CISC 371 Class 32

SVM – Soft Margins

Texts: [1] pp. 432–434

Main Concepts:

- *Mis-classified data vectors: false positives and false negatives*
- *Slack variables for inequality constraints*
- *Dual formulation of slack variables*

Sample Problem, Machine Inference: What is the optimal hyperplane for data that are not linearly separable data?

One common problem in data analysis is that data may be not be linearly separable. The simplest possible data are a list of real numbers, or 1D data. Suppose that we draw 10 numbers from a first Gaussian distribution and another 10 numbers from a second distribution, with the first set having Label -1 and the second set having Label +1. An example, in which the numbers are linearly separable, is shown in Figure 32.1(A). We can then add two more x_j values, labelled in a way that makes the data not linearly separable; this example is shown in Figure 32.1(B).

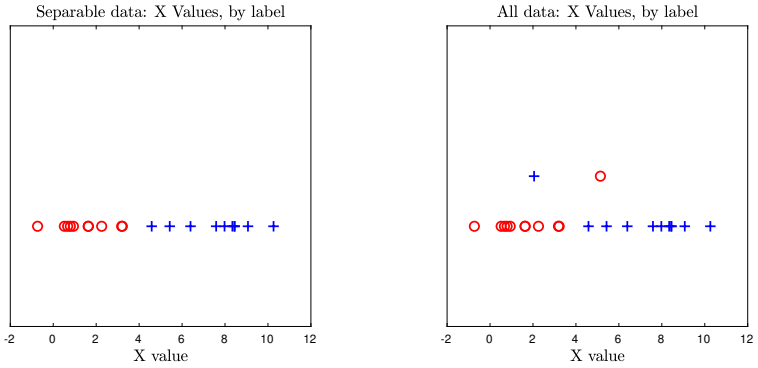


Figure 32.1: 1D data for linear separability, with Label -1 data shown as red circles and Label +1 data shown as blue crosses. (A) The data are linearly separable by a threshold of $\theta \approx 4$. (B) Additional data, which using a simple threshold would be classified as a false negative and a false positive, are not linearly separable.

We can use the data in Figure 32.1(A) to compute an SVM. This gives us a weight "vector", which is a scalar weight w, and a bias value b. For each input x_j, we can compute the score as

$$z_j = x_j w + b \tag{32.1}$$

We can then transform the inputs to scores. Applying Equation 32.1 to the 1D data, we find the scores of the linearly separable data as shown in Figure 32.2(A) and all of the data as shown in Figure 32.2(B). The data with scores less than 0 are mapped to Class -1 and the data with non-negative scores are mapped to Class +1. It is plain, from Figure 32.2(A) figure, that the support data are: the x_i value with the smallest positive score, which is the left-most blue circle; and x_i value with the largest negative score, which is the right-most red cross.

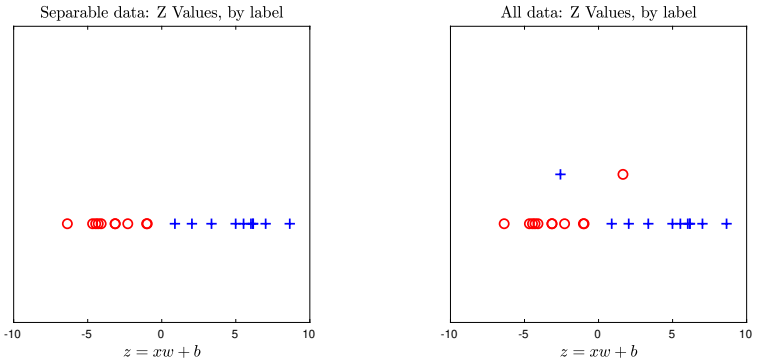


Figure 32.2: 1D data, written as x_j, are mapped to scores $z_j = x_j w + b$ for classification. Data with Label -1 data shown as red circles and Label +1 data shown as blue crosses. (A) The scores are linearly separable by a threshold of $\theta = 0$. (B) Additional data, which using a simple threshold would be classified as a false negative and a false positive, are not linearly separable.

Let us explore the false negative value in Figure 32.2(B) because it is a potential support vector for the SVM. Although this value has a Label +1, it has a score $u_j < 0$ that implies it should placed in Class -1. One way we can manage this false-negative score is to modify the equality constraint on a support vector. Recall that, if the natural number s indexes a support vector, the optimization requires that each support vector satisfies the equality constraint

$$\forall i \in \mathbb{N}_s : y_i(\vec{x}_i^T \vec{w} + b) = 1 \tag{32.2}$$

One way that we could modify this condition, to address the difficulty of the single false-negative value in our 1D data, is to add some *slack* to the left-hand side of Equation 32.2. Because

the score of this false-negative value is some $z_j < 0$, we could allow the score to increase by some positive number until the equality constraint of Equation 32.2 was met. It is common, in the literature describing an SVM, to use the Greek symbol ξ as the slack term. Our modified constraint on a support vector would be

$$\forall i \in \mathbb{N}_s : y_i(\vec{x}_i^T \vec{w} + b) + \xi_i = 1 \tag{32.3}$$

Adding this slack term raises other concerns, such as:

- how large ξ_s can be
- how many support vectors can have slack
- whether there is a limit on the total slack that is possible

32.1 Slack Variables: Primal Formulation

Let us recall how inequalities are defined in mathematics. In many constructions of the natural numbers and the real numbers from predicate logic, the inequality relation is defined in terms of equality. For example, for real numbers $a \in \mathbb{R}$ and $b \in \mathbb{R}$, we might say that $a \geq b$. The underlying definition is that there is some non-negative real number ξ such that $a + \xi = b$. The logical statement for this assertion is

$$(a \leq b) \rightarrow (\exists_{\xi \in \mathbb{R}_+} (a + \xi = b))$$

The name for ξ is a *slack variable*. We can use this concept to improve the inequality constraints in an SVM by introducing one slack variable for each inequality constraint and gathering these variables into a vector $\vec{\xi}$. In an SVM, we need to write a slack variable ξ_j with the constraint $\xi_j \geq 0$.

We arrived at the use of a slack variable by considering a specific instance of a potential support vector in a simple data set. We can generalize this idea so that it applies to every data vector $\vec{x}_j$. when we formulated the SVM problem, we extended the equality constraint of Equation 32.2 to be the inequality constraint

$$y_j(\vec{x}_j^T \vec{w} + b) \geq 1 \tag{32.4}$$

The addition of a slack variable ξ_j to Equation 32.4 would be

$$y_j(\vec{x}_j^T \vec{w} + b) + \xi_j \geq 1 \quad \text{with} \quad \xi_j \geq 0 \tag{32.5}$$

We prefer to write an inequality constraint as a level set, especially with respect to the level 0. We can do this by writing Equation 32.5 as

$$1 - y_j(\vec{x}_j^T \vec{w} + b) - \xi_j \leq 0 \quad \text{with} \quad \xi_j \geq 0 \tag{32.6}$$

The matrix-vector version of Equation 32.6 is a modification of Equation 30.5, in which we replace the "where" condition with an inequality constraint. We can write one set of constraints for the scores and another set for the non-negativity of the slack variables, so that the constraints are

$$\begin{array}{rcl} \vec{1} - YX\vec{w} - b\vec{y} - \vec{\xi} & \leq & \vec{0} \\ -\vec{\xi} & \leq & \vec{0} \end{array} \tag{32.7}$$

At this stage, there are two choices for how to incorporate the slack variables in the vector $\vec{\xi}$. If we include them as free variables, then there is a possibility that the optimization problem becomes infeasible because there may not be a value $\xi_j \geq 0$ for some slack variables.

The alternative is to use the slack variables as a *regularization* term. The common solution is to use a regularization argument $C > 0$ that is provided by the user. In practice, rather than regularizing the L_2 norm, which is the sum of squares of the slack variables ξ_j, the L_1 norm – sum of the absolute values of the slack variables – is used.

In the SVM literature, a slightly different formula is common. Because

$$(\xi_j \geq 0) \rightarrow (|\xi_j| = \xi_j)$$

we can write the L_1 norm of the slack variables as

$$\|\vec{\xi}\|_1 \stackrel{\text{def}}{=} \sum_{j=1}^{m} |\xi_j| = \sum_{j=1}^{m} \xi_j \tag{32.8}$$

The regularization term becomes the differentiable formula

$$C \sum_{j=1}^{m} \xi_j \tag{32.9}$$

We can differentiate Equation 32.9 with respect to the vector $\vec{\xi}$ by considering each entry. The derivative of Equation 32.9 with respect to ξ_j is the constant C, because the contribution of every other ξ_j is constant with respect to ξ_j. Gathering these terms as a vector, we have

$$\frac{\partial}{\partial \vec{\xi}} \left[C \sum_{j=1}^{m} \xi_j \right] = \begin{bmatrix} C \\ C \\ \vdots \\ C \end{bmatrix} = C\vec{1} \tag{32.10}$$

The objective function for the SVM is

$$f(\vec{w}) = \frac{1}{2}\vec{w}^T\vec{w} \tag{32.11}$$

The primal Lagrange function uses Equation 32.11 and Equation 32.7. We need to add two more terms: a Lagrange multiplier β_j for each slack variable ξ_j, and the regularization term of Equation 32.9. The slack version of the Lagrange function is

$$\mathcal{L}(\vec{w}, b, \vec{\xi}, \vec{\alpha}, \vec{\beta}) = \frac{1}{2}\vec{w}^T\vec{w} + \vec{\alpha}^T[\vec{1} - YX\vec{w} - b\vec{y} - \vec{\xi}] + C\sum_{i=1}^{m}\xi_j - \vec{\beta}^T\vec{\xi} \tag{32.12}$$

32.2 Slack Variables: Dual Formulation

The dual formulation of the slack primal function in Equation 32.12 is found by differentiating the primal function with respect to the variables $\vec{e}$, b, and $\vec{\xi}$, and then setting the transposes equal to zero for the KKT stationary point. The first two differentiations produce the same result as we found for the ordinary dual formulation. The third term, from Equation 32.12 and Equation 32.10, is

$$\frac{\partial\mathcal{L}}{\partial\vec{\xi}} = C\underline{1} - \vec{\alpha}^T - \vec{\beta}^T \tag{32.13}$$

Transposing Equation 32.13, setting it equal to zero, re-arranging terms, and again imposing the KKT conditions gives us

$$\vec{\alpha} + \vec{\beta} = C\vec{1} \quad \text{with} \quad \vec{\beta} \geq \vec{0} \tag{32.14}$$

An important observation is that we can use the concept of a slack variable *a second time*. We see that the Lagrange multipliers $\vec{\beta}$ in Equation 32.14 are acting as slack variables for the Lagrange multipliers $\vec{\alpha}$. We can therefore re-write Equation 32.14 as

$$\vec{\alpha} \leq C\vec{1} \tag{32.15}$$

Equation 32.15 gives us a second constraint on the original Lagrange multipliers $\vec{\alpha}$. Originally, the KKT conditions required that $\vec{\alpha} \geq \vec{0}$; Equation 32.15 now imposes $\vec{\alpha} \leq C\vec{1}$. We can combine these into a *box constraint* on the Lagrange multipliers, which is

$$\vec{0} \leq \vec{\alpha} \leq C\vec{1} \tag{32.16}$$

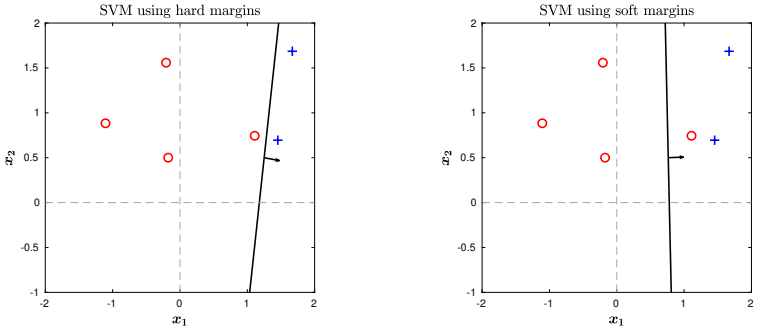
The dual formulation of the SVM problem with slack variables is therefore a single additional constraint on the Lagrange multipliers, so the formulation is

$$\begin{aligned} \mathcal{L}_D(\vec{\alpha}, b) &= -\frac{1}{2}\vec{\alpha}^T Y X X^T Y \vec{\alpha} + \vec{1}^T \vec{\alpha} \\ \text{with} \quad \vec{0} \leq \vec{\alpha} &\leq C\vec{1} \\ \vec{\alpha}^T \vec{y} &= 0 \end{aligned} \tag{32.17}$$

Equation 32.17 can be solved using the SMO algorithm, restricting $0 \leq \alpha_j \leq C$ at each step.

32.3 Example: Slack Variables for Simple 2D Data

A example of the use of slack variables is for simple 2D data, which are shown in Figure 32.3. As above, data with Label -1 are shown as red circles and data with Label +1 are shown as blue crosses. The ideal SVM hyperplane does not account for the possibility that some of the data may be labeled incorrectly. An alternative hyperplane allows for labeling errors. A single regularization term C changed the hyperplane in Figure 32.3(A) to the hyperplane in Figure 32.3(B). The slack variables, when constrained by a single regularization term, are also called *soft margins*.



(A) **(B)**

Figure 32.3: 2D data for linear separability, with Class -1 data shown as red circles and Class +1 data shown as blue crosses. (A) Strict linear separability of the data in which the hyperplane is shown as a black line. (B) Soft linear separability allows for incorrectly labeled data.

References

[1] Hastie T, Tibshirani R, Friedman J: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, 2009