SVM: retraining with just the support vectors?

David S. Rosenberg

1 Question

Consider the following formulation of the SVM objective function:

$$J(w) = \sum_{i=1}^{n} \ell(w^T x_i, y_i) + \lambda \|w\|^2,$$

for $\lambda > 0$ and where the loss function is the hinge loss $\ell(\hat{y}, y) = (1 - \hat{y} y_i)_+$, where $(x)_+ = x\mathbb{1}[x \geq 0]$ refers to the "positive part" of x. This differs from our usual objective $J'(w) = \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n} \ell(w^T x_i, y_i)$, but the two will produce the same set of solutions as we vary the hyperparameters $\lambda, c \in (0, \infty)$.

We know from the duality theory of SVMs that the minimizer of $J(w)$ can be written as

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

where some subset of the α_i^*'s may be exactly 0. For prediction, we don't need to save the (x_i, y_i) points for which $\alpha_i^* = 0$. One natural question is, what happens if we remove these points from the training set and re-fit the model? Perhaps the solution doesn't change at all?

2 Answer to an easier question

We can't show that dropping all points with $\alpha_i^* = 0$ from the training set won't change the answer. But here we show something a bit weaker: if we drop all training points that are on the "good side of the margin", then the solution does not change. In other words, we can drop all training points for which $y_i x_i^T w^* > 1$ and still end up with the same trained model. The set of training examples for which $y_i x_i^T w^* > 1$ all have $\alpha_i^* = 0$, but there may be some points with $\alpha_i^* = 0$ for which $y_i x_i^T w^* = 1$, and so wouldn't be excluded. Here's the proof of our claim:

Without loss of generality, index the x_i's so that $x_{m+1}, \ldots, x_n$ are all the points on the "good side of the margin" (i.e. $y_i x_i^T w^* > 1$). Then we know that $\alpha_{m+1}^*, \ldots, \alpha_n^* = 0$. Let's define

$$J_1(w) = \sum_{i=1}^{m} \ell(w^T x_i, y_i) + \lambda \|w\|^2$$

and let

$$J_2(w) = \sum_{m+1}^{n} \ell(w^T x_i, y_i).$$

Note that $J(w) = J_1(w) + J_2(w)$. The claim is that if w^* is the minimizer of $J(w)$, then it is also the minimizer of $J_1(w)$. We'll do this with a local analysis of J and J_1 around w^*. The relation $y_i x_i^T w^* > 1$ holds for each $i = m+1, \ldots, n$. Moreover, since $y_i x_i^T w$ is a continuous function of w for each i, there is some ε-ball around w^* for which $y_i x_i^T w > 0$ for all i and for all w in the ball. Thus in that ball, i.e. for all $\{w \mid \|w - w^*\| < \varepsilon\}$, we have $\ell(w^T x_i, y_i) = \left(1 - y_i w^T x_i\right)_+ = 0$, and so $J_2(w) \equiv 0$. Thus in that ball, $J_1(w) = J(w)$. Since w^* is a local minimizer of $J(w)$ in the ball, it is also a local minimizer of $J_1(w)$. By convexity of $J_1(w)$, w^* is a global minimizer of J_1, and so the solution is unchanged by dropping the training points on the good side of the margin.

3 Challenge

What happens if exclude all points with $\alpha_i^* = 0$? Either show that we may end up with a different solution w^* or show that the solution is unchanged.