SVM: RETRAINING WITH JUST THE SUPPORT VECTORS?

DAVID S. ROSENBERG

1. Setup

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 $(x)_+=x1(x\geq 0)$ refers to the "positive part" of x. This differs from our usual objective $J'(w)=\frac{1}{2}||w||^2+\frac{c}{n}\sum_{i=1}^{n}\ell(w^Tx_i,y_i)$, 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? We can't quite show that, 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 $y_ix_i^Tw^*>1$ and still end up with the same trained model. The set of training examples for which $y_ix_i^Tw^*>1$ all have $\alpha_i^*=0$, but there may be some points with $\alpha_i^*=0$ for which $y_ix_i^Tw^*=1$, and so wouldn't be excluded. Here's the proof of our claim:

Without loss of generality, index the x_i's so $x_{m+1},\ldots,x_n$ are all on the "good side of the margin" (i.e. $y_ix_i^Tw^*>1$). 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 a minimizer of $J(w)$, then it is also a minimizer of $J_1(w)$. We'll do this with a local analysis of J and J_1 around w^*. The relation $y_ix_i^Tw^*>1$ holds for each $i=m+1,\ldots,n$. Moreover, since $y_ix_i^Tw$ is a continuous function of w for each i, these inequality relations will also hold for w in an ε-ball around w^*, for small enough $\varepsilon>0$. Thus in that ball, $\ell(w^Tx_i,y_i)=\left(1-y_iw^Tx_i\right)_+=0$, and so $J_2(w)\equiv 0$ for $\|w-w^*\|<\varepsilon$. Thus in that ball, $J_1(w)=J(w)$, and so w^* is a local minimizer

Date: October 31, 2021 (Revised November 10, 2022).

of $J_1(w)$. By convexity of $J_1(w)$, w^* is also a global minimizer of J_1, and so the solution is unchanged by dropping the training points on the good side of the margin.

This argument fails if we only have $y_i x_i^T w^* \geq 1$ for some i, and so we have not shown that we can just drop all training points with $\alpha_i^* = 0$.