Soft-Margin Linear SVM

David S. Rosenberg

09-02-2015

1 Formulating SVM as a QP

Here we consider the "soft-margin linear SVM", which is the linear predictor that minimizes the SVM or "hinge" loss subject to an ℓ_2-regularization penalty:

$$\min_{w \in \mathbf{R}^d} \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n}(1 - y_i w^T x_i)_+. \tag{1.1}$$

Rather than the typical λ regularization parameter attached to the ℓ_2 penalty, for SVMs it's traditional to have a "c" parameter attached to the empirical risk component. The larger c is, the more we care about maximizing the margin, compared to finding a "simple" hypothesis w with small ℓ_2-norm.

This is an unconstrained optimization problem, which is nice. But the objective function is not differentiable, which makes it difficult to work with. So we formulate an equivalent problem with a differentiable objective, though we'll have to add new constaints to do so. Note that 1.1is trivially equivalent to

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} & \xi_i = \left(1 - y_i w^T x_i\right)_+ . \end{array}$$

It is clear after a moment's thought that this is equivalent to

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} & \xi_i \geq 0 \text{ for } i = 1, \ldots, n \\ & \xi_i \geq \left(1 - y_i w^T x_i\right) \text{ for } i = 1, \ldots, n \end{array}$$

We now have a differentiable objective function with $2n$ affine constraints. This is a quadratic program that can be solved by any off-the-shelf QP solver.

Note that even though this last formulation does not directly define ξ_i to be the hinge loss for example i, because of the minimization, the optimal ξ_i^* will exactly be the loss on example i.

2 Compute the Lagrangian Dual

The Lagrangian for this formulation is

$$\begin{aligned}L(w,\xi,\alpha,\lambda) =& \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1 - y_i w^T x_i - \xi_i) - \sum_i \lambda_i \xi_i \\ =& \frac{1}{2}w^T w + \sum \xi_i\left(\frac{c}{n} - \alpha_i - \lambda_i\right) + \sum_{i=1}^{n}\alpha_i(1 - y_i w^T x_i)\end{aligned}$$

From our study of Lagrangian duality, we know that the original problem can now be expressed as

$$\inf_{w,\xi}\sup_{\alpha,\lambda\succeq 0} L(w,\xi,\alpha,\lambda).$$

Since our constraints are affine, by Slater's condition we have strong duality so long as the problem is feasible (i.e. so long as there is at least one point in the feasible set). The constraints are satisfied by $w = 0$ and $\xi_i = 1$ for $i = 1,\ldots,n$, so **we have strong duality**. Thus we get the same result if we solve the following dual problem:

$$\sup_{\alpha,\lambda\succeq 0}\inf_{w,\xi} L(w,\xi,\alpha,\lambda).$$

As usual, we capture the inner optimization in the Lagrange dual objective: $g(\alpha,\lambda) = \inf_{w,\xi} L(w,\xi,\alpha,\lambda)$.

Note that if $\frac{c}{n} - \alpha_i - \lambda_i \neq 0$, then the Lagrangian is unbounded below (by taking $\xi_i \to \pm\infty$) and thus the infimum is $-\infty$. For any given (α,λ), the function $(w,\xi) \mapsto L(w,\xi,\alpha,\lambda)$ is convex and differentiable, thus we have an optimal point if and only if $\partial_w L = \partial_\xi L = 0$ at that point. That is:

$$\begin{aligned}\partial_w L = 0 \quad &\iff \quad w - \sum_{i=1}^{n}\alpha_i y_i x_i = 0 \iff w = \sum_{i=1}^{n}\alpha_i y_i x_i \qquad (2.1)\\ \partial_\xi L = 0 \quad &\iff \quad \frac{c}{n} - \alpha_i - \lambda_i = 0 \iff \alpha_i + \lambda_i = \frac{c}{n}.\end{aligned}$$

Note that one of the conditions is $\alpha_i + \lambda_i = \frac{c}{n}$. We previously noted that when this does not hold, the infimum is $-\infty$ (i.e. there is no minimum – so these observations are consistent). Substituting these conditions back into L, the second term disappears, while the first and third terms become

$$\begin{aligned}\frac{1}{2}w^T w &= \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j \\ \sum_{i=1}^{n}\alpha_i(1 - y_i w^T x_i) &= \sum_{i=1}^{n}\alpha_i - \sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_j^T x_i.\end{aligned}$$

Putting it together, the dual function is

$$g(\alpha,\lambda) = \begin{cases}\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_j^T x_i & \frac{c}{n} - \alpha_i - \lambda_i = 0\\ -\infty & \text{otherwise.}\end{cases}$$

Thus we can write the dual problem as

$$\begin{aligned}&\sup_{\alpha,\lambda}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_j^T x_i\\ &\text{s.t.}\alpha_i+\lambda_i=\frac{c}{n}\\ &\quad\alpha_i\geq 0\qquad\lambda_i\geq 0.\end{aligned}$$

We can actually eliminate the λ variables, replacing the three constraints by $0 \leq \alpha_i \leq \frac{c}{n}$:

$$\begin{aligned}&\sup_{\alpha}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_j^T x_i\\ &\text{s.t.}\alpha_i\in[0,\frac{c}{n}].\end{aligned}$$

When written in standard form, this has a quadratic objective in n unknowns and $2n$ constraints. Note that these constraints have a particularly simple form: they are called **box constraints**. If α^* is a solution to the dual problem, then by strong duality and 2.1, the optimal solution to the primal problem is given by

$$w^*=\sum_{i=1}^{n}\alpha_i^* y_i x_i.$$

Since $\alpha_i \in [0, \frac{c}{n}]$, we see that c controls the amount of weight we can put on any single example.

2.1 Consequences of Complementary Slackness

Since we have strong duality, we have the following **complementary slackness** conditions:

$$\alpha_i^*(1-y_i\left(w^*\right)^T x_i-\xi_i^*)=0 \tag{2.2}$$

$$\lambda_i^*\xi_i^*=\left(\frac{c}{n}-\alpha_i^*\right)\xi_i^*=0 \tag{2.3}$$

For convenience, let's write $f^*(x) = \left(w^*\right)^T x$. As we noted above, ξ_i^* is the hinge loss on example i. When $\xi_i^* = 0$, we're either "at the margin" (i.e. $y_i f^*(x_i) = 1$ or on the good side of the margin ($y_i f^*(x_i) > 1$).

Note that $w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$ only depends on those examples for which $\alpha_i^* > 0$ (recall that $\alpha_i^* \geq 0$ by constraint). These examples are called **support vectors**. By the complementary slackness condition, $\alpha_i^* > 0$ implies $1 - y_i f^*(x_i) - \xi_i^* = 0$. So $y_i f^*(x_i) = 1 - \xi_i^* \leq 1$, where the last inequality follows from $\xi_i^* \geq 0$. So the margin is at most 1.

If $y_i f^*(x_i) < 1$ then the margin loss $\xi_i^* > 0$, which implies that $\alpha_i^* = \frac{c}{n}$ (by 2.3).

If $\alpha_i^* = 0$ then $\xi_i^* = 0$, so $y_i f^*(x_i) \geq 1$.

$\alpha_i^* \in (0, c/n)$ implies $\xi_i^* = 0$, by 2.3. Then by 2.2, $y_i f^*(x_i) = 1$. So the prediction is right on the margin.

Summary table:

$$\begin{aligned} \alpha_i^* = 0 \quad &\implies \quad y_i f^*(x_i) \geq 1 \\ \alpha_i^* \in \left(0, \frac{c}{n}\right) \quad &\implies \quad y_i f^*(x_i) = 1 \\ \alpha_i^* = \frac{c}{n} \quad &\implies \quad y_i f^*(x_i) \leq 1 \end{aligned}$$

Conversely

$$\begin{aligned} y_i f^*(x_i) < 1 \quad &\implies \quad \alpha_i^* = \frac{c}{n} \\ y_i f^*(x_i) = 1 \quad &\implies \quad \alpha_i^* \in \left[0, \frac{c}{n}\right] \\ y_i f^*(x_i) > 1 \quad &\implies \quad \alpha_i^* = 0 \end{aligned}$$

3 Kernelization

Since we've found that

$$w^* = \sum_{i=1}^n \alpha_i^* y_i x_i,$$

we know that predictions will be of the form

$$f^*(x) = (w^*)^T x = \sum_{i=1}^n \alpha_i^* y_i x_i^T x.$$

Let's

..

we know that w^* can be written as a linear combination of the x_i's. Let's write this as we can plug this expression back into the primal problem:

$$\min_{w \in \mathbf{R}^d} \frac{1}{2}||w||^2 + \frac{c}{n} \sum_{i=1}^n (1 - y_i w^T x_i)_+ \tag{3.1}$$

References