CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Support Vector Machines—Non-Separable Classification and Regression

Lecturer: Michael I. Jordan *Scribe: Vito Dai*

1 On Homework...

In the homework set, we solve a classification problem by explicitly finding the transformation $\Phi(x)$ that maps an ellpitical kernel in $\Re^2$ to a inner product in $\Re^3$, then running logistic regression in this higher dimension.

$$\Phi(x) = \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix} \tag{1}$$

There are several problems with this approach. First while working explictly in $\Re^3$ might be reasonable, the number of dimensions grows quickly as we allow more degrees of freedom in the kernel, e.g. allowing translations of the ellipse center. Moreover, the transformation $\Phi(x)$ is not one-to-one. There are also overfit problems associated with looking at the data, then arbitrarily choosing a kernel that fits the data.

The key to the success of SVM is the kernel trick, where we avoid computations in the higher dimensional space by computing only inner products there, which translates into evaluation of a kernel function in the original low dimensional space. In other words, we try to take advantage of the fact that $\Phi(x_1)^T\Phi(x_2) = k(x_1, x_2)$.

2 SVM Non-Separable Classification

To implement the optimal margin classifier when the data are not separable, we introduce slack variables ξ_i that allow data to cross margin boundaries.

$$y_i(x_i^T w + b) \geq 1 - \xi_i \qquad \xi_i \geq 0 \tag{2}$$

When $\xi_i = 0$, the data lie on or outside margin boundaries. $0 < \xi_i \leq 1$ means the data lie inside the margin, but on the correct side of the classifying boundary. $\xi_i > 1$ means the data crossed the classification boundary, and could be considered a classification error on the training data. These properties of ξ_i are summarized in Figure 1. The sum of ξ_i over all i is an upper bound on the training classification errors, which we would like to minimize. Thus the non-separable optimal margin primal problem can be expressed as follows:

$$\begin{aligned} min \qquad & \frac{1}{2}w^T w + \frac{C}{m}\sum_{i=1}^{m} \xi_i \\ s.t. \qquad & y_i(x_i^T w + b) - 1 + \xi_i \geq 0 \end{aligned} \tag{3}$$

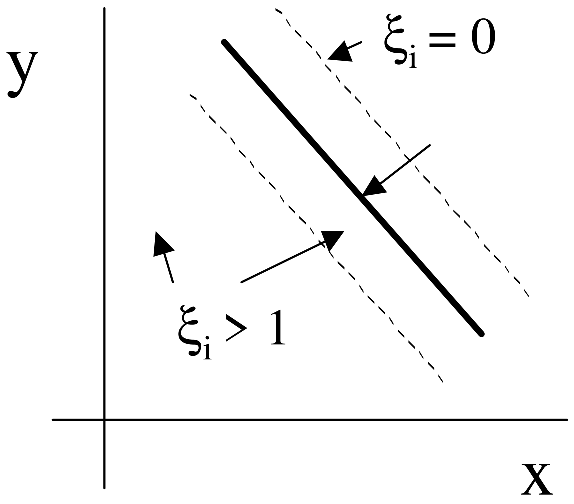

Figure 1: Properties of slack variables ξ_i

To form the dual problem we write down the Lagrangian,

$$\mathcal{L} = \frac{1}{2}w^T w + \frac{C}{m}\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i\left[y_i(x_i^T w + b) - 1 + \xi_i\right] - \sum_{i=1}^{m}\mu_i\xi_i \tag{4}$$

and solve by setting the gradient of the primal variables w, b, ξ to zero.

$$\frac{\delta\mathcal{L}}{\delta w} = 0 \quad\Rightarrow\quad w = \sum_i \alpha_i y_i x_i \tag{5}$$

$$\frac{\delta\mathcal{L}}{\delta b} = 0 \quad\Rightarrow\quad \sum_i \alpha_i y_i = 0 \tag{6}$$

$$\frac{\delta\mathcal{L}}{\delta\xi_i} = 0 \quad\Rightarrow\quad \alpha_i = \frac{C}{m} - \mu_i \tag{7}$$

As in the separable case, the optimal vector w is a linear combination of the data points. Substituting these values back into (4), we can write down the dual problem:

$$D : \Theta(\alpha) = \sum_i \alpha_i - \frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j x_i^T x_j \tag{8}$$

Note that this equation is exactly the same as the dual problem for separable case because the μ_i have been cancelled out. The main difference lies in the constraints on α_i and μ_i.

$$\mu_i \geq 0 \tag{9}$$

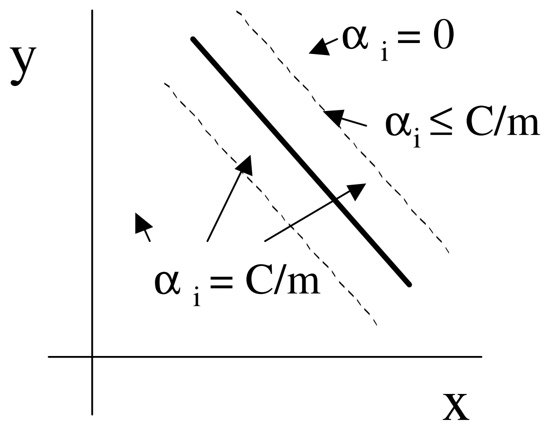

Figure 2: Property of support vector coefficients α_i

$$0 \leq \alpha_i \leq \frac{C}{m} \qquad \textit{from } (7) \textit{ and } (9) \tag{10}$$
$$\alpha_i(y_i(x_i^T w + b) - 1 + \xi_i) \geq 0 \tag{11}$$
$$\mu_i \xi_i = 0 \tag{12}$$

The box constraint on α_i in (10) is new, though not any more difficult to solve using quadratic programming. As before, when the data lie strictly outside margin boundaries, $\alpha_i = 0$. These data points are not included in the set of support vectors that define margin boundaries. When the data lie on the margin boundary, we have $0 \leq \alpha_i \leq \frac{C}{m}$. These are considered support vectors as before, but the maximum contribution is bounded by the fraction $\frac{C}{m}$. Finally, for points that lie inside the boundary, and on the wrong side of the classifier, $\alpha_i = \frac{C}{m}$. All such points are also support vectors, and they contribute a maximum weight of $\frac{C}{m}$ no matter how far away they are from the margin boundary. All these properties of α_i are summarized in Figure 2.

3 Sequential Minimal Optimization (SMO)

To solve for the α_i requires quadratic programming techniques. However, instead of optimizing over all α_i simultaneously, it is possible to do iterative pairwise optimizations over α_i and α_j while keeping all other $\alpha_k, k \neq i, k \neq j$ constant. This is called sequential minimal optimization and it works because the optimization is over a simple quadratic surface, so climbing uphill locally is guaranteed to climb uphill globally as well. Doing pairwise optimization is extremely attractive because we can solve for α_i and α_j analytically:

$$case1 : y_i = y_j \qquad \alpha_i + \alpha_j = \gamma(\alpha_k : k \neq i, k \neq j) \tag{13}$$
$$case2 : y_i \neq y_j \qquad \alpha_i - \alpha_j = \gamma(\alpha_k : k \neq i, k \neq j) \tag{14}$$

If the pairwise optimium lies outside our box constraint, then one of the α is clipped to zero. The difficulty of the SMO technique lies in the picking of α_i and α_j; we need to pick them such that the optimization does

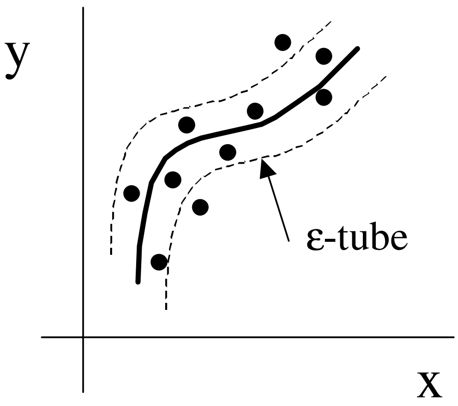

Figure 3: SVM regression with ϵ tube margin

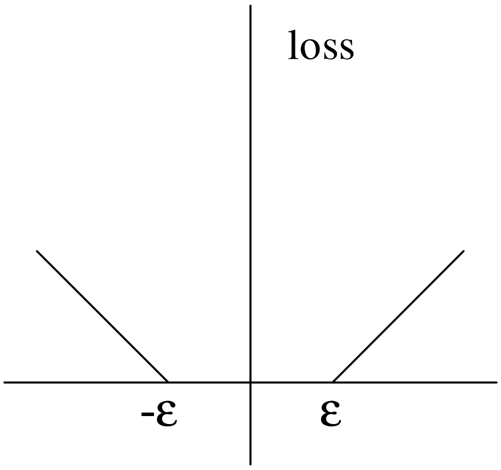

Figure 4: ϵ sensitive loss

move us. The first heuristic is to choose α_i that violates the KKT conditions, then chose α_j that maximizes the uphill climb in relation to this α_i. The SMO technique, in fact, works very well in practice.

4 SVM Regression

The SVM framework can also be adapted to perform regression. Data are allowed to lie within an ϵ-tube of the kernel without penalty by defining two slack variables ξ_i and ξ_i^* for the lower and upper boundary of the tube respectively as shown in Figure 3 and Figure 4. The primal SVM regression problem can be stated as follows:

$$min\frac{1}{2}w^T w + C\sum_i(\xi_i + \xi_i^*) \quad \xi_i > 0, \xi_i^* > 0 \tag{15}$$

$$
\begin{aligned}
s.t. y_i - w^T x_i - b &\leq \epsilon + \xi_i && (16) \\
w^T x_i + b - y_i &\leq \epsilon + \xi_i^* && (17)
\end{aligned}
$$

The dual problem is formed and solved as before.