The SVM Optimization Formulation with Kernels

In order to show how we can learn an SVM using only kernel functions, we will rely on a result called the Representer Theorem. Recall the generalized optimization formulation of the SVM:

$$\min_{\mathbf{w}} \{ f(\langle \mathbf{w}, \psi(\mathbf{x}_1) \rangle, \langle \mathbf{w}, \psi(\mathbf{x}_2) \rangle, \ldots, \langle \mathbf{w}, \psi(\mathbf{x}_m) \rangle) + R(\|\mathbf{w}\|) \} \tag{1}$$

The Representer Theorem is then,

Theorem (Representer Theorem): *Given a set of vectors* $S := \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ *residing in a vector space* $\mathcal{X}$ *and mapping* $\psi : \mathcal{X} \to \mathcal{F}$ *where* $\mathcal{F}$ *is a Reproducing Kernel Hilbert Space, then there exists a vector* $\boldsymbol{\alpha} \in \mathbb{R}^m$ *such that*

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i \psi(\mathbf{x}_i)$$

is the solution to Equation 1

Proof:

First, note that the vectors $\psi(\mathbf{x}_1), \psi(\mathbf{x}_2), \ldots, \psi(\mathbf{x}_m)$ span a subspace of $\mathcal{F}$, which we denote as F. That is, $F \subseteq \mathcal{F}$. Though not proven rigorously here, it follows that $\mathcal{F}$ can be formed by

$$\mathcal{F} = F \oplus F^{\perp}$$

where $F^{\perp}$ is the orthogonal compliment of F. That is $F^{\perp}$ is the set of all vectors orthogonal to vectors in F. Thus, every vector $\mathbf{f} \in \mathcal{F}$ can be written as

$$\mathbf{f} = \mathbf{f}_{\|} + \mathbf{f}^{\perp}$$

where $\mathbf{f}_{\|} \in F$ and $\mathbf{f}^{\perp} \in F^{\perp}$.

Now, let $\mathbf{w}^*$ be the solution to Equation 1. By the previous fact, $\mathbf{w}^*$ can be formed by

$$\mathbf{w}^* = \sum_{i=1}^{m} \alpha_i \psi(\mathbf{x}_i) + \mathbf{u}$$

where $\mathbf{u} \in F^{\perp}$. That is, $\mathbf{w}^*$ can be formed by taking a linear combination of the $\psi(\mathbf{x}_i)$ vectors added to at least one vector $\mathbf{u}$ that is orthogonal to all of the $\psi(\mathbf{x}_i)$.

We'll denote the linear combination of $\psi(\mathbf{x}_i)$ vectors used to construct $\mathbf{w}^*$ as $\mathbf{w}$. That is,

$$\mathbf{w} := \sum_{i=1}^{m} \alpha_i \psi(\mathbf{x}_i)$$

Thus,

$$\mathbf{w}^* = \mathbf{w} + \mathbf{u}$$

$$\implies \mathbf{w} = \mathbf{w}^* - \mathbf{u}$$

Now, since $\mathbf{u}$ is orthogonal to all $\psi(\mathbf{x}_i)$ vectors it follows that for all i,

$$\langle \mathbf{u}, \psi(\mathbf{x}_i) \rangle = 0$$

$$\implies \langle \mathbf{u}, \alpha_i \psi(\mathbf{x}_i) \rangle = 0$$

$$\implies \sum_{j=1}^{m} \langle \mathbf{u}, \alpha_j \psi(\mathbf{x}_j) \rangle = 0$$

$$\implies \left\langle \mathbf{u}, \sum_{j=1}^{m} \alpha_j \psi(\mathbf{x}_j) \right\rangle = 0 \qquad \text{by linearity of the inner-product}$$

$$\implies \langle \mathbf{u}, \mathbf{w} \rangle = 0$$

Now, looking at the value of the f function evaluated on $\mathbf{w}$ we see that

$$\begin{aligned} f(\langle \mathbf{w}, \psi(\mathbf{x}_1) \rangle, \ldots, \langle \mathbf{w}, \psi(\mathbf{x}_m) \rangle) &= f(\langle \mathbf{w}^* - \mathbf{u}, \psi(\mathbf{x}_1) \rangle, \ldots, \langle \mathbf{w}^* - \mathbf{u}, \psi(\mathbf{x}_m) \rangle) \\ &= f(\langle \mathbf{w}^*, \psi(\mathbf{x}_1) \rangle - \langle \mathbf{u}, \psi(\mathbf{x}_1) \rangle, \ldots, \langle \mathbf{w}^*, \psi(\mathbf{x}_m) \rangle - \langle \mathbf{u}, \psi(\mathbf{x}_1) \rangle) \\ &= f(\langle \mathbf{w}^*, \psi(\mathbf{x}_1) \rangle, \ldots, \langle \mathbf{w}^*, \psi(\mathbf{x}_m) \rangle) \end{aligned}$$

Thus, we see that f evaluated on $\mathbf{w}$ is equal to f evaluated on the optimal vector $\mathbf{w}^*$.

To prove that $\mathbf{w}$ is also an optimal vector, we need to show that R evaluated on $\mathbf{w}^*$ is greater than or equal to R evaluated on $\mathbf{w}$. Since we have established that $\mathbf{w} \perp \mathbf{u}$, then it follows that

$$\|\mathbf{w}^*\|^2 = \|\mathbf{w}\|^2 + \|\mathbf{u}\|^2$$

$$\geq \|\mathbf{w}\|^2$$

Since R is, by definition, a monotonically increasing function, it follows that

$$R\left(\|\mathbf{w}^*\|^2\right) \geq R\left(\|\mathbf{w}\|^2\right)$$

□

Thus, with the Representer Theorem, we have show that our solution to the SVM optimization formulation can be expressed as the sum of the weighted example vectors. That is,

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i \psi(\mathbf{x}_i)$$

Then, the inner product $\langle \mathbf{w}, \psi(\mathbf{x}_i) \rangle$ can be expressed as

$$\langle \mathbf{w}, \psi(\mathbf{x}_i) \rangle = \left\langle \sum_{j=1}^{m} \alpha_j \psi(\mathbf{x}_j), \psi(\mathbf{x}_i) \right\rangle$$

$$= \sum_{j=1}^{m} \langle \alpha_j \psi(\mathbf{x}_j), \psi(\mathbf{x}_i) \rangle$$

$$= \sum_{j=1}^{m} \alpha_j \langle \psi(\mathbf{x}_j), \psi(\mathbf{x}_i) \rangle$$

$$= \sum_{j=1}^{m} \alpha_j K(\mathbf{x}_j, \mathbf{x}_i)$$

Similary,

$$
\begin{aligned}
\|\mathbf{w}\|^2 &= \langle \mathbf{w}, \mathbf{w} \rangle \\
&= \left\langle \sum_{j=1}^{m} \alpha_j \psi(\mathbf{x}_j), \sum_{i=1}^{m} \alpha_i \psi(\mathbf{x}_i) \right\rangle \\
&= \sum_{j=1}^{m} \alpha_j \left\langle \psi(\mathbf{x}_j), \sum_{i=1}^{m} \alpha_i \psi(\mathbf{x}_i) \right\rangle \\
&= \sum_{j=1}^{m} \sum_{i=1}^{m} \alpha_j \alpha_i \left\langle \psi(\mathbf{x}_j), \psi(\mathbf{x}_i) \right\rangle \\
&= \sum_{j=1}^{m} \sum_{i=1}^{m} \alpha_j \alpha_i K(\mathbf{x}_j, \mathbf{x}_i)
\end{aligned}
$$

Therefore, the SVM optimization can be formulated only in terms of the kernel function as follows

$$
\min_{\alpha} \left\{ f\left(\sum_{j=1}^{m} \alpha_j K(\mathbf{x}_j, \mathbf{x}_1), \ldots, \sum_{j=1}^{m} \alpha_j K(\mathbf{x}_j, \mathbf{x}_m) \right) + R\left(\sqrt{\sum_{j=1}^{m} \sum_{i=1}^{m} \alpha_j \alpha_i K(\mathbf{x}_j, \mathbf{x}_i)} \right) \right\} \tag{2}
$$

The Gram Matrix

In practice, an efficient implementation of an SVM would first compute the value of the Kernel function for every pair of training vectors. Each value is stored in an $m \times m$ matrix G called the **Gram** matrix. That is, G is defined as

$$
G_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)
$$

We know show how to compute $\|\mathbf{w}\|^2$ and $\langle \mathbf{w}, \mathbf{x}_i$ in terms of the Gram matrix.

Claim 1:

$$
\sum_{j=1}^{m} \alpha_j K(\mathbf{x}_j, \mathbf{x}_i) = (G\boldsymbol{\alpha})_i
$$

Proof:

$$G\boldsymbol{\alpha} = \alpha_1 \begin{bmatrix} G_{1,1} \\ G_{2,1} \\ \vdots \\ G_{m,1} \end{bmatrix} + \alpha_2 \begin{bmatrix} G_{1,2} \\ G_{2,2} \\ \vdots \\ G_{m,2} \end{bmatrix} + \cdots + \alpha_m \begin{bmatrix} G_{1,m} \\ G_{2,m} \\ \vdots \\ G_{m,m} \end{bmatrix}$$

$$= \begin{bmatrix} \alpha_1 G_{1,1} + \alpha_2 G_{1,2} + \cdots + \alpha_m G_{1,m} \\ \alpha_1 G_{2,1} + \alpha_2 G_{2,2} + \cdots + \alpha_m G_{2,m} \\ \vdots \\ \alpha_1 G_{m,1} + \alpha_2 G_{m,2} + \cdots + \alpha_m G_{m,m} \end{bmatrix}$$

Thus we see that the *i*th value of this vector is

$$\alpha_1 G_{i,1} + \alpha_2 G_{i,2} + \cdots + \alpha_m G_{i,m} = \alpha_1 K(\mathbf{x}_i, \mathbf{x}_1) + \alpha_2 K(\mathbf{x}_i, \mathbf{x}_2) + \cdots + \alpha_m K(\mathbf{x}_i, \mathbf{x}_m)$$

$$= \sum_{j=1}^{m} \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$= \sum_{j=1}^{m} \alpha_j K(\mathbf{x}_j, \mathbf{x}_i)$$

□

Claim 2:

$$\sum_{j=1}^{m} \sum_{i=1}^{m} \alpha_j \alpha_i K(\mathbf{x}_j, \mathbf{x}_i) = \boldsymbol{\alpha}^T G \boldsymbol{\alpha}$$

Proof:

From Claim 1,

$$G\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 G_{1,1} + \alpha_2 G_{1,2} + \cdots + \alpha_m G_{1,m} \\ \alpha_1 G_{2,1} + \alpha_2 G_{2,2} + \cdots + \alpha_m G_{2,m} \\ \vdots \\ \alpha_1 G_{m,1} + \alpha_2 G_{m,2} + \cdots + \alpha_m G_{m,m} \end{bmatrix}$$

Now,

$$\boldsymbol{\alpha}^T G\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_m] \begin{bmatrix} \alpha_1 G_{1,1} + \alpha_2 G_{1,2} + \cdots + \alpha_m G_{1,m} \\ \alpha_1 G_{2,1} + \alpha_2 G_{2,2} + \cdots + \alpha_m G_{2,m} \\ \vdots \\ \alpha_1 G_{m,1} + \alpha_2 G_{m,2} + \cdots + \alpha_m G_{m,m} \end{bmatrix}$$

$$= \alpha_1(\alpha_1 G_{1,1} + \alpha_2 G_{1,2} + \cdots + \alpha_m G_{1,m}) + \cdots + \alpha_m(\alpha_1 G_{m,1} + \alpha_2 G_{m,2} + \cdots + \alpha_m G_{m,m})$$

$$= \sum_{j=1}^{m} \sum_{i=1}^{m} \alpha_j \alpha_i G_{j,i}$$

$$= \sum_{j=1}^{m} \sum_{i=1}^{m} \alpha_j \alpha_i K(\mathbf{x}_j, \mathbf{x}_i)$$

□

SVM Optimization with Kernels

Hard-SVM with Kernels

For the hard-SVM Equation 2 becomes the optimization problem

$$\begin{aligned} &\text{minimize} && \sum_{j=1}^{m} \alpha_j \alpha_i \sum_{i=1}^{m} K(\mathbf{x}_j, \mathbf{x}_i) \\ &\text{subject to} && \forall i \; y_i \sum_{j=1}^{m} \alpha_j K(\mathbf{x}_j, \mathbf{x}_i) \geq 1 \end{aligned}$$

From Claims 1 and 2, this becomes the optimization problem

$$\begin{aligned} &\text{minimize} && \boldsymbol{\alpha}^T G\boldsymbol{\alpha} \\ &\text{subject to} && \forall i \; y_i (G\boldsymbol{\alpha})_i \geq 1 \end{aligned}$$

Soft-SVM with Kernels

For the soft-SVM Equation 2 becomes the optimization problem

$$\text{minimize} \qquad \sum_{j=1}^{m} \sum_{i=1}^{m} K(\mathbf{x}_j, \mathbf{x}_i) + \frac{\lambda}{m} \sum_{i=1}^{m} \max \left\{ 0, 1 - y \sum_{j=1}^{m} \alpha_j K(\mathbf{x}_j, \mathbf{x}_i) \right\}$$

From Claims 1 and 2, this becomes the optimization problem

$$\text{minimize} \qquad \boldsymbol{\alpha}^T G \boldsymbol{\alpha} + \frac{\lambda}{m} \sum_{i=1}^{m} \max \{0, 1 - y_i (G\boldsymbol{\alpha})_i\}$$