SVM Example

Dan Ventura

March 12, 2009

Abstract

We try to give a helpful simple example that demonstrates a linear SVM and then extend the example to a simple non-linear case to illustrate the use of mapping functions and kernels.

1 Introduction

Many learning models make use of the idea that any learning problem can be made easy with the right set of features. The trick, of course, is discovering that "right set of features", which in general is a very difficult thing to do. SVMs are another attempt at a model that does this. The idea behind SVMs is to make use of a (nonlinear) mapping function Φ that transforms data in input space to data in feature space in such a way as to render a problem linearly separable. The SVM then automatically discovers the optimal separating hyperplane (which, when mapped back into input space via Φ^{-1}, can be a complex decision surface). SVMs are rather interesting in that they enjoy both a sound theoretical basis as well as state-of-the-art success in real-world applications.

To illustrate the basic ideas, we will begin with a linear SVM (that is, a model that assumes the data is linearly separable). We will then expand the example to the nonlinear case to demonstrate the role of the mapping function Φ, and finally we will explain the idea of a kernel and how it allows SVMs to make use of high-dimensional feature spaces while remaining tractable.

2 Linear Example – when Φ is trivial

Suppose we are given the following positively labeled data points in $\Re^2$:

$$\left\{ \begin{pmatrix} 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ -1 \end{pmatrix}, \begin{pmatrix} 6 \\ 1 \end{pmatrix}, \begin{pmatrix} 6 \\ -1 \end{pmatrix} \right\}$$

and the following negatively labeled data points in $\Re^2$ (see Figure 1):

$$\left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \end{pmatrix} \right\}$$

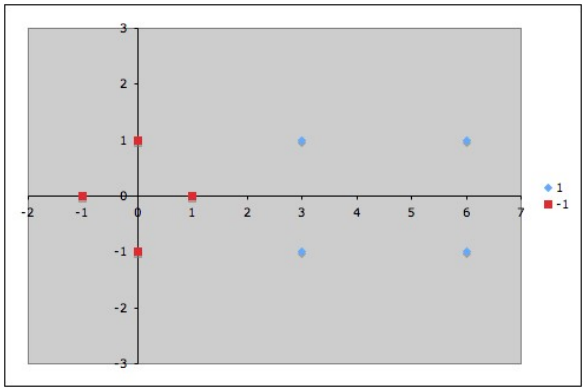

Figure 1: Sample data points in $\Re^2$. Blue diamonds are positive examples and red squares are negative examples.

We would like to discover a simple SVM that accurately discriminates the two classes. Since the data is linearly separable, we can use a linear SVM (that is, one whose mapping function $\Phi()$ is the identity function). By inspection, it should be obvious that there are three support vectors (see Figure 2):

$$\left\{ s_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, s_2 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}, s_3 = \begin{pmatrix} 3 \\ -1 \end{pmatrix} \right\}$$

In what follows we will use vectors augmented with a 1 as a bias input, and for clarity we will differentiate these with an over-tilde. So, if $s_1 = (10)$, then $\tilde{s_1} = (101)$. Figure 3 shows the SVM architecture, and our task is to find values for the α_i such that

$$\begin{aligned} \alpha_1\Phi(s_1)\cdot\Phi(s_1) + \alpha_2\Phi(s_2)\cdot\Phi(s_1) + \alpha_3\Phi(s_3)\cdot\Phi(s_1) &= -1 \\ \alpha_1\Phi(s_1)\cdot\Phi(s_2) + \alpha_2\Phi(s_2)\cdot\Phi(s_2) + \alpha_3\Phi(s_3)\cdot\Phi(s_2) &= +1 \\ \alpha_1\Phi(s_1)\cdot\Phi(s_3) + \alpha_2\Phi(s_2)\cdot\Phi(s_3) + \alpha_3\Phi(s_3)\cdot\Phi(s_3) &= +1 \end{aligned}$$

Since for now we have let $\Phi() = I$, this reduces to

$$\begin{aligned} \alpha_1\tilde{s_1}\cdot\tilde{s_1} + \alpha_2\tilde{s_2}\cdot\tilde{s_1} + \alpha_3\tilde{s_3}\cdot\tilde{s_1} &= -1 \\ \alpha_1\tilde{s_1}\cdot\tilde{s_2} + \alpha_2\tilde{s_2}\cdot\tilde{s_2} + \alpha_3\tilde{s_3}\cdot\tilde{s_2} &= +1 \\ \alpha_1\tilde{s_1}\cdot\tilde{s_3} + \alpha_2\tilde{s_2}\cdot\tilde{s_3} + \alpha_3\tilde{s_3}\cdot\tilde{s_3} &= +1 \end{aligned}$$

Now, computing the dot products results in

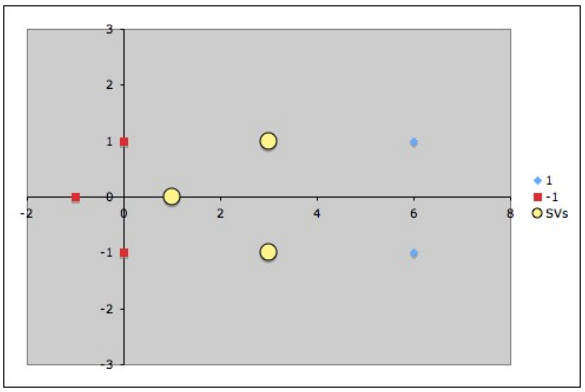

Figure 2: The three support vectors are marked as yellow circles.

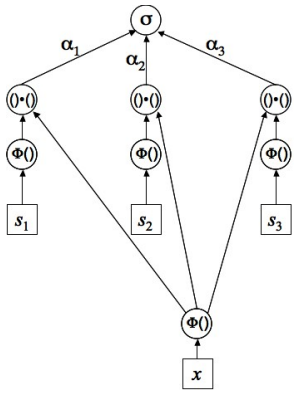

Figure 3: The SVM architecture.

$$\begin{aligned} 2\alpha_1 + 4\alpha_2 + 4\alpha_3 &= -1 \\ 4\alpha_1 + 11\alpha_2 + 9\alpha_3 &= +1 \\ 4\alpha_1 + 9\alpha_2 + 11\alpha_3 &= +1 \end{aligned}$$

A little algebra reveals that the solution to this system of equations is $\alpha_1 = -3.5, \alpha_2 = 0.75$ and $\alpha_3 = 0.75$.

Now, we can look at how these α values relate to the discriminating hyperplane; or, in other words, now that we have the α_i, how do we find the hyperplane that discriminates the positive from the negative examples? It turns out that

$$\begin{aligned} \tilde{w} &= \sum_i \alpha_i \tilde{s}_i \\ &= -3.5 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + 0.75 \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} + 0.75 \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix} \end{aligned}$$

Finally, remembering that our vectors are augmented with a bias, we can equate the last entry in $\tilde{w}$ as the hyperplane offset b and write the separating hyperplane equation $y = wx + b$ with $w = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $b = -2$. Plotting the line gives the expected decision surface (see Figure 4).

2.1 Input space vs. Feature space

3 Nonlinear Example – when Φ is non-trivial

Now suppose instead that we are given the following positively labeled data points in $\Re^2$:

$$\left\{ \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ -2 \end{pmatrix}, \begin{pmatrix} -2 \\ -2 \end{pmatrix}, \begin{pmatrix} -2 \\ 2 \end{pmatrix} \right\}$$

and the following negatively labeled data points in $\Re^2$ (see Figure 5):

$$\left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix} \right\}$$

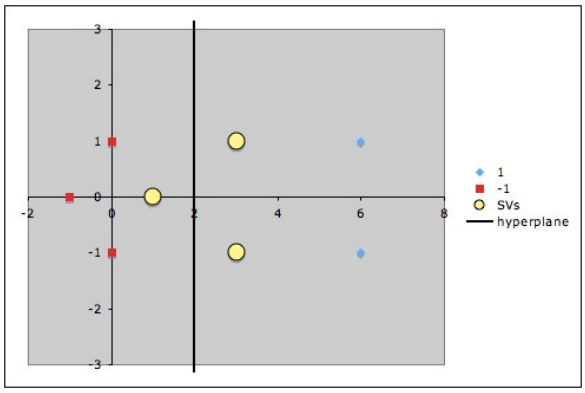

Figure 4: The discriminating hyperplane corresponding to the values $\alpha_1 = -3.5, \alpha_2 = 0.75$ and $\alpha_3 = 0.75$.

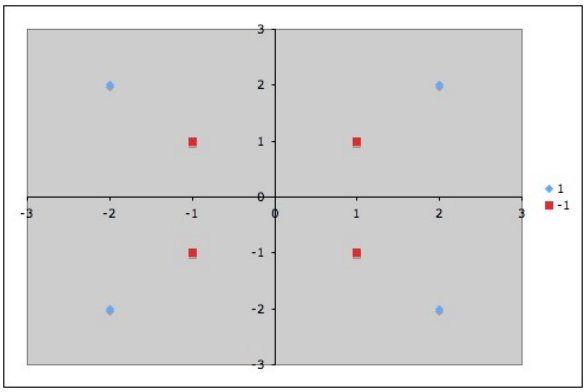

Figure 5: Nonlinearly separable sample data points in $\Re^2$. Blue diamonds are positive examples and red squares are negative examples.

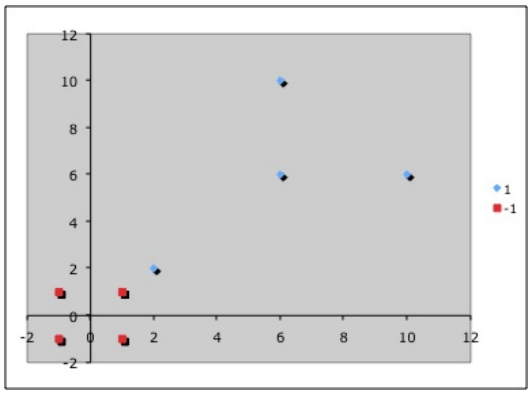

Figure 6: The data represented in feature space.

Our goal, again, is to discover a separating hyperplane that accurately discriminates the two classes. Of course, it is obvious that no such hyperplane exists in the input space (that is, in the space in which the original input data live). Therefore, we must use a nonlinear SVM (that is, one whose mapping function Φ is a nonlinear mapping from input space into some feature space). Define

$$\Phi_1 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{cases} \begin{pmatrix} 4 - x_2 + |x_1 - x_2| \\ 4 - x_1 + |x_1 - x_2| \end{pmatrix} & \text{if } \sqrt{x_1^2 + x_2^2} > 2 \\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} & \text{otherwise} \end{cases} \tag{1}$$

Referring back to Figure 3, we can see how Φ transforms our data before the dot products are performed. Therefore, we can rewrite the data in feature space as

$$\left\{ \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 6 \\ 2 \end{pmatrix}, \begin{pmatrix} 6 \\ 6 \end{pmatrix}, \begin{pmatrix} 2 \\ 6 \end{pmatrix} \right\}$$

for the positive examples and

$$\left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix} \right\}$$

for the negative examples (see Figure 6). Now we can once again easily identify the support vectors (see Figure 7):

$$\left\{ s_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, s_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \right\}$$

We again use vectors augmented with a 1 as a bias input and will differentiate them as before. Now given the [augmented] support vectors, we must again find values for the α_i. This time our constraints are

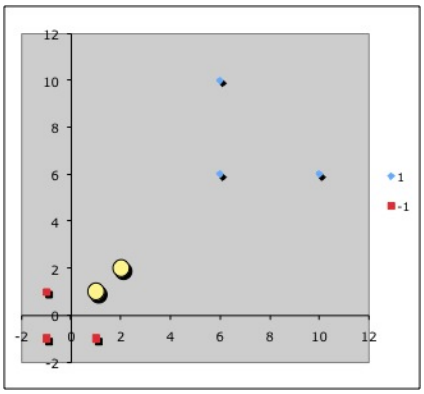

Figure 7: The two support vectors (in feature space) are marked as yellow circles.

$$\begin{aligned} \alpha_1 \Phi_1(s_1) \cdot \Phi_1(s_1) + \alpha_2 \Phi_1(s_2) \cdot \Phi_1(s_1) &= -1 \\ \alpha_1 \Phi_1(s_1) \cdot \Phi_1(s_2) + \alpha_2 \Phi_1(s_2) \cdot \Phi_1(s_2) &= +1 \end{aligned}$$

Given Eq. 1, this reduces to

$$\begin{aligned} \alpha_1 \tilde{s_1} \cdot \tilde{s_1} + \alpha_2 \tilde{s_2} \cdot \tilde{s_1} &= -1 \\ \alpha_1 \tilde{s_1} \cdot \tilde{s_2} + \alpha_2 \tilde{s_2} \cdot \tilde{s_2} &= +1 \end{aligned}$$

(Note that even though Φ_1 is a nontrivial function, both s_1 and s_2 map to themselves under Φ_1. This will not be the case for other inputs as we will see later.)

Now, computing the dot products results in

$$\begin{aligned} 3\alpha_1 + 5\alpha_2 &= -1 \\ 5\alpha_1 + 9\alpha_2 &= +1 \end{aligned}$$

And the solution to this system of equations is $\alpha_1 = -7$ and $\alpha_2 = 4$.

Finally, we can again look at the discriminating hyperplane in input space that corresponds to these α.

$$\tilde{w} = \sum_i \alpha_i \tilde{s_i}$$

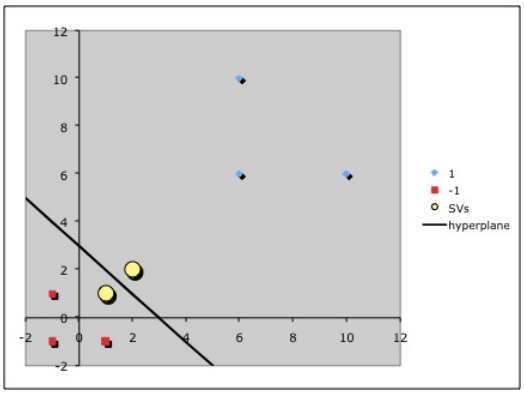

Figure 8: The discriminating hyperplane corresponding to the values $\alpha_1 = -7$ and $\alpha_2 = 4$

$$\begin{aligned} &= -7\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + 4\begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 \\ 1 \\ -3 \end{pmatrix} \end{aligned}$$

giving us the separating hyperplane equation $y = wx + b$ with $w = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $b = -3$. Plotting the line gives the expected decision surface (see Figure 8).

3.1 Using the SVM

Let's briefly look at how we would use the SVM model to classify data. Given x, the classification $f(x)$ is given by the equation

$$f(x) = \sigma\left(\sum_i \alpha_i \Phi(s_i) \cdot \Phi(x)\right) \tag{2}$$

where $\sigma(z)$ returns the sign of z. For example, if we wanted to classify the point $x = (4, 5)$ using the mapping function of Eq. 1,

$$\begin{aligned} f\begin{pmatrix} 4 \\ 5 \end{pmatrix} &= \sigma\left(-7\Phi_1\begin{pmatrix} 1 \\ 1 \end{pmatrix} \cdot \Phi_1\begin{pmatrix} 4 \\ 5 \end{pmatrix} + 4\Phi_1\begin{pmatrix} 2 \\ 2 \end{pmatrix} \cdot \Phi_1\begin{pmatrix} 4 \\ 5 \end{pmatrix}\right) \\ &= \sigma\left(-7\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + 4\begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}\right) \\ &= \sigma(-2) \end{aligned}$$

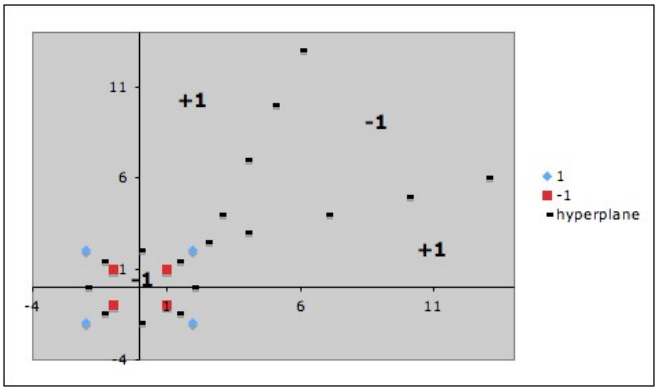

Figure 9: The decision surface in input space corresponding to Φ_1. Note the singularity.

and thus we would classify $x = (4, 5)$ as negative. Looking again at the input space, we might be tempted to think this is not a reasonable classification; however, it is what our model says, and our model is consistent with all the training data. As always, there are no guarantees on generalization accuracy, and if we are not happy about our generalization, the likely culprit is our choice of Φ. Indeed, if we map our discriminating hyperplane (which lives in feature space) back into input space, we can see the effective decision surface of our model (see Figure 9). Of course, we may or may not be able to improve generalization accuracy by choosing a different Φ; however, there is another reason to revisit our choice of mapping function.

4 The Kernel Trick

Our definition of Φ in Eq. 1 preserved the number of dimensions. In other words, our input and feature spaces are the same size. However, it is often the case that in order to effectively separate the data, we must use a feature space that is of (sometimes very much) higher dimension than our input space. Let us now consider an alternative mapping function

$$\Phi_2 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \frac{\left(x_1^2+x_2^2\right)-5}{3} \end{pmatrix} \tag{3}$$

which transforms our data from 2-dimensional input space to 3-dimensional feature space. Using this alternative mapping, the data in the new feature space looks like

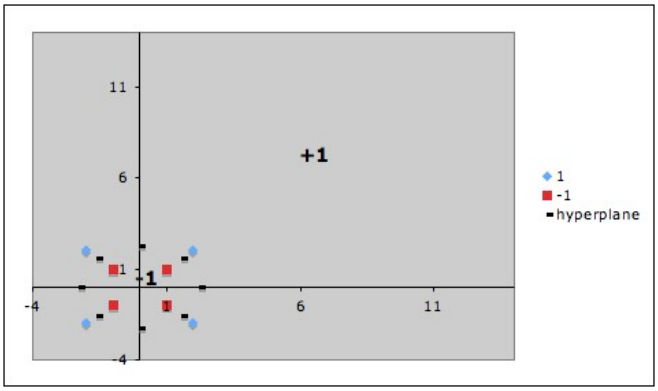

Figure 10: The decision surface in input space corresponding Φ_2.

$$\left\{ \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ -2 \\ 1 \end{pmatrix}, \begin{pmatrix} -2 \\ -2 \\ 1 \end{pmatrix}, \begin{pmatrix} -2 \\ 2 \\ 1 \end{pmatrix} \right\}$$

for the positive examples and

$$\left\{ \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix} \right\}$$

for the negative examples. With a little thought, we realize that in this case, all 8 of the examples will be support vectors with $\alpha_i = \frac{1}{46}$ for the positive support vectors and $\alpha_i = \frac{-7}{46}$ for the negative ones. Note that a consequence of this mapping is that we do not need to use augmented vectors (though it wouldn't hurt to do so) because the hyperplane in feature space goes through the origin, $y = wx+b$, where $w = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ and $b = 0$. Therefore, the discriminating feature, is x_3, and Eq. 2 reduces to $f(x) = \sigma(x_3)$.

Figure 10 shows the decision surface induced in the input space for this new mapping function.

Kernel trick.

5 Conclusion

What kernel to use? Slack variables. Theory. Generalization. Dual problem. QP.