SVM INTUITION

WILLIAM MCCLOSKEY

Given a binary classification problem on a set of data points $(x_1, y_1), \ldots, (x_m, y_m)$, $x \in \mathbb{R}^n$, $y \in \{-1, 1\}$, the support vector machine is obtained by minimizing the objective

$$w^*,\, b^* = \arg\min L(w, b)$$

$$L(w, b) = \lambda \|w\|_2^2 + \frac{1}{n} \sum_{i=1}^{n} \ell_{w,b}(x_i, y_i)$$

$$\ell_{w,b}(x, y) = \max\left(0, 1 - y(w^T x + b)\right)$$

The prediction $\hat{y}$ of the SVM is obtained by taking the sign of $w^T x + b$.

The distance of a vector x to the hyperplane $P_{w,b}$ defined by $w^T x + b = 0$ is given by $\text{dist}_{P_{w,b}}(x) = \frac{|w^T x + b|}{\|w\|_2}$. So we can rewrite the loss

$$\ell_{w,b}(x, y) = \max\left(0, 1 - y\hat{y}\|w\|_2 \text{dist}_{P_{w,b}}(x)\right).$$

The vectors x_i that contribute to the objective are called support vectors, i.e. those vectors for which $\ell_{w,b}(x_i, y_i)$ is nonzero. These support vectors are vectors that are incorrectly classified ($\hat{y} \neq y$) or are within a distance $\|w\|_2^{-1}$ of the hyperplane $P_{w,b}$.

An important property of the SVM is that optimal weight vector w^* is a linear combination of support vectors. The usual derivation of this property using Lagrange multipliers can somewhat obscure the geometry, so we hope to make the picture clear here.

First, we establish that the weight vector w^* is in the span of the data points $V = \text{span}\{x_1, \ldots, x_m\}$. We can write $w^* = v + v_p$ where $v \in V$ and $v_p \in V^\perp$. Now $v_p^T x_i = 0$ for all x_i, so each term $\ell_{w,b}(x_i, y_i)$ in the objective does not depend on v_p. However, by orthogonality $\|w^*\|_2^2 = \|v\|_2^2 + \|v_p\|_2^2$, so v_p must be zero for otherwise w^* would not be optimal. As a consequence

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

where $\alpha_i \in \mathbb{R}$. Finally, note only support vectors contribute to the objective $L(w, b)$. By strict convexity of $L(w, b)$, the solution w^* is unique and therefore must only depend on support vectors. So we must have $\alpha_i = 0$ for all non-support vectors α_i.

Thus, we have shown that w^* is a linear combination of support vectors. (For infinite dimensional spaces, simply replace the transpose with an inner product or kernel.)