Support Vector Machines

Greg Strabel

May 5, 2019

1 Linear SVM

Recall that a hyperplane H in $\mathbb{R}^P$ is defined by a normal vector, $\beta \in \mathbb{R}^P$, and an offset from the origin, $\delta \in \mathbb{R}$, so that the hyperplane is the set $H = \{x : \beta \cdot x - \delta = 0\}$. Given a set of data $\{y_i, x_i\}_{i=1}^n$ where $x_i \in \mathbb{R}^P$ and $y_i \in \{-1, 1\}$, the objective is to find a hyperplane that separates the data according to the y values. In order for a separating hyperplane to generalize well to additional data, we prefer the hyperplane that maximizes the distance between it and the closest points on either side. Given the normal vector β and offset δ, there are two auxiliary hyperplanes $H_{-1} = \{x : \beta \cdot x - \delta = -1\}$ and $H_1 = \{x : \beta \cdot x - \delta = 1\}$ that are parallel to the hyperplane H and both lie a distance of $\|\beta\|^{-1}$ from it; this distance is called the margin. Thus, we can formulate our objective as finding the values of β and δ that satisfy $y_i (\beta \cdot x_i - \delta) \geq 1$ for all i and minimize $\|\beta\|$.

One problem with this formulation remains: the data may not be linearly separable. To accommodate this case we introduce slack variables that allow individual data points to lie on the wrong sides of the two auxiliary hyperplanes. For each i, we have a slack variable $\zeta_i \geq 0$ so that $y_i (\beta \cdot x_i - \delta) \geq 1 - \zeta_i$.

There is a trade-off between increasing the margin (distance between the auxiliary hyperplanes) and allowing greater slackness for data points on the wrong sides of the hyperplanes. We control this trade-off through a parameter λ. Thus, the problem to solve is:

$$\min_{\beta,\delta} \frac{1}{2}\|\beta\|^2 + \lambda \sum_{i=1}^{n} \zeta_i \tag{1}$$

subject to the constraints:

$$\forall i \ : \ y_i (\beta \cdot x_i - \delta) \geq 1 - \zeta_i \ \wedge \ \zeta_i \geq 0 \tag{2}$$

The Lagrangian for this problem is:

$$L = \frac{1}{2}\|\beta\|^2 + \lambda \sum_{i=1}^{n} \zeta_i - \sum_{i=1}^{n} \gamma_i \left[y_i (\beta \cdot x_i - \delta) - 1 + \zeta_i \right] - \sum_{i=1}^{n} \eta_i \zeta_i \tag{3}$$

The first order conditions are:

$$\frac{\partial L}{\partial \beta_j} = \beta_j - \sum_{i=1}^{n} \gamma_i y_i x_{ij} = 0 \tag{4}$$

$$\frac{\partial L}{\partial \delta} = \sum_{i=1}^{n} \gamma_i y_i = 0 \tag{5}$$

$$\frac{\partial L}{\partial \zeta_i} = \lambda - \gamma_i - \eta_i = 0 \tag{6}$$

Condition (4) implies:

$$\|\beta\|^2 = \sum_{i=1}^{n} \gamma_i y_i \beta \cdot x_i \tag{7}$$

and

$$\|\beta\|^2 = \sum_{i=1}^{n} \sum_{k=1}^{n} \gamma_i \gamma_k y_i y_k x_i \cdot x_k \tag{8}$$

Substituting into the Lagrangian, we have

$$\begin{aligned} L &= \frac{1}{2}\|\beta\|^2 + \lambda \sum_{i=1}^{n} \zeta_i - \|\beta\|^2 + \delta \sum_{i=1}^{n} \gamma_i y_i + \sum_{i=1}^{n} \gamma_i - \sum_{i=1}^{n} \gamma_i \zeta_i - \sum_{i=1}^{n} (\lambda - \gamma_i)\, \zeta_i \\ &= -\frac{1}{2}\|\beta\|^2 + \sum_{i=1}^{n} \gamma_i \\ &= \sum_{i=1}^{n} \gamma_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \gamma_i \gamma_k y_i y_k x_i \cdot x_k \end{aligned} \tag{9}$$

It follows that the Lagrangian dual problem is:

$$\max_{\gamma} \sum_{i=1}^{n} \gamma_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \gamma_i \gamma_k y_i y_k x_i \cdot x_k \tag{10}$$

subject to the constraints:

$$\forall i \ : \ 0 \leq \gamma_i \leq \lambda \tag{11}$$

and

$$\sum_{i=1}^{n} \gamma_i y_i = 0 \tag{12}$$

This is a convex programming problem and can be solved by standard numerical techniques. The Kharush-Kuhn-Tucker conditions for a solution require that:

$$\gamma_i \left[y_i \left(\beta \cdot x_i - \delta \right) - 1 + \zeta_i \right] = 0 \tag{13}$$

which implies that if $y_i(\beta \cdot x_i - \delta) > 1$ then $\gamma_i = 0$. From condition (4) it follows that

$$\beta = \sum_{i=1}^{n} \gamma_i y_i x_i \tag{14}$$

where $\gamma_i \neq 0$ only if observation i lies on the wrong side of the corresponding auxiliary hyperplane.

2 Nonlinear SVM

The hyperplane from the previous section was constructed in the input space $\mathbb{R}^P$. Instead, we could consider a function ϕ that maps $\mathbb{R}^P$ to a transformed feature space $\mathbb{R}^{P'}$ and find a hyperplane in this space. If we replace x_i with $\phi(x_i)$ throughout the derivations in the previous section we arrive at the "kernel trick" for general nonlinear SVMs:

The Lagrangian dual problem is now:

$$\max_{\gamma} \sum_{i=1}^{n} \gamma_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \gamma_i \gamma_k y_i y_k \phi(x_i) \cdot \phi(x_k) \tag{15}$$

subject to the constraints:

$$\forall i \ : \ 0 \leq \gamma_i \leq \lambda \tag{16}$$

and

$$\sum_{i=1}^{n} \gamma_i y_i = 0 \tag{17}$$

It follows that the SVM will depend on ϕ only through terms of the form $\phi(x_i) \cdot \phi(x_k)$. This implies that we do not even need to specify ϕ; it is enough to specify values of the dot product using a kernel function, K. Recall that given a nonempty set $\mathfrak{X}$, $K : \mathfrak{X} \times \mathfrak{X} \to \mathbb{R}$ is a kernel if it is symmetric and positive definite:

$$\sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j K(x_i, x_j) \geq 0 \ \ \forall n \in \mathbb{N}, \ \forall c_i, c_j \in \mathbb{R}, \ \forall x_i, x_j \in \mathfrak{X} \tag{18}$$

Kernels commonly used in nonlinear SVMs include:

1. Homogeneous polynomial: $K(x_i, x_j) = (x_i \cdot x_j)^d$
2. Inhomogeneous polynomial: $K(x_i, x_j) = (x_i \cdot x_j + 1)^d$
3. Gaussian radial basis function: $K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$ for $\gamma > 0$

Note that to classify points using the nonlinear SVM and the kernel trick, we now use:

$$\mathbf{sign} \sum_{i=1}^{n} \gamma_i y_i K(x_i, x) \tag{19}$$