CS 536 : Worked SVMs - Bias Values as the Thorn in my Side 16:198:536

Consider the following data set of two points: $(x_1, y_1) = (4, +1), (x_2, y_2) = (6, -1)$. Let's build a 1-D classifier that allows us to classify new values of x as members of the positive or negative class. Intuitively, if $x < 5$ it should be the positive class, if $x > 5$ it should be the negative class, and $x = 5$ should be the boundary between these two classes. This gives an overall 'best' classifier of:

$$\text{classify}(x) = \text{sign}(-x + 5). \tag{1}$$

How is this reflected/recoverable from the SVM? Recall that for the primal, we are trying to recover and weight value w and a bias value b such that they satisfy

$$\begin{aligned} &\min_{w,b} \ \frac{1}{2}w^2 \\ &\text{(s.t.)} \ (+1)(4w + b) \geq 1 \\ &\qquad\ \ (-1)(6w + b) \geq 1. \end{aligned} \tag{2}$$

One approach would be to try to manipulate the two constraint inequalities and try to get more explicit control over w, perhaps turn this into a one dimensional problem based on b instead of w. Alternately, we shift to look at the dual problem. The dual asks us to solve

$$\begin{aligned} &\max_{\alpha_1,\alpha_2} \ [\alpha_1 + \alpha_2] - \frac{1}{2}\left[\alpha_1^2(4 * 4) + \alpha_2^2(6*^6) - \alpha_1\alpha_2 4 * 6 - \alpha_2\alpha_1 6 * 4\right] \\ &\text{(s.t.)} \ \alpha_1 - \alpha_2 = 0 \\ &\qquad\ \ \alpha_1 \geq 0 \\ &\qquad\ \ \alpha_2 \geq 0. \end{aligned} \tag{3}$$

We can actually solve this in relatively short order. First, simplifying,

$$\begin{aligned} &\max_{\alpha_1,\alpha_2} \ [\alpha_1 + \alpha_2] - \frac{1}{2}\left[16\alpha_1^2 + 36\alpha_2^2 - 48\alpha_1\alpha_2\right] \\ &\text{(s.t.)} \ \alpha_1 = \alpha_2 \\ &\qquad\ \ \alpha_1 \geq 0 \\ &\qquad\ \ \alpha_2 \geq 0, \end{aligned} \tag{4}$$

we see that because $\alpha_1 = \alpha_2$, we can actually reduce this to a one dimensional problem,

$$\begin{aligned} &\max_{\alpha} \ 2\alpha - 2\alpha^2 \\ &\text{(s.t.)} \ \alpha \geq 0. \end{aligned} \tag{5}$$

In this case, the maximum occurs at $\alpha = \alpha_1 = \alpha_2 = 1/2$. In this usual way, this recovers the optimal weight as

$$w^* = \alpha_1(+1)(4) + \alpha_2(-1)(6) = (1/2)(4 - 6) = -1. \tag{6}$$

But what of the bias value?

To recover the bias value, we need to find one of the support vectors, where the inequality constraint in the primal will be satisfied with equality - this is a point that lies exactly on the margin. In this case, a support vector is one where $\alpha_i > 0$ (which is in fact both of our points). For such a point, we get that, in the general form,

$$y^i(\underline{w}.\underline{x}^i + b) = 1, \tag{7}$$

or (since $y^i = \pm 1, y^i * y^i = 1$),

$$\underline{w}.\underline{x}^i + b = y^i, \tag{8}$$

and

$$b = y^i - \underline{w}.\underline{x}^i. \tag{9}$$

In this case, taking the support vector $(x_1, y_1) = (4, +1)$, and the recovered weight value of -1, we can recover the bias directly as

$$b = (+1) - (-1)(4) = 1 + 4 = 5. \tag{10}$$

This completes the classifier as

$$\text{classify}(x) = \text{sign}(wx + b) = \text{sign}(-x + 5). \tag{11}$$

What if we had more points?

Suppose we expanded this so that the data set was $(x_1, y_1) = (4, +1)$, $(x_2, y_2) = (6, -1)$, and $(x_3, y_3) = (10, -1)$. Note, this new point is completely consistent with the classifier from the previous case, we expect that it will not be a support vector in the final classifier.

Writing out the primal in this case, we now want to solve for three α values (working in one dimension makes things easy to visualize but the fact that the dimension scales like the number of data points is increasingly frustrating). The dual becomes

$$\begin{aligned} \max_{\alpha_1,\alpha_2,\alpha_3} \ & [\alpha_1 + \alpha_2 + \alpha_3] - \frac{1}{2}\left[16\alpha_1^2 + 36\alpha_2^2 + 100\alpha_3^2 - 24\alpha_1\alpha_2 - 40\alpha_1\alpha_3 - 24\alpha_2\alpha_1 + 60\alpha_2\alpha_3 - 40\alpha_3\alpha_1 + 60\alpha_3\alpha_2\right] \\ \text{(s.t.)} \ & \alpha_1 - \alpha_2 - \alpha_3 = 0 \\ & \alpha_1 \geq 0, \alpha_2 \geq 0, \alpha_3 \geq 0, \end{aligned} \tag{12}$$

which simplifies somewhat to

$$\begin{aligned} \max_{\alpha_1,\alpha_2,\alpha_3} \ & [\alpha_1 + \alpha_2 + \alpha_3] - \frac{1}{2}\left[16\alpha_1^2 + 36\alpha_2^2 + 100\alpha_3^2 - 48\alpha_1\alpha_2 - 80\alpha_1\alpha_3 + 120\alpha_2\alpha_3\right] \\ \text{(s.t.)} \ & \alpha_1 - \alpha_2 - \alpha_3 = 0 \\ & \alpha_1 \geq 0, \alpha_2 \geq 0, \alpha_3 \geq 0. \end{aligned} \tag{13}$$

In this case, we see that the problem can be simplified by taking $\alpha_1 = \alpha_2 + \alpha_3$, which will reduce it from a three variable problem to a two variable problem:

$$\begin{aligned} \max_{\alpha_2,\alpha_3} \ & [\alpha_2 + \alpha_3 + \alpha_2 + \alpha_3] - \frac{1}{2}\left[16(\alpha_2^2 + 2\alpha_2\alpha_3 + \alpha_3^2) + 36\alpha_2^2 + 100\alpha_3^2 - (\alpha_2 + \alpha_3)(48\alpha_2 + 80\alpha_3) + 120\alpha_2\alpha_3\right] \\ \text{(s.t.)} \ & \alpha_2 \geq 0, \alpha_3 \geq 0, \end{aligned} \tag{14}$$

or

$$\begin{aligned} \max_{\alpha_2,\alpha_3} \ & 2\left[\alpha_2 + \alpha_3\right] - 2\left[\alpha_2^2 + 6\alpha_2\alpha_3 + 9\alpha_3^2\right] \\ \text{(s.t.)} \ & \alpha_2 \geq 0, \alpha_3 \geq 0, \end{aligned} \tag{15}$$

which simplifies even further to

$$\begin{aligned} \max_{\alpha_2,\alpha_3} \ & 2\left[\alpha_2 + \alpha_3\right] - 2(\alpha_2 + 3\alpha_3)^2 \\ \text{(s.t.)} \ & \alpha_2 \geq 0, \alpha_3 \geq 0, \end{aligned} \tag{16}$$

For any given value of α_3, we can imagine finding the α_2 that maximizes the objective function. In this case, taking the derivative for α_2 and setting it equal to zero yields a solution of $\alpha_2 = (1/2)(1-6\alpha_3)$, which simplifies (substituting that in for α_2) the problem finally to one variable:

$$\begin{aligned} &\max_{\alpha_3} \frac{1}{2} - 4\alpha_3 \\ &\text{(s.t.)}\ (1/2)(1-6\alpha_3) \geq 0, \alpha_3 \geq 0, \end{aligned} \tag{17}$$

Note, the additional constraint comes from the fact that we need $\alpha_2 \geq 0$.

However, this final form of the dual is clearly solved when $\alpha_3 = 0$, which gives a total solution of

$$\begin{aligned} &\alpha_3 = 0 \\ &\alpha_2 = (1/2)(1-6\alpha_3) = 1/2 \\ &\alpha_1 = \alpha_2 + \alpha_3 = 1/2. \end{aligned} \tag{18}$$

This recovers the exact same solution as before, a weight value of $w = (+1)(1/2)(4)+(-1)(1/2)(6)+(-1)(0)(10) = -1$, and noting that x_1 and x_2 are support vectors but x_3 is not, we have that the bias value we recover from the support vectors is exactly the same. Hence a final estimator of

$$\text{classify}(x) = \text{sign}(-x+5). \tag{19}$$

But what if the data points weren't separable?

Suppose we expanded this so that the data set was $(x_1, y_1) = (4, +1)$, $(x_2, y_2) = (6, -1)$, and $(x_3, y_3) = (10, +1)$. Note, in this case, the data is actually not separable with a straight linear classifier. What are we to do?

Well the thing we are to do is to extend it into a higher dimension through the use of a kernel. Consider for instance the Gaussian kernel,

$$K(x, x') = e^{\frac{1}{4}(x-x')^2}. \tag{20}$$

It's convenient to denote $K_{i,j}$ as $K(x_i, x_j)$. Setting up the dual in this case, we have

$$\begin{aligned} &\max_{\alpha_1,\alpha_2,\alpha_3} \ [\alpha_1+\alpha_2+\alpha_3] - \frac{1}{2}\left[\alpha_1^2 K_{1,1} - 2\alpha_1\alpha_2 K_{1,2} + 2\alpha_1\alpha_3 K_{1,3} + \alpha_2^2 K_{2,2} - 2\alpha_2\alpha_3 K_{2,3} + \alpha_3^2 K_{3,3}\right] \\ &\text{(s.t.)}\ \alpha_1 - \alpha_2 + \alpha_3 = 0 \\ &\qquad \alpha_1, \alpha_2, \alpha_3 \geq 0. \end{aligned} \tag{21}$$

As before, we see that $\alpha_2 = \alpha_1 + \alpha_3$, and this can be used to simplify the system over all:

$$\begin{aligned} &\max_{\alpha_1,\alpha_2,\alpha_3} \ 2\left[\alpha_1+\alpha_3\right] - \frac{1}{2}\left[\alpha_1^2 K_{1,1} - 2(\alpha_1+\alpha_3)(\alpha_1 K_{1,2} + \alpha_3 K_{2,3}) + 2\alpha_1\alpha_3 K_{1,3} + (\alpha_1+\alpha_3)^2 K_{2,2} + \alpha_3^2 K_{3,3}\right] \\ &\text{(s.t.)}\ \alpha_1, \alpha_3 \geq 0. \end{aligned} \tag{22}$$

Substituting all the kernel values in, this simplifies to

$$\begin{aligned} &\max_{\alpha_1,\alpha_3} \ 2(\alpha_1+\alpha_3) - \alpha_1^2\left(1-\frac{1}{e}\right) - \alpha_3^2\left(1-\frac{1}{e^4}\right) - \alpha_1\alpha_3\left(1+\frac{1}{e^9}-\frac{1}{e^4}-\frac{1}{e}\right) \\ &\text{(s.t.)}\ \alpha_1, \alpha_3 \geq 0. \end{aligned} \tag{23}$$

At this point, it is best to switch to numeric approximations, and the above can be solved explicitly for $\alpha_3 \approx 0.617797, \alpha_1 \approx 1.28197$, and these combine to give a final solutions of

$$\begin{aligned} &\alpha_1 \approx 1.28197 \\ &\alpha_2 \approx 1.89977 \\ &\alpha_3 \approx 0.617797 \end{aligned} \tag{24}$$

Note that each of these are non-zero, thus each represents a support vector. How to reconstruct the classifier from this? Recall that the 'kernelized' version of $\underline{w}.\underline{x}$ is now given by

$$\underline{w}.\underline{x} = \sum_i \alpha_i y_i K(\underline{x}^i, \underline{x}) \approx 1.28197 e^{-\frac{1}{4}(4-x)^2} - 1.89977 e^{-\frac{1}{4}(6-x)^2} + 0.617797 e^{-\frac{1}{4}(10-x)^2}. \tag{25}$$

We still need to determine the bias value, however. Choosing any one of the support vectors $\underline{x}^i, y^i$, we have in the usual way that

$$b = y^i - \underline{w}.\underline{x}^i \tag{26}$$

or in the case of $i = 2$ in this case, we have

$$b = (-1) - \left[1.28197 e^{-\frac{1}{4}(4-6)^2} - 1.89977 e^{-\frac{1}{4}(6-6)^2} + 0.617797 e^{-\frac{1}{4}(10-6)^2}\right] \approx 0.41684. \tag{27}$$

And in fact, it doesn't matter which support vector you pick, you get the same value of the bias from each. This gives us our final 'kernalized' $w.x$ and bias value of b, which combine to give our total classifier:

$$\text{classify}(x) = \text{sign}\left(1.28197 e^{-\frac{1}{4}(4-x)^2} - 1.89977 e^{-\frac{1}{4}(6-x)^2} + 0.617797 e^{-\frac{1}{4}(10-x)^2} + 0.41684\right). \tag{28}$$

And this correctly classifies each of the training data points. Additionally, the boundary between the classification regions is given by the solution to

$$1.28197 e^{-\frac{1}{4}(4-x)^2} - 1.89977 e^{-\frac{1}{4}(6-x)^2} + 0.617797 e^{-\frac{1}{4}(10-x)^2} + 0.41684 = 0, \tag{29}$$

which is approximately at $x \approx 5$ and $x \approx 8$.

The General SVM Procedure

For a general training data set, solving for the SVM proceeds in the following way:

- Compute the relevant kernel values
- Set up the Dual SVM Problem
- Solve for the values of $\alpha_1, \ldots, \alpha_m$
  - Either algebraically simplify the problem giving the constraints, as shown above
  - Or propose a solution $\underline{\alpha}$, and then attempt to tweak and modify it to improve the objective function
  - The SMO Algorithm: iteratively choose pairs α_i, α_j and determine the optimal values for α'_i, α'_j that maximize the objective function, holding the other $\underline{\alpha}$ constant, and maintaining the equality constraint and positivity constraints, $\alpha'_i, \alpha'_j \geq 0$
- Select a support vector, $\underline{x}^i, y^i$ for which $\alpha_i > 0$
- Solve for the bias value, b with

$$b = y^i - \sum_j \alpha_j y^j K(\underline{x}^j, \underline{x}^i) \tag{30}$$

- Construct the final classifier

$$\text{classify}(\underline{x}) = \text{sign}\left(\sum_i \alpha_i y^i K(\underline{x}^i, \underline{x}) + b\right) \tag{31}$$