1 Support Vector Machines

Support vector machine is a kernelized large margin linear classifier. For binary classification problem $\mathcal{Y} = \{-1, +1\}$, we are interested in finding a linear decision bondary, parameterized by $w \in \mathbb{R}^d, b \in \mathbb{R}$, that separates the training data points by maximizing the worst case distance (margin) of each data point to the decision boundary. Given dataset $\{(x_i, y_i)\}_{i=1}^n$, we are interested in solving the following quadratic programming problem,

$$\min_{w,b} \frac{1}{2} \|w\|^2$$
$$\text{subject to } y_i(w^T x_i + b) \geq 1 \quad i = 1, 2, \cdots, n$$

To derive the dual problem, we write the Lagrangian,

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2} \|w\|^2 + \sum_{i=1}^n \alpha_i \left[1 - y_i(w^T x_i + b)\right]$$

where $\alpha = \{\alpha_i\}_{i=1}^n$ are the dual variables. Solve for $\inf_{w,b} \mathcal{L}(w, b, \alpha)$ to arrive at the dual objective. In particular, first order optimality condition gives $w = \sum_{i=1}^n \alpha_i y_i x_i$ and it must be that $0 = \sum_{i=1}^n \alpha_i y_i$. Therefore, we arrive at the dual problem,

$$\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j$$
$$\text{subject to } \alpha_i \geq 0 \quad i = 1, 2, \cdots, n \qquad \text{(dual feasibility)}$$
$$\sum_{i=1}^n \alpha_i y_i = 0 \qquad \text{(from } \nabla_b \mathcal{L} = 0\text{)}$$

The dual problem can be solved more efficiently than the primal problem using coordinate descent. The decision rule can be written entirely using dot products between input vectors,

$$f(x) = \sum_{i=1}^n \alpha_i y_i x_i^T x + b$$

We observe that optimization as well as prediction uses input vectors via dot products only. We are motivated to use feature mapping ϕ to map input vectors to a higher dimensional space in hope that the lifted space is linearly separable. The kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ allows us to compute dot product $\phi(x_i)^T \phi(x_j)$ efficiently and even without ever defining the exact forms of ϕ. We can substitute k whenever inner product is used and arrive at a large margin classifier over implicitly defined nonlinear feature mapping ϕ, i.e. support vector machines.