CSC 446 Notes: Lecture 7

Support Vector Machine

Support Vector Machine (SVM) is one of the most widely used classification methods. SVM is different from other classifiers that we have covered so far. SVM cares only about the data points near the class boundary and finds a hyperplane that maximizes the margin between the classes.

Training Linear SVM

Let the input be a set of N training vectors $\{\mathbf{x_n}\}_{n=1}^N$ and corresponding class labels $\{y_n\}_{n=1}^N$, where $\mathbf{x_n} \in \mathbb{R}^D$ and $y_n \in \{-1, 1\}$. Initially we assume that the two classes are linearly separable. The hyperplane separating the two classes can be represented as:

$$\mathbf{w}^T\mathbf{x}_n + b = 0,$$

such that:

$$\begin{aligned} \mathbf{w}^T\mathbf{x}_n + b \geq 1 \quad &\text{for} \quad y_n = +1, \\ \mathbf{w}^T\mathbf{x}_n + b \leq -1 \quad &\text{for} \quad y_n = -1. \end{aligned}$$

Let H_1 and H_2 be the two hyperplanes (Figure 1) separating the classes such that there is no other data point between them. Our goal is to maximize the margin M between the two classes. The objective function:

$$\begin{aligned} \max_{\mathbf{w}} \ & M \\ \text{s.t.} \ & y^n(\mathbf{w}^T\mathbf{x_n} + b) \geq M, \\ & \mathbf{w}^T\mathbf{w} = 1. \end{aligned}$$

The margin M is equal to $\frac{2}{\|\mathbf{w}\|}$. We can rewrite the objective function as:

$$\begin{aligned} \min_{\mathbf{w}} \ & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{s.t.} \ & y^n(\mathbf{w}^T\mathbf{x_n} + b) \geq 1 \end{aligned}$$

Now, let's consider the case when the two classes are not linearly separable. We introduce slack variables $\{\xi_n\}_{n=1}^N$ and allow few points to be on the wrong side of the hyperplane at some cost. The modified objective function:

$$\begin{aligned} \min_{\mathbf{w}} \ & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\xi_n \\ \text{s.t.} \ & y^n(\mathbf{w}^T\mathbf{x_n} + b) + \xi_n \geq 1, \\ & \xi_n \geq 0, \ \ \forall n. \end{aligned}$$

The parameter C can be tuned using development set. This is the primal optimization problem for SVM.

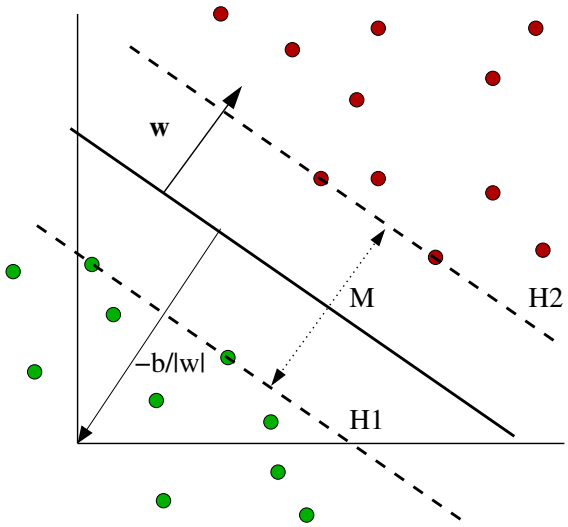

Figure 1: The figure shows a linear SVM classifier for two linearly separable classes. The hyperplane $\mathbf{w}^T x+b$ is the solid line between H_1 and H_2, and the the margin is M.

The Lagrangian for the primal problem:

$$L\left(\mathbf{w}, b, \xi, \alpha, \mu\right) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\xi_n - \sum_n \alpha_n\left[y_n(\mathbf{w}^T\mathbf{x}_n + b)\right] - \sum_n \alpha_n\xi_n + \sum_n \alpha_n - \sum_n \mu_n\xi_n,$$

where α_n and μ_n, $1 \leq n \leq N$ are Lagrange multipliers.

Differentiating the Lagrangian with respect to the variables:

$$\frac{\partial}{\partial \mathbf{w}} L\left(\mathbf{w}, b, \xi, \alpha, \mu\right) = \mathbf{w} - \sum_n \alpha_n y_n \mathbf{x}_n = 0$$

$$\frac{\partial}{\partial b} L\left(\mathbf{w}, b, \xi, \alpha, \mu\right) = -\sum_n \alpha_n y_n = 0$$

$$\frac{\partial}{\partial \xi_n} L\left(\mathbf{w}, b, \xi, \alpha, \mu\right) = C - \alpha_n - \mu_n = 0$$

Solving these equations, we get:

$$\mathbf{w} = \sum_n \alpha_n y_n \mathbf{x}_n \tag{1}$$

$$\sum_n \alpha_n y_n = 0$$

$$\alpha_n = C - \mu_n \tag{2}$$

We now plug-in these values to get the dual function and cancelling out some terms:

$$\begin{aligned} g(\alpha, \mu) &= \frac{1}{2}\sum_n\sum_m \alpha_n\alpha_m y_n y_m \mathbf{x_n}^T\mathbf{x_m} - \sum_n\sum_n \alpha_n\alpha_m y_n y_m \mathbf{x_n}^T\mathbf{x_m} + \sum_n \alpha_n \\ &= \sum_n \alpha_n - \frac{1}{2}\sum_n\sum_m \alpha_n\alpha_m y_n y_m \mathbf{x_n}^T\mathbf{x_m} \end{aligned} \tag{3}$$

Using the equation (2) and (3) and the KKT conditions, we obtain the dual optimization problem:

$$\begin{aligned} \max_{\alpha} \ & \sum_n \alpha_n - \frac{1}{2} \sum_n \sum_m \alpha_n \alpha_m y_n y_m \mathbf{x_n}^T \mathbf{x_m} \\ \text{s.t. } & 0 \leq \alpha_n \leq C. \end{aligned}$$

The dual optimization problem is concave and easy to solve. The dual variables (α_n) lie within a box with side C. We usually vary two values α_i and α_j at a time and numerically optimize the dual function. Finally, we plug in the values of the α_n^*'s to the equations (1) to obtain the primal solution $\mathbf{w}^*$.

Convex Optimization Review

Let we are given an optimization problem:

$$\begin{aligned} \min_{x} \ & f_0(x) \\ \text{s.t. } & f_i(x) \leq 0, \ \text{ for } i \in 1, 2, \ldots, K, \end{aligned}$$

where f_0 and f_i ($i \in \{1, 2, \ldots, K\}$) are convex functions. We call this optimization problem the 'primal' problem.

The Lagrangian is:

$$L(x, \lambda) = f_0(x) + \sum_{i=1}^{K} \lambda_i f_i(x)$$

The Lagrange dual function:

$$g(\lambda) = \min_x L(x, \lambda)$$

The dual function $g(\lambda)$ is concave and hence easy to solve. We can obtain the minima of a convex primal optimization problem by maximizing the dual function $g(\lambda)$. The dual optimization problem:

$$\begin{aligned} \max_{\lambda} \ & g(\lambda) \\ \text{s.t. } & \lambda_i \geq 0, \ \text{ for } i \in 1, 2, \ldots, K. \end{aligned}$$

Karush-Kuhn-Tucker (KKT) Conditions

The Karush-Kuhn-Tucker (KKT) conditions are the conditions for optimality in primal and dual functions. If f_0 and f_i's are convex, differentiable, and the feasible set has some interior points (satisfies Slater condition), the x^* and λ_i^*'s are the optimal solutions of the primal and dual problems if and only if they satisfy the following conditions:

$$\begin{aligned} f_i(x^*) &\leq 0 \\ \lambda_i^* &\geq 0, \ \forall i \in 1, \ldots, K \\ \frac{\partial}{\partial x} L(x^*, \lambda_1^*, \ldots, \lambda_K^*) &= 0 \\ \lambda_i^* f_i(x^*) &= 0 \end{aligned}$$