CS 5526 – Virginia Tech
Homework 2

Brennon Bortz
(PID: brennon, Campus: Blacksburg)

1 April 2015

Code for this assignment is available at `https://github.com/brennon/cs5526-hw2`.

Written Problems

1. (10 points) Prove that the dimensionality of the feature space for the inhomogeneous polynomial kernel of degree q is

$$m = \binom{d+q}{q}$$

Solution: Per the text, the mapping $\phi : \mathbb{R}^d \to \mathbb{R}^m$ is given as the vector

$$\phi(\mathbf{x}) = (\ldots, a_n \mathbf{x}^{\mathbf{n}}, \ldots)^T = \cdots$$

where the variable $\mathbf{n} = (n_0, \ldots, n_d)$ ranges over all the possible assignments such that $|\mathbf{n}| = q$ (the sum of $\mathbf{n}$ equals q). This is equivalent to the number of monomial terms in a multinomial expansion, and can be counted by the stars and bars method.

To illustrate this, let $q = 5$ and $d = 3$. We are searching for all assignments of non-negative integers to the values $n_0, \ldots, n_d$ such that they sum to q. Arrange 5 stars (*) in a row. Place 4 bars (as there are $d + 1$ terms in $n_0, \ldots, n_d$) (—) between any of the stars, to the left of the entire row of stars, or to the right of the entire row of stars. Moving from left to right, groups of stars are counted and assigned to each n_i. Where there are no stars between two bars, the value of 0 is given to the corresponding n_i.

There are $\binom{d+1+q-1}{q} = \binom{d+q}{q}$ ways to arrange the stars and bars, and thus there are $\binom{d+q}{q}$ assignments of non-negative integers to the values $n_0, \ldots, n_d$ such that they sum to q.

2. (12 points) Given the three points $\mathbf{x}_1 = (2.5, 1)^T$, $\mathbf{x}_1 = (3.5, 4)^T$, and $\mathbf{x}_1 = (2, 2.1)^T$.

   (a) Compute the kernel matrix for the Gaussian kernel assuming that $\sigma^2 = 5$.

$$
\begin{aligned}
K(\mathbf{x}_1, \mathbf{x}_1) &= \exp\left(\frac{\left\|\begin{pmatrix} 2.5 & 1 \end{pmatrix} - \begin{pmatrix} 2.5 & 1 \end{pmatrix}\right\|^2}{10}\right) \\
&= \exp\left(\frac{\left\|\begin{pmatrix} 0 & 0 \end{pmatrix}\right\|^2}{10}\right) \\
&= \exp\left(-\frac{0}{10}\right) \\
&= \exp(0) \\
&= 1
\end{aligned}
$$

Similar computations show that $K(\mathbf{x}_1, \mathbf{x}_1) = K(\mathbf{x}_2, \mathbf{x}_2) = K(\mathbf{x}_3, \mathbf{x}_3) = 1$.

$$
\begin{aligned}
K(\mathbf{x}_2, \mathbf{x}_1) = K(\mathbf{x}_1, \mathbf{x}_2) &= \exp\left(\frac{\left\|\begin{pmatrix} 2.5 & 1 \end{pmatrix} - \begin{pmatrix} 3.5 & 4 \end{pmatrix}\right\|^2}{10}\right) \\
&= \exp\left(\frac{\left\|\begin{pmatrix} -1 & -3 \end{pmatrix}\right\|^2}{10}\right) \\
&= \exp\left(-\frac{\sqrt{10}^2}{10}\right) \\
&= \exp(-1) \\
&= 0.3679
\end{aligned}
$$

$$
\begin{aligned}
K(\mathbf{x}_3, \mathbf{x}_1) = K(\mathbf{x}_1, \mathbf{x}_3) &= \exp\left(\frac{\left\|\begin{pmatrix} 2.5 & 1 \end{pmatrix} - \begin{pmatrix} 2.5 & 2.1 \end{pmatrix}\right\|^2}{10}\right) \\
&= \exp\left(\frac{\left\|\begin{pmatrix} 0.5 & -1.1 \end{pmatrix}\right\|^2}{10}\right) \\
&= \exp\left(-\frac{\sqrt{1.46}^2}{10}\right) \\
&= \exp(-0.146) \\
&= 0.8642
\end{aligned}
$$

$$
\begin{aligned}
K(\mathbf{x}_3, \mathbf{x}_2) = K(\mathbf{x}_2, \mathbf{x}_3) &= \exp\left(\frac{\left\|\begin{pmatrix} 3.5 & 4 \end{pmatrix} - \begin{pmatrix} 2 & 2.1 \end{pmatrix}\right\|^2}{10}\right) \\
&= \exp\left(\frac{\left\|\begin{pmatrix} 1.5 & 1.9 \end{pmatrix}\right\|^2}{10}\right) \\
&= \exp\left(-\frac{\sqrt{5.86}^2}{10}\right) \\
&= \exp\left(-0.586\right) \\
&= 0.5565
\end{aligned}
$$

Putting these together, we have

$$
K = \begin{bmatrix} 1 & 0.3679 & 0.8642 \\ 0.3679 & 1 & 0.5565 \\ 0.8642 & 0.5565 & 1 \end{bmatrix}
$$

Solution:

(b) Compute the distance of the point $\phi(\mathbf{x}_1)$ from the mean in feature space.

Solution:

$$
\|\phi(\mathbf{x}_1) - \mu_\phi\|^2 = K(\mathbf{x}_1, \mathbf{x}_1) - \frac{2}{n}\sum_{j=1}^{n} K(\mathbf{x}_1, \mathbf{x}_j) + \frac{1}{n^2}\sum_{a=1}^{n}\sum_{b=1}^{n} K(\mathbf{x}_a, \mathbf{x}_b)
$$

$$
\begin{aligned}
\frac{1}{n^2}\sum_{a=1}^{n}\sum_{b=1}^{n} K(\mathbf{x}_a, \mathbf{x}_b) &= \frac{1}{3^2}\left[1 + 0.3679 + 0.8642 + 0.3679 + 1 + 0.5565 + 0.8642 + 0.5565 + 1\right] \\
&= \frac{6.5572}{9} \\
&= 0.7308
\end{aligned}
$$

$$
\begin{aligned}
\frac{2}{n}\sum_{j=1}^{n} K(\mathbf{x}_1, \mathbf{x}_j) &= \frac{2}{3}\left[1 + 0.3679 + 0.8642\right] \\
&= \frac{2.2321}{3} = 1.4881
\end{aligned}
$$

$$
\begin{aligned}
\|\phi(\mathbf{x}_1) - \mu_\phi\|^2 &= 1.0 - 1.4881 + 0.7308 \\
&= 0.2427 \\
\|\phi(\mathbf{x}_1) - \mu_\phi\| &= 0.2427^2 = 0.0589
\end{aligned}
$$

(c) Compute the dominant eigenvector and eigenvalue for the kernel matrix from (a).

Solution:

MATLAB gives the following eigenvalues for the matrix in (a): 0.1084, 0.674, and 2.2176. The corresponding eigenvector for the domninant eigenvalue (2.2176) is:

$$\begin{pmatrix} 0.6002 \\ 0.4754 \\ 0.6433 \end{pmatrix}$$

3. (30 points) Consider the dataset in Figure 21.9 (shown in the text), which has points from two classes c_1 (triangles) and c_2 (circles). Answer the questions below.

(a) Find the equations for the two hyperplanes h_1 and h_2.

Solution:

We use the points $x_1 = (\ 6 \quad 0\ )$ and $x_2 = (\ 5 \quad 2\ )$ on h_1 to find the equation for h_1

$$h_1 = \mathbf{w}^T\mathbf{x} + b = w_1x_1 + w_2x_2 + b = 0$$

Rearranging terms

$$x_2 = -\frac{w_1}{w_2}x_1 - \frac{b}{w_2}$$

where $-\frac{w_1}{w_2}$ is the slope of the line, and $-\frac{b}{w_2}$ is the intercept along the second dimension.

$$-\frac{w_1}{w_2} = \frac{2-0}{5-6} = -\frac{2}{1}$$

which implies that $w_1 = 2$ and $w_2 = 1$. We compute the offset b directly

$$b = -2x_1 - 1x_2 = -2 \cdot 6 - 1 \cdot 0 = -12$$

Thus, $\mathbf{w} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ is the weight vector, and $b = -12$ is the bias, and the equation of the hyperplane is given as

$$h(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = (\ 2 \quad 1\ ) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - 12 = 0$$

We use the points $x_1 = (\ 2 \quad 0\ )$ and $x_2 = (\ 5 \quad 5\ )$ on h_1 to find the equation for h_1

$$-\frac{w_1}{w_2} = \frac{5-0}{5-2} = \frac{5}{3}$$

which implies that $w_1 = -5$ and $w_2 = 3$. We compute the offset b directly

$$b = -(-5)x_1 - 3x_2 = 5 \cdot 5 - 3 \cdot 5 = 10$$

Thus, $\mathbf{w} = \begin{pmatrix} -5 \\ 3 \end{pmatrix}$ is the weight vector, and $b = 10$ is the bias, and the equation of the hyperplane is given as

$$h(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = \begin{pmatrix} -5 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + 10 = 0$$

(b) Show all the support vectors for h_1 and h_2.

Solution:

The support vectors for h_1 are $\begin{pmatrix} 2 & 6 \end{pmatrix}$, $\begin{pmatrix} 3 & 4 \end{pmatrix}$, and $\begin{pmatrix} 6 & 2 \end{pmatrix}$.

The support vectors for h_2 are $\begin{pmatrix} 3 & 4 \end{pmatrix}$ and $\begin{pmatrix} 7 & 6 \end{pmatrix}$.

(c) Which of the two hyperplanes is better at separating the two classes based on the margin computation?

To find the margin of each hyperplane, we must first find the canonical hyperplanes for h_1 and h_2.

The equation of the separating hyperplane h_1 is

$$h_1(\mathbf{x}) = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \mathbf{x} - 12 = 0$$

Consider the support vector $\mathbf{x}^* = (3, 4)^T$, with class $y^* = -1$. To find the canonical hyperplane equation, we have to rescale the weight vector and bias by the scalar s

$$s = \frac{1}{y^* h_1(\mathbf{x}^*)} = \frac{1}{-1\left(\begin{pmatrix} 2 \\ 1 \end{pmatrix}^T \begin{pmatrix} 3 \\ 4 \end{pmatrix} - 12\right)} = \frac{1}{2}$$

Thus, the rescaled weight vector is

$$\mathbf{w} = \frac{1}{2}\begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

and the rescaled bias is

$$b = \frac{1}{2} \cdot -12 = -6$$

The canonical form of the hyperplane is therefore

$$h_1(\mathbf{x}) = \mathbf{x} - 6$$

and the margin of the canonical hyperplane for h_1 is

$$\delta^* = \frac{y^* h(\mathbf{x}^*)}{\|\mathbf{w}\|} = \frac{1}{1} = 1$$

The equation of the separating hyperplane h_2 is

$$h_2(\mathbf{x}) = \begin{pmatrix} -5 \\ 3 \end{pmatrix} \mathbf{x} + 10 = 0$$

Consider the support vector $\mathbf{x}^* = (3, 4)^T$, with class $y^* = -1$. To find the canonical hyperplane equation, we have to rescale the weight vector and bias by the scalar s

$$s = \frac{1}{y^* h_2(\mathbf{x}^*)} = \frac{1}{-1\left(\begin{pmatrix} -5 \\ 3 \end{pmatrix}^T \begin{pmatrix} 3 \\ 4 \end{pmatrix} + 10\right)} = \frac{1}{-1(-3+10)} = -\frac{1}{7}$$

Thus, the rescaled weight vector is

$$\mathbf{w} = -\frac{1}{7}\begin{pmatrix} -5 \\ 3 \end{pmatrix} = \begin{pmatrix} 5/7 \\ -3/7 \end{pmatrix}$$

and the rescaled bias is

$$b = -\frac{1}{7} \cdot 10 = -\frac{10}{7}$$

The canonical form of the hyperplane is therefore

$$h_2(\mathbf{x}) = \begin{pmatrix} 5/7 \\ -3/7 \end{pmatrix}^T \mathbf{x} - \frac{10}{7} = \begin{pmatrix} 0.7143 \\ -0.4286 \end{pmatrix}^T \mathbf{x} - 1.4286$$

and the margin of the canonical hyperplane for h_2 is

$$\delta^* = \frac{y^* h(\mathbf{x}^*)}{\|\mathbf{w}\|} = \frac{1}{\sqrt{0.7143^2 + (-0.4286)^2}} = \frac{1}{\sqrt{0.5102 + 0.1387}} = \frac{1}{0.6489} = -1.5411$$

Thus, the hyperplane h_2 is better at separating the two classes based on these margin computations.

Solution:

(d) Find the equation of the best separating hyperplane for this dataset, and show the corresponding support vectors. You can do this without having to solve the Lagrangian by considering the convex hull of each class and the possible hyperplanes at the boundary of the two classes.

Solution:

Based on the latter approach suggested in the problem, two of the points on the best separating hyperplane are $x_1 = (4.5, 2)$ and $x_2 = (5.5, 6)$. Following the logic from part (a) we use these points to find the equation for this new separating hyperplane h_3

$$h_3 = \mathbf{w}^T \mathbf{x} + b = w_1 x_1 + w_2 x_2 + b = 0$$

Rearranging terms

$$x_2 = -\frac{w_1}{w_2}x_1 - \frac{b}{w_2}$$

where $-\frac{w_1}{w_2}$ is the slope of the line, and $-\frac{b}{w_2}$ is the intercept along the second dimension.

$$-\frac{w_1}{w_2} = \frac{6-2}{5.5-4.5} = \frac{4}{1}$$

which implies that $w_1 = -4$ and $w_2 = 1$. We compute the offset b directly

$$b = 4x_1 - 1x_2 = 4 \cdot 4.5 - 1 \cdot 2 = 18 - 2 = 16$$

Thus, $\mathbf{w} = \begin{pmatrix} -4 \\ 1 \end{pmatrix}$ is the weight vector, and $b = 16$ is the bias, and the equation of the hyperplane is given as

$$h(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = \begin{pmatrix} -4 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + 16 = 0$$

4. (18 points) Given the 10 points in Table 21.2 (shown in the text), along with their classes and their Lagranian multipliers (α_i), answer the following questions:

   (a) What is the equation of the SVM hyperplane $h(x)$?

   Solution:

   We compute the weight vector for the hyperplane with

$$\begin{aligned}
\mathbf{w} &= \sum_{i,\alpha_i>0} \alpha_i y_i \mathbf{x}_i \\
&= 0.414\begin{pmatrix} 4 \\ 2.9 \end{pmatrix} - 0.018\begin{pmatrix} 2.5 \\ 1 \end{pmatrix} + 0.018\begin{pmatrix} 3.5 \\ 4 \end{pmatrix} - 0.414\begin{pmatrix} 2 \\ 2.1 \end{pmatrix} \\
&= \begin{pmatrix} 0.846 \\ 0.3852 \end{pmatrix}
\end{aligned}$$

   The final bias is the average of the bias obtained from each support vector using

$$b_i = y_i - \mathbf{w}^T\mathbf{x}_i$$

$$
\begin{aligned}
b_1 &= 1 - \begin{pmatrix} 0.846 \\ 0.3852 \end{pmatrix}^T \begin{pmatrix} 4 \\ 2.9 \end{pmatrix} = -1.22692 \\
b_4 &= -1 - \begin{pmatrix} 0.846 \\ 0.3852 \end{pmatrix}^T \begin{pmatrix} 2.5 \\ 1 \end{pmatrix} = -2.7298 \\
b_7 &= 1 - \begin{pmatrix} 0.846 \\ 0.3852 \end{pmatrix}^T \begin{pmatrix} 3.5 \\ 4 \end{pmatrix} = -0.4202 \\
b_9 &= -1 - \begin{pmatrix} 0.846 \\ 0.3852 \end{pmatrix}^T \begin{pmatrix} 2 \\ 2.1 \end{pmatrix} = -1.88308 \\
b &= \text{avg}(b_i) = -1.575
\end{aligned}
$$

Thus, the optimal hyperplane is given as follows:

$$
h(\mathbf{x}) = \begin{pmatrix} 0.846 \\ 0.3852 \end{pmatrix}^T \mathbf{x} - 1.575 = 0
$$

(b) What is the distance of x_6 from the hyperplane? Is it within the margin of the classifier?

Solution:

To compute this distance *and* determine whether or not it lies in the margin, we must first find the canonical form of the separating hyperplane. The equation of the separating hyperplane h is

$$
h(\mathbf{x}) = \begin{pmatrix} 0.846 \\ 0.3852 \end{pmatrix}^T \mathbf{x} - 1.575 = 0
$$

Consider the support vector $\mathbf{x}^* = \mathbf{x}_1 = (4, 2.9)^T$, with class $y^* = 1$. To find the canonical hyperplane equation, we have to rescale the weight vector and bias by the scalar s

$$
s = \frac{1}{y^* h_1(\mathbf{x}^*)} = \frac{1}{1\left(\begin{pmatrix} 0.846 \\ 0.3852 \end{pmatrix}^T \begin{pmatrix} 4 \\ 2.9 \end{pmatrix} - 1.575\right)} = \frac{1}{4.5018 - 1.575} = \frac{1}{2.92608} = 0.3418
$$

Thus, the rescaled weight vector is

$$
\mathbf{w} = 0.3418 \begin{pmatrix} 0.846 \\ 0.3852 \end{pmatrix} = \begin{pmatrix} 0.2892 \\ 0.1317 \end{pmatrix}
$$

and the rescaled bias is

$$
b = 0.348 \cdot -1.575 = -0.5383
$$

The canonical form of the hyperplane is therefore

$$h(\mathbf{x}) = \begin{pmatrix} 0.2892 \\ 0.1317 \end{pmatrix}^T \mathbf{x} - 0.5383$$

and the margin of the canonical hyperplane for h is

$$\delta^* = \frac{y^* h(\mathbf{x}^*)}{\|\mathbf{w}\|} = \frac{1}{\sqrt{0.2892^2 + 0.1317^2}} = \frac{1}{\sqrt{0.0836 + 0.0173}} = \frac{1}{\sqrt{0.1009}} = 0.3176$$

The distance from $\mathbf{x}_6$ to the hyperplane is then given by

$$\begin{aligned} \delta^* &= \frac{y_6(\mathbf{x}_6)}{\|\mathbf{w}\|} = \frac{-1 \begin{pmatrix} 0.2892 \\ 0.1317 \end{pmatrix}^T \begin{pmatrix} 1.9 \\ 1.9 \end{pmatrix} - 0.5383}{0.3176} \\ &= \frac{-1 \cdot 0.79971 - 0.5383}{0.3176} \\ &= \frac{-1.33801}{0.3176} \\ &= -4.2129 \end{aligned}$$

While the SVM misclassifies $\mathbf{x}_6$, it does not lie within the SVM margin.

(c) Classify the point $\mathbf{z} = (3, 3)_T$ using $h(x)$ from above.

Solution:

We classify the point $\mathbf{z}$ by

$$\begin{aligned} \hat{y} &= \text{sign}(h(\mathbf{z})) \\ &= \text{sign}(\mathbf{w}^T \mathbf{z} + b) \\ &= \text{sign}\left(\begin{pmatrix} 0.2892 \\ 0.1317 \end{pmatrix}^T \begin{pmatrix} 3 \\ 3 \end{pmatrix} - 0.5383 \right) \\ &= \text{sign}(1.2627 - 0.5383) \\ &= \text{sign}(0.7244) \\ &= + \end{aligned}$$

5. (30 points) Create a binary (2 feature) dataset where the target (2-class) variable encodes the XOR function. Design and implement a SVM (with a suitable kernel) to learn a classifier for this dataset. For full credit, explain the kernel you selected, and the support vectors picked by the algorithm. Redo all the above with multiple settings involving more than 2 features. Ensure that your kernel is able to model XOR in all these dimensions. Now begin deleting the non-support vectors from your dataset and relearn the classifier. What do you observe? Does the margin increase or decrease? What will happen to the margin if the support vectors are removed from the dataset? Will the margin increase or decrease?

Solution: Code for this question is provided at the URL listed at the beginning of this document.

I selected a Gaussian kernel for use with SVM, and trained the classifier using the dual SVM stochastic gradient descent algorithm provided by Zaki and Meira. In addition, the SVM was trained with $C = 10$ and $\epsilon = .00001$. The SVM was trained and tested on datasets with input feature dimensions ranging from 2 to 32. Figure 1 shows accuracies obtained using hold-out cross-validation on datasets across this range of input feature dimensions. All datasets included 500 examples, with 67% of examples used for training and 33% used for testing.

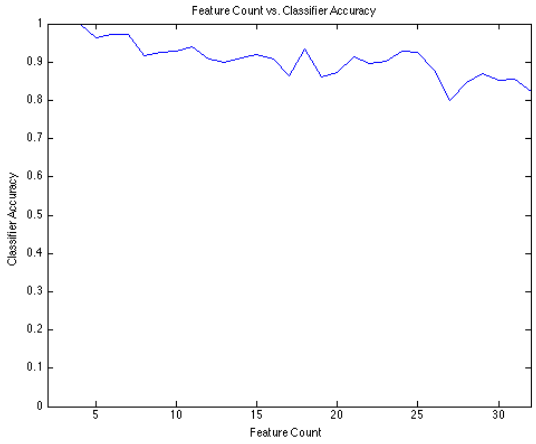

Figure 1: Input dimensionality vs. accuracy.

Overall, the classifier performs very well, with a mean accuracy of 91% across all dimensionalities. At the higher end of the range of input dimensionality, accuracy begins to drop, but never below 80%. With an input feature dimensionality of 32, a dataset with 500 examples can only represent .0001% of possible examples, in the best case scenario. Thus, even at this extreme end of the range, an accuracy of $> 80\%$ is very respectable.

The support vectors selected by the classifier are mostly the positive examples (those that pass the XOR test) on one side of the separating hyperplane, and the zero vector and points with a 'smaller' number of ones. Figure 2 characterizes the selected support vectors for a classifier where the input space is in $\mathbb{R}^8$. Here, most of the support vectors are those points with one one, and on the other side of the hyperplane we have support vectors with 0, 2, 3, and 4 ones. No points with > 4 ones are selected as support vectors.

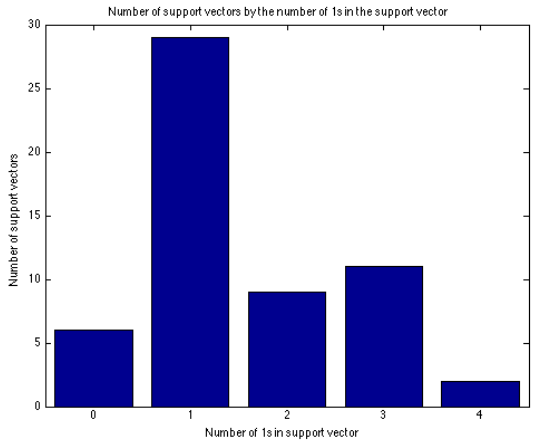

Figure 2: Support vector characteristics.

Removing both points that are support vectors and points that are not support vectors does little to affect the accuracy of the classifier, as Figures 3 and 4 demonstrate, respectively. As I have selected a kernel that produces an infinite feature space, it is impossible to reconstruct $\mathbf{w}$ in the input space, and thus impossible to determine the margin, per Zaki and Meira. In fact, it is even difficult to characterize the concept of a margin in infinite feature space.

My intuition tells me the following, however. As points that are not support vectors are removed from the dataset, the original support vectors still remain, and thus the margin should remain unchanged. On the other hand, as points that *are* support vectors are removed from the dataset, the SVM has a wider 'gap' across which to reach in order to find support vectors, and thus the margin should widen, in turn. Figure 4 reinforces this intuition, as when most support vectors are removed from the dataset, the accuracy of the classifier trends toward 100%.

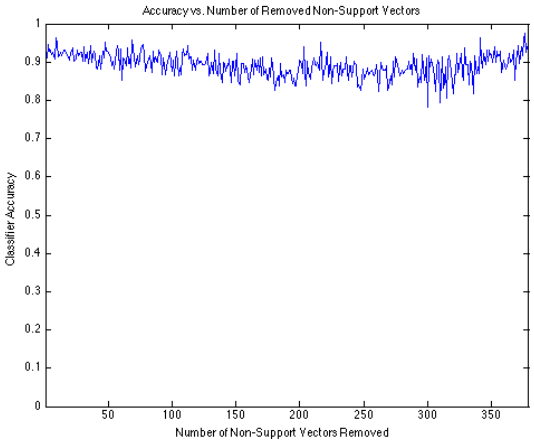

Figure 3: Effect of removing points that *are not* support vectors from the dataset.

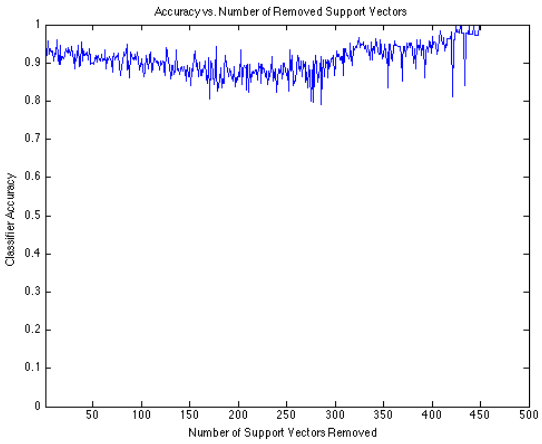

Figure 4: Effect of removing points that *are* support vectors from the dataset.