell}\sum_i \xi_i$$

$$\text{subject to} \quad \|\Phi(\mathbf{x}_i) - c\|^2 \le R^2 + \xi_i, \quad \xi_i \ge 0 \text{ for } i \in [\ell].$$

This leads to the dual

$$\min_{\boldsymbol{\alpha}} \quad \sum_{ij} \alpha_i\alpha_j k(\mathbf{x}_i,\mathbf{x}_j) - \sum_i \alpha_i k(\mathbf{x}_i,\mathbf{x}_i)$$

$$\text{subject to} \quad 0 \le \alpha_i \le \frac{1}{\nu\ell}, \quad \sum_i \alpha_i = 1$$



and the solution

$$c = \sum_i \alpha_i \Phi(\mathbf{x}_i),$$

corresponding to a decision function of the form

$$f(\mathbf{x}) = \operatorname{sgn}\left(R^2 - \sum_{ij}\alpha_i\alpha_j k(\mathbf{x}_i,\mathbf{x}_j) + 2\sum_i \alpha_i k(\mathbf{x}_i,\mathbf{x}) - k(\mathbf{x},\mathbf{x})\right).$$

Equivalence of Spheres

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \sum_{ij} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) - \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}_i) \\ \text{subject to} \quad & 0 \le \alpha_i \le \frac{1}{\nu\ell}, \ \sum_i \alpha_i = 1 \end{aligned}$$

- If kernel $k(\mathbf{x}, \mathbf{y})$depends on only $\mathbf{x} - \mathbf{y}$, $k(\mathbf{x}, \mathbf{x})$ is constant
  - e.g., Gaussian Kernel $k(\mathbf{x}, \mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|^2/c}$
- In these cases, spheres equivalent to origin separation

$$\min_{\boldsymbol{\alpha}} \frac{1}{2} \sum_{ij} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \text{ subject to } 0 \le \alpha_i \le \frac{1}{\nu\ell}, \qquad \sum_i \alpha_i = 1.$$

Generalization Bound

- 3.4, 3.5: $\min_{w \in F, \boldsymbol{\xi} \in \mathbb{R}^{\ell}, \rho \in \mathbb{R}} \quad \frac{1}{2}\|w\|^2 + \frac{1}{\nu\ell}\sum_i \xi_i - \rho$, 3.10: $f(\mathbf{x}) = \operatorname{sgn}\left(\sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) - \rho\right).$

  subject to $(w \cdot \Phi(\mathbf{x}_i)) \geq \rho - \xi_i, \quad \xi_i \geq 0.$

- 3.12:

$$\rho = (w \cdot \Phi(\mathbf{x}_i)) = \sum_j \alpha_j k(\mathbf{x}_j, \mathbf{x}_i).$$

Definition 2. *Let f be a real-valued function on a space $\mathcal{X}$. Fix $\theta \in \mathbb{R}$. For $\mathbf{x} \in \mathcal{X}$ let $d(\mathbf{x}, f, \theta) = \max\{0, \theta - f(\mathbf{x})\}$. Similarly for a training sequence $\mathbf{X} := (\mathbf{x}_1, \ldots, \mathbf{x}_\ell)$, we define*

$$\mathcal{D}(\mathbf{X}, f, \theta) = \sum_{\mathbf{x} \in \mathbf{X}} d(\mathbf{x}, f, \theta).$$

Theorem 1 (generalization error bound). *Suppose we are given a set of ℓ examples $\mathbf{X} \in \mathcal{X}^{\ell}$ generated i.i.d. from an unknown distribution P, which does not contain discrete components. Suppose, moreover, that we solve the optimization problem, equations 3.4 and 3.5 (or equivalently equation 3.11) and obtain a solution f_w given explicitly by equation (3.10). Let $R_{w,\rho} := \{\mathbf{x}\colon f_w(\mathbf{x}) \geq \rho\}$ denote the induced decision region. With probability $1-\delta$ over the draw of the random sample $\mathbf{X} \in \mathcal{X}^{\ell}$, for any $\gamma > 0$,*

$$P\left\{\mathbf{x}'\colon \mathbf{x}' \notin R_{w,\rho-\gamma}\right\} \leq \frac{2}{\ell}\left(k + \log\frac{\ell^2}{2\delta}\right), \tag{5.7}$$

where

$$k = \frac{c_1 \log\left(c_2 \hat{\gamma}^2 \ell\right)}{\hat{\gamma}^2} + \frac{2\mathcal{D}}{\hat{\gamma}} \log\left(e\left(\frac{(2\ell-1)\hat{\gamma}}{2\mathcal{D}} + 1\right)\right) + 2, \tag{5.8}$$

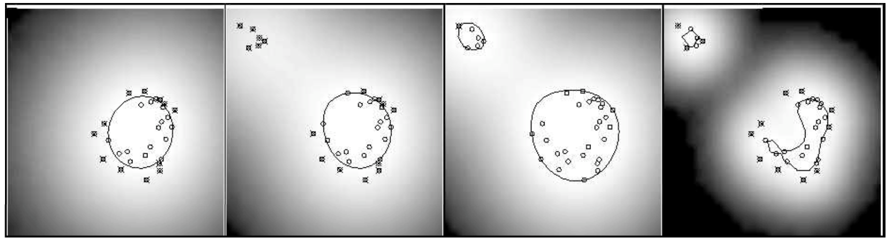
$c_1 = 16c^2$, $c_2 = \ln(2)/\left(4c^2\right)$, $c = 103$, $\hat{\gamma} = \gamma/\|w\|$, $\mathcal{D} = \mathcal{D}(\mathbf{X}, f_{w,0}, \rho) = \mathcal{D}(\mathbf{X}, f_{w,\rho}, 0)$, and ρ is given by equation (3.12).

Generalization Comments

- Covering number argument, strong connections to soft-margin binary classification bounds
- Bound is loose and thus not directly applicable in practice
    - $c = 103$, too large by a factor of >50

Experiments: Toy Data

- Synthetic 2-D data, Gaussian kernel $k(\mathbf{x}, \mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|^2/c}$



ν, width c	0.5, 0.5	0.5, 0.5	0.1, 0.5	0.5, 0.1
frac. SVs/OLs	0.54, 0.43	0.59, 0.47	0.24, 0.03	0.65, 0.38
margin $\rho/\|w\|$	0.84	0.70	0.62	0.48

Figure 1: (First two pictures) A single-class SVM applied to two toy problems; $\nu = c = 0.5$, domain: $[-1, 1]^2$. In both cases, at least a fraction of ν of all examples is in the estimated region (cf. Table 1). The large value of ν causes the additional data points in the upper left corner to have almost no influence on the decision function. For smaller values of ν, such as 0.1 (third picture) the points cannot be ignored anymore. Alternatively, one can force the algorithm to take these outliers (OLs) into account by changing the kernel width (see equation 3.3). In the *fourth picture*, using $c = 0.1, \nu = 0.5$, the data are effectively analyzed on a different length scale, which leads the algorithm to consider the outliers as meaningful points.

Experiments: Digits

- 9298 digits, 16x16=256 dimensionality, last 2007 are test

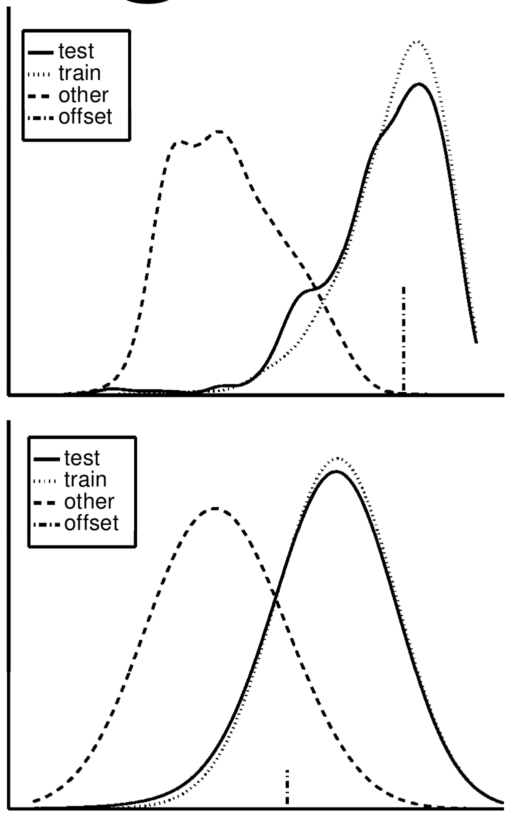


Figure 3: Experiments on the U.S. Postal Service OCR data set. Recognizer for digit 0; output histogram for the exemplars of 0 in the training/test set, and on test exemplars of other digits. The *x*-axis gives the output values, that is, the argument of the sgn function in equation 3.10. For $\nu = 50\%$ (top), we get 50% SVs and 49% outliers (consistent with proposition 3), 44% true positive test examples, and zero false positives from the "other" class. For $\nu = 5\%$ (bottom), we get 6% and 4% for SVs and outliers, respectively. In that case, the true positive rate is improved to 91%, while the false-positive rate increases to 7%. The offset ρ is marked in the graphs. Note, finally, that the plots show a Parzen windows density estimate of the output histograms. In reality, many examples sit exactly at the threshold value (the nonbound SVs). Since this peak is smoothed out by the estimator, the fractions of outliers in the training set appear slightly larger than it should be.

Experiments: Digits

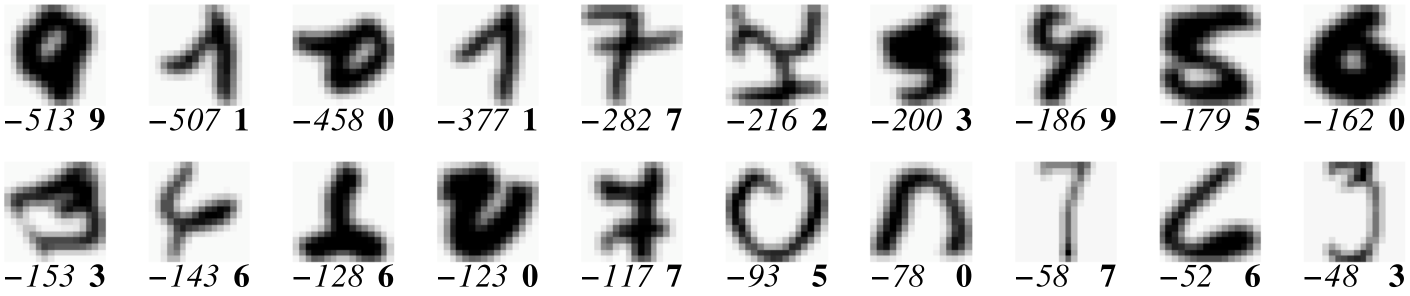


Figure 5: Outliers identified by the proposed algorithm, ranked by the negative output of the SVM (the argument of equation 3.10). The outputs (for convenience in units of 10^{-5}) are written underneath each image in italics; the (alleged) class labels are given in boldface. Note that most of the examples are “difficult” in that they are either atypical or even mislabeled.