Support Vector Machines

These notes are based on Mohri, Rostamizadeh and Talwalkar (2012).

Some Convex Optimization. Consider

$$\min_x f(x) \quad \text{subject to} \quad g_i(x) \leq 0 \quad i = 1, \ldots, m.$$

Define the Lagrangian

$$\mathcal{L} = f(x) + \sum_j \alpha_j g_j(x).$$

The *dual function* is define by

$$F(\alpha) = \inf_x \mathcal{L}.$$

A central result in convex optimization is that the original problem can be solved by maximizing F subject to $\alpha_i \geq 0$ and $\alpha_i g(x_i) = 0$.

Hyperplanes and SVM's. Suppose we have data $(X_1, Y_1), \ldots, (X_n, Y_n)$ that can be separated by a hyperplane. Let $b + w^T x = 0$ be such a hyperplane. Note that $Y_i(b + X_i^T w) \geq 1$ for all i. Any re-scaled version of the hyper-plane is the same classifier. So re-scale the hyper-plane so that

$$\min_i |b + w^T X_i| = 1.$$

If x_0 is any point, then using some simple algebra, we find that the distance to the hyperplane is

$$\frac{|b + w^T x_0|}{||w||}.$$

We call the distance to the closest point, the *margin* ρ. Since $|\min_i |b + w^T X_i| = 1$, we see that

$$\rho = \min_i \frac{|w^T X_i + b|}{||w||} = \frac{1}{||w||}.$$

The support vector machine (SVM) is the hyperplane that maximized the margin. But maximizing $1/||w||$ is the same is minimizing $||w||$ which is the same as minimizing $(1/2)||w||^2$. So finding the SVM corresponds to:

$$\min_{w,b} \quad \frac{1}{2}||w||^2 \qquad \text{subject to } Y_i(w^T X_i + b) \geq 1 \quad i = 1, \ldots, n.$$

The Lagrangian for this problem is

$$\mathcal{L} = \frac{1}{2}||w||^2 - \sum_i \alpha_i [Y_i(w^T X_i + b) - 1]$$

where $\alpha_i \geq 0$ and $\alpha_i[Y_i(w^T X_i + b) - 1] = 0$. If we set $\nabla_w \mathcal{L} = 0$ and $\nabla_b \mathcal{L} = 0$ we get the two equations

$$w = \sum_i \alpha_i Y_i X_i = 0$$
$$0 = \sum_i \alpha_i Y_i.$$

If we insert $w = \sum_i \alpha_i Y_i X_i$ into $\mathcal{L}$ and use the fact that $\sum_i \alpha_i Y_i = 0$ we get

$$\mathcal{L} = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j Y_i Y_j (X_i^T X_j).$$

This leads to the optimization

$$\text{maximize } \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j Y_i Y_j (X_i^T X_j)$$

subject to $\alpha_i \geq 0$ and $\alpha_i[Y_i(w^T X_i + b) - 1] = 0$. Note two importnat facts: (i) this is a quadratic program so it can be solved quickly and (ii) we don't need the X_i's we only need the inner products $X_i^T X_j$.

Consider the constraint $\alpha_i[Y_i(w^T X_i + b) - 1] = 0$. If $\alpha_i > 0$ then $Y_i(w^T X_i + b) = 1$ which implies that this point lies on the boundary of the margin. Such a point is called a support vector. On the other hand, if $Y_i(w^T X_i + b) > 1$ then $\alpha_i = 0$. Since $w = \sum_i \alpha_i Y_i X_i$ this means that the hyperplane only depends on the support vectors.

If (X_i, Y_i) is a support vector then $W^T X_i + b = Y_i$. Since $w = \sum_j \alpha_j Y_j X_j$, we see that

$$b = Y_i - \sum_j \alpha_j Y_j X_j^T X_i.$$

Multiply by $\alpha_i Y_i$ and sum to get

$$\sum_i \alpha_i Y_i b = \sum_i \alpha_i Y_i^2 - \sum_{i,j} \alpha_i \alpha_j Y_i Y_j (X_i^T X_j).$$

Since $Y_i^2 = 1$, $w = \sum_i \alpha_i Y_i X_i$ and $\sum_i \alpha_i Y_i = 0$ this implies that

$$0 = \sum_i \alpha_i - ||w||^2.$$

The margin ρ is $1/||w||$ so that

$$\rho^2 = \frac{1}{||w||^2} = \frac{1}{\sum_i \alpha_i} = \frac{1}{||\alpha||_1}.$$

The Non-separable Case. Usually, the data are not linearly separable. So we can't assume that $Y_i(w^T X_i + b) \geq 1$. We introduce slack variables $\xi_i \geq 0$ and instead require

$$Y_i(W_i^T X_i + b) \geq 1 - \xi_i.$$

This allows points to be incorrectly classified. But it also allows points to be correctly classified but be inside the margin. We change the optimization problem to

$$\min_{w,b,\xi} \quad \frac{1}{2}||w||^2 + C\sum_i \xi_i$$

subject to $Y_i(w^T X_i + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$. The constant $C \geq 0$ controls the amount of slack that is allowed.

The Lagrangian is

$$\mathcal{L} = \frac{1}{2}||w||^2 + C\sum_i \xi_i - \sum_i \alpha_i[Y_i(w^T X_i + b) - 1 + \xi_i] - \sum_i \beta_i \xi_i.$$

Setting the derivative to 0 leads to the conditions

$$\begin{aligned} w &= \sum_i \alpha_i Y_i X_i \\ 0 &= \sum_i \alpha_i Y_i \\ C &= \alpha_i + \beta_i \\ 0 &= \alpha_i \text{ or } Y_i(w^T X_i + b) = 1 - \xi_i \\ 0 &= \beta_i \text{ or } \xi_i = 0. \end{aligned}$$

When $\alpha_i > 0$ we call X_i a support vector. If $\alpha_i \neq 0$ then

$$Y_i(w^T X_i + b) = 1 - \xi_i.$$

If $\xi_i = 0$ then X_i lies on the marginal hyperplane. If $\xi_i \neq 0$ then $\beta_i = 0$ which implies $\alpha_i = C$. In summary, support vectors lie on the marginal hyperplane or $\alpha_i = C$.

The dual problem has a simple form:

$$\max_\alpha \quad \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j Y_i Y_j X_i^T X_j$$

subject to $0 \leq \alpha_i \leq C$ and $\sum_i \alpha_i Y_i = 0$. Again, it is a quadratic program and only involves inner products of the X_i.

Since the VC dimension of hyperplane classifiers is $d+1$, we know that, with probability at least $1-\delta$,

$$R(h) \leq R(\widehat{h}) + \sqrt{\frac{2(d+1)\log(en/(n+1))}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}. \tag{1}$$

But this bound does not use the structure of SVM's. For this, we turn to margin theory.

Margins. Recall that the margin is

$$\rho = \min_i \frac{Y_i(w^T X_i + b)}{||w||}.$$

We can improve the VC bound using the margin.

Theorem 1 *Suppose that the sample space is contained in* $\{x : ||x|| \leq r\}$. *Let* $\mathcal{H}$ *be the set of hyperplanes satisfying* $||w|| \leq \Lambda$ *and* $\min_i |w^T X_i| = 1$. *Then* $\mathrm{VC}(\mathcal{H}) \leq r^2\Lambda^2$.

Proof. Suppose that $\{x_1, \ldots, x_d\}$ can be shattered. Then for $y \in \{-1, +1\}^d$ there exists w such that $1 \leq y_i(w^T x_i)$ for all i. Sum over i to get

$$d \leq w^T \sum_i y_i x_i \leq ||w|| \;\; ||\sum_i y_i x_i|| \leq \Lambda \; ||\sum_i y_i x_i||.$$

This holds for all choices of y_i. So it holds if Y_i is drawn uniformly over $\{-1, +1\}$. Thus $\mathbb{E}[Y_i Y_j] = \mathbb{E}[Y_i][Y_j] = 0$ for $i \neq j$ and $\mathbb{E}[Y_i Y_i] = 1$. So

$$\begin{aligned} d &\leq \Lambda \mathbb{E}||\sum_{i=1}^d Y_i x_i|| \leq \Lambda \sqrt{\mathbb{E}||\sum_i Y_i x_i||^2} \\ &= \Lambda \sqrt{\sum_{i,j} \mathbb{E}[Y_i Y_j] x_i^T x_j} = \Lambda \sqrt{\sum_i x_i^T x_i} \\ &\leq \Lambda \sqrt{d r^2} = \Lambda r \sqrt{d} \end{aligned}$$

so that $d \leq r^2 \Lambda$. $\square$

If the data are separable, the hyperplane satisfies $||w|| = 1/\rho$ so that $\Lambda^2 = 1/\rho^2$ and hence $d \leq r^2/\rho^2$. Plugging this into (1) we get

$$R(h) \leq R(\widehat{h}) + \sqrt{\frac{2r^2 \log((en\rho^2)/r^2)}{n\rho^2}} + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{2}$$

which is dimension independent.

Nonparametric SVM's. We can get a nonparametric SVM using RKHS's by replacing x with a feature map $\Phi(x)$. Recall that $\Phi(x_1)^T \Phi(x_2) = K(x_1, x_2)$. So we get a nonparaametric

SVM by solving

$$\max_{\alpha} \quad \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j Y_i Y_j K(X_i, X_j)$$

subject to $0 \leq \alpha_i \leq C$ and $\sum_i \alpha_i Y_i = 0$. The classifier is

$$h(x) = \text{sign}\left(\sum_i Y_i K(X_i, x) + b\right).$$

This is a nonlinear (nonparametric) classifer.