Support Vector Machines

Carlos C. Rodríguez
http://omega.albany.edu:8008/

October 17, 2004

Finding the Hyperplane with Largest Margin

Let us assume that we have n labeled examples $(x_1, y_1), \ldots, (x_n, y_n)$ with labels $y_i \in \{1, -1\}$. We want to find the hyperplane $< w, x > +b = 0$ (i.e. with parameters (w, b)) satisfying the followin three conditions:

1. The scale of (w, b) is fixed so that the plane is in canonical position w.r.t. $\{x_1, \ldots, x_n\}$. i.e.,
$$\min_{i \leq n} | < w, x_i > +b| = 1$$

2. The plane with parameters (w, b) separates the $+1$'s from the the -1's. i.e.,
$$y_i(< w, x_i > +b) \geq 0 \text{ for all } i \leq n$$

3. The plane has maximum margin $\rho = 1/|w|$. i.e., minimum $|w|^2$.

Of course there may not be a separating plane for the observed data. Let us assume, for the time being, that the data is in fact linearly separable and we'll take care of the general (more realistic) case later.

Clearly 1 and 2 combine into just one condition:

$$y_i(< w, x_i > +b) \geq 1 \text{ for all } i \leq n.$$

Thus, we want to solve the following optimization problem,

$$\text{minimize } \frac{1}{2}|w|^2$$

over all $w \in R^d$ and $b \in R$ subject to,

$$y_i(< w, x_i > +b) - 1 \geq 0 \text{ for all } i \leq n.$$

This is a very simple quadratic programming problem. There are readily available algorithms of complexity $O(n^3)$ that can be used for solving this problem. For example the so called interior point algorithms that are variations of

the Karmarkar algorithm for linear programming will do. But, when n and d are large (tens of thousands) even the best QP methods will fail. A very desirable characteristics of SVMs is that most of the data ends up being irrelevant. The relevant data are only the points that end up exactly on the margin of the optimal classifier and these are often a very small fraction of n.

KKT-Theory

The problem that we need to solve is a special case of the general problem of minimizing a convex function $f(x)$ subject to n inequality constraints $g_j(x) \geq 0$ for $j = 1, 2, \ldots, n$ where the functions g_j are also convex. Let's call this problem (CO). Notice that in our case $x = (w, b) \in R^{d+1}$ and the constraints are linear in the unknowns x. Don't get confused with our previous x_i's.

The characterization of the solution to the convex optimization (CO) problem is given by the so called Karush-Kuhn-Tucker conditions.

Theorem (KKT-Conditions)

$$\bar{x} \text{ solves the (CO) problem}$$

if and only if there exists

$$\bar{\lambda} = (\bar{\lambda}_1, \ldots, \bar{\lambda}_n) \geq 0$$

a vector of non-negative Lagrange multipliers so that

$$(\bar{x}, \bar{\lambda}) \text{ is a saddle point of the Lagrangean,}$$

$$L(x, \lambda) = f(x) - \sum_{j=1}^{n} \lambda_j g_j(x).$$

i.e., for all x and for all $\lambda \geq 0$ we have,

$$L(\bar{x}, \lambda) \leq L(\bar{x}, \bar{\lambda}) \leq L(x, \bar{\lambda}).$$

Before proving (half of) this theorem notice that there is an easy to understand intuitive reason behind this result. Think of the term added (subtracted actually) to $f(x)$ to form the Lagrangean L, as a penalty for an x that violates the constraints. In fact, if $g_j(x) < 0$, the term $-\lambda_j g_j(x) > 0$ can be made arbitrarily large by increasing λ_j. Thus, the minimizer of $L(x, \lambda)$ over x will have to make $g_j(x) \geq 0$. On the other hand if $g_j(x) > 0$ then it is best to take $\lambda_j = 0$ to maximize $L(x, \lambda)$ as a function of λ. It is possible to show that, the saddle point condition is equivalent to,

$$\max_x \min_{\lambda \geq 0} L(x, \lambda) = L(\bar{x}, \bar{\lambda}) = \min_{\lambda \geq 0} \max_x L(x, \lambda).$$

Proof: Let us show that the saddle point condition is in fact sufficient for solving the (CO) problem. That it is also necessary depends on Farkas's Lemma and it is much more difficult to prove. We need to show that the saddle point condition implies,

1. for all $j \leq n$,
$$g_j(\bar{x}) \geq 0$$
and,

2. for all x that satisfies the constraints,
$$f(\bar{x}) \leq f(x)$$

To show 1, suppose that the ith constraint is violated. Then by taking
$$\lambda_i > \bar{\lambda}_i$$
and
$$\lambda_j = \bar{\lambda}_j \text{ for all } j \neq i$$
we get,
$$L(\bar{x}, \lambda) > L(\bar{x}, \bar{\lambda})$$
which contradicts the saddle point condition.

To show 2, take $\lambda = 0$ on the left hand side of the saddle point condition and take x satisfying the constraints on the right. Then,
$$f(\bar{x}) = L(\bar{x}, 0) \leq L(x, \bar{\lambda}) \leq f(x).$$
Which proves 2. •

When all, the objective function $f(x)$, and the constrainning functions $g_j(x)$ are differentiable (they are infinitely differentiable in the case of SVMs) the condition for a saddle point is simply that at that point the tangent plane to the surface $z = L(x, \lambda)$ is parallel to the (x, λ) plane. The saddle point of L can be obtained by solving the system of equations,
$$\begin{aligned} \nabla_x L(x, \lambda) &= 0, \text{i.e., } \nabla f(x) = \sum_j \lambda_j \nabla g_j(x) \\ \textit{and} \qquad & \lambda_j g_j(x) = 0 \text{ for all } j \leq n \text{from where, } \nabla_\lambda L(x, \lambda) = 0. \end{aligned}$$
The second set of equations are known as *complementarity conditions* and are a consequence of the constraint that $\lambda \geq 0$.

The Dual

The min max = max min characterization of the saddle point of the Lagrangean L provides an alternative way to find the solution of the (CO) problem. Instead of minimizing $f(x)$ subject to the $g_j(x) \geq 0$ constraints one can equivalently maximize $W(\lambda)$, where
$$W(\lambda) = \min_x L(x, \lambda)$$
subject to the constraint that $\lambda \geq 0$. This provides an alternative route to the same saddle point of L.

The Support Vectors of SVMs

Let us apply the KKT-conditions to our original problem of finding the separating hyperplane with maximum margin. The Lagrangean in this case is,

$$L(w, b, \lambda) = \frac{1}{2} \sum_{i=1}^{d} w_i^2 - \sum_{j=1}^{n} \lambda_j \left\{ y_j (< w, x_j > + b) - 1 \right\},$$

and the KKT-conditions for optimality are,

$$\begin{aligned} \nabla_w L &= 0, \text{i.e., } w = \sum_{j=1}^{n} \lambda_j y_j x_j \\ \nabla_b L &= 0, \text{i.e., } \sum_{j=1}^{n} \lambda_j y_j = 0 \\ \lambda_j \left\{ y_j (< w, x_j > + b) - 1 \right\} &= 0, \text{ for all } j \leq n. \end{aligned}$$

These provide a complete characterization of the optimal plane. The normal w must be a linear combination of the observed vectors x_j, that's the first set of equations. The coefficients of this linear combination must add up to 0, that's the second equation. Finally the complementarity conditions tell us that the only non-zero Lagrange multipliers λ_j are those associated to the vectors x_j right on the margin, i.e., such that,

$$y_j (< w, x_j > + b) = 1.$$

These are called *support vectors* and they are the only ones needed since

$$w = \sum_{j \in J_0} \lambda_j y_j x_j$$

where $J_0 = \{j : x_j \text{ is a s.v.}\}$. The support vectors are the observations x_j at the exact distance $\rho = 1/|w|$ from the separating plane. The number of such vectors is usually much smaller than n and that makes it possible to consider very large numbers of examples with x_j having many coordinates.

The Dual Problem for SVMs

The dual problem for SVMs turns out to be even simpler than the primal and its formulation shows the way to a magnificent non-linear generalization. For a given vector λ of Lagrange multipliers, the minimizer of $L(w, b, \lambda)$ w.r.t. (w, b) must satisfy the optimality conditions obtained above, i.e., w is a l.c. of the x_j's with coefficients $\lambda_j y_j$ that must add up to zero. Hence, replacing these conditions into $L(w, b, \lambda)$ we obtain the dual formulation,

$$\text{maximize } W(\lambda)$$

where,

$$W(\lambda) = \sum_{j=1}^{n} \lambda_j - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j < x_i, x_j >$$

and the $\lambda \geq 0$ satisfying,

$$\sum_{j=1}^{n} \lambda_j y_j = 0.$$

Maximizing $W(\lambda)$ over $\lambda \geq 0$ s.t. the above simple linear constraint is satisfied, is the preferred form to feed a QP algorithm. Once an optimal λ is obtained we find w as the l.c. of the x_j as above and we find b by recalling that the plane must be in canonical position so,

$$\min_{i \leq n} y_i(< w, x_i > +b) = 1 = y_j(< w, x_j > +b) \text{ for all } j \in J_0$$

and we get,

$$b = y_j - < w, x_j > .$$

Multiplying through by λ_j and adding over j we find,

$$b = \frac{-\sum_{i,j} \lambda_i \lambda_j y_i y_j < x_i, x_j >}{\sum_j \lambda_j}$$

and it can be readily checked that this value coincides with the value of the Lagrange multiplier β associated to the constraint $\sum_i \lambda_i y_i = 0$ (just find $\nabla_\lambda \mathcal{L} = 0$ for the Lagrangean associated to the dual, i.e. $\mathcal{L} = W(\lambda) - \beta \sum_i \lambda_i y_i$). The optimal values of β and of λ are often returned by the modern QP solvers based on interior point algorithms.