CS 578 Lecture Notes

Qilei Zhang

Mar 1, 2021

1 Support Vector Machine

1.1 Review

1.1.1 Margin

The smallest distance from all training points to hyperplane.

$$\text{Margin of } (w,b) = \min_{n\in[N]} \frac{y_n(w^\top \phi(x_n)+b)}{||w||_2} \tag{1}$$

1.1.2 Maximal Margin Classifier

In maximal margin classifier, in order to classify the data, we will use a separating hyperplane, which has the farthest (max) minimum (min) distance from the observations.

$$\max_{w,b} \min_{n\in[N]} \frac{y_n(w^\top \phi(x_n)+b)}{||w||_2} = \max_{w,b} \frac{1}{||w||_2} \min_{n\in[N]} y_n(w^\top \phi(x_n)+b) \tag{2}$$

1.1.3 Re-scaling

If we re-scale w and b, we are still maximizing objective and the optimization result will not change. So, we can re-scale w and b and make Margin = 1, the above problem can be rewritten as:

$$\begin{aligned} &\max_{w,b} \frac{1}{||w||_2} \\ &\text{s.t. } y_n(w^\top \phi(x_n)+b) = 1(\text{or} \geq 1) \end{aligned} \tag{3}$$

This maximization problem is equivalent to the following minimization problem:

$$\begin{aligned} &\min_{w,b} \frac{1}{2}||w||_2^2 \\ &\text{s.t. } y_n(w^\top \phi(x_n)+b) \geq 1 \end{aligned} \tag{4}$$

1.2 Hard Margin

The above statement is the formulation of Hard Margin SVM optimization problem (which is convex in nature). The constraints state that we need to prevent data points from falling into the margin, i.e. $\forall n$:

$$\begin{aligned} &(w^\top \phi(x_n)+b) \geq 1, \text{if } y_n = 1, \text{or} \\ &(w^\top \phi(x_n)+b) \leq -1, \text{if } y_n = -1 \end{aligned} \tag{5}$$

In other words, it states that each data point must lie on the correct side of the margin.

1.3 Soft Margin

So far we have assumed that the dataset is perfectly linearly separable, which doesn't really happen in real scenario. To extend SVM to cases in which the data are not linearly separable, we should introduce the slack variable ξ_n and then redefine our inequality constraint as:

$$y_n(w^\top \phi(x_n) + b) \geq 1 - \xi_n, \forall \xi_n \geq 0 \tag{6}$$

2 Optimization Problem

2.1 Primal Problem

2.1.1 Slack Variable

The Slack Variable indicates how much the point can violate the margin, which helps to define 3 types of data points:

- If $\xi = 0$ then the corresponding point ξ is on the margin or further away.
- If $0 < \xi < 1$ then the point ξ is within the margin and classified correctly (Correct side of the hyperplane).
- If $\xi \geq 1$ then the point is misclassified and present at the wrong side of the hyperplane.

The ξ is the misclassification penalty. Hence we want to minimize it during optimization.

2.1.2 Objective Function

Thus, we can write the final objective function as,

$$\begin{aligned} \min_{w,b,\{\xi_n\}} & \frac{1}{2}||w||_2^2 + C\sum_n \xi_n \\ \text{s.t. } & y_n(w^\top \phi(x_n) + b) \geq 1 - \xi_n \\ & \xi_n \geq 0, \forall n \end{aligned} \tag{7}$$

Here, C balances the cost of misclassification. It controls the tread-off between maximizing the margin and minimizing the loss.

2.1.3 Hinge Loss

Next, we can change the inequality constraint to equality constraint by rewriting the equation $y_n(w^\top \phi(x_n) + b) \geq 1 - \xi_n$ in following way:

$$\xi_n = \max(0, 1 - y_n(w^\top \phi(x_n) + b) \tag{8}$$

We can incorporate this directly to the objective function itself and calculate the loss function as,

$$\min_{w,b} \frac{||w||_2^2}{2} + C\sum_n \underbrace{\max(0, 1 - y_n(w^\top \phi(x_n) + b)}_{\text{Hinge Loss}} \tag{9}$$

That is to say that SVM is to minimize the Hinge Loss with L_2 regularizer.

2.2 Lagrangian Duality

However, this way we are not able use the objective function to solve for non-linear cases. Hence we will find an equivalent problem named Dual Problem and solve that using Lagrange Multipliers. The generalized primal form is:

$$\begin{aligned} & \min_w F(w) \\ \text{s.t. } & h_j(w) \leq 0, \forall j = \{1, 2, \ldots, J\} \end{aligned} \tag{10}$$

The Lagrangian Function is defined as:

$$L(w;\lambda_j) = F(w) + \sum_{j=1}^{J} \lambda_j h_j(w), \forall \lambda_j \geq 0 \tag{11}$$

Here λ are called Lagrange multiplier vectors associated with the problem.

2.3 Classical Result

When we look at the following function:

$$\begin{aligned} H(w) &= \max_{\lambda_j \geq 0} L(w;\lambda_j) \\ &= \max_{\lambda_j \geq 0} F(w) + \sum_{j=1}^{J} \lambda_j h_j(w) \end{aligned} \tag{12}$$

$H(w)$ is an extension to an entire domain without changing the minimum value of $F(w)$. In the first case, we require all the λ to be non-negative. Suppose we have an $h_j(w)$ strictly negative and λ_j is zero. So the objective function remains the same. In other words, in these range, w satisfies all the constraints. In the second case, there is at least one $h_j(w)$ strictly positive. We can keep $\lambda \to +\infty$ to make $H(w)$ as large as possible. Essentially, the optimal function of the $h_j(w)$ at the region in which at least one constraint violates $h_j(w) \leq 0$ is positive infinity. Specifically,

$$H(w) = \begin{cases} F(w), & \text{if } \forall j, h_j(w) \leq 0, \\ +\infty, & \text{otherwise} \end{cases} \tag{13}$$

Thus, we can solve the following equivalent problem as equation (10) without enforcing all the constraints, but it still will give the exactly same result as solution of the primal problem.

$$\min_w H(w) = \min_w \max_{\lambda_j \geq 0} L(w;\lambda_j) \tag{14}$$

2.4 Dual Problem

Dual problem looks very similar to primal problem except we switch max and min.

$$\max_{\lambda_j \geq 0} \min_w L(w;\lambda_j) \tag{15}$$

Because we take the minimum value of the w, it is no longer the variable of this function. This is a function depends on λ:

$$\begin{aligned} G(\lambda) &= \min_w L(w;\lambda_j) \\ &= \min_w F(w) + \sum_{j=1}^{J} \lambda_j h_j(w) \end{aligned} \tag{16}$$

Thus, we have two oberservations:

1. $G(\lambda)$ is a concave function with respect to λ. Because if we fix w at w_i, function becomes $F(w_i) + \sum_{j=1}^{J} \lambda_j h_j(w_i)$ as shown in the Figure 1. It takes minimal with respect to all w. Therefore, the equation (16) is the lower envelope of all the functions, where is concave, even when $F(w)$ is non-convex.

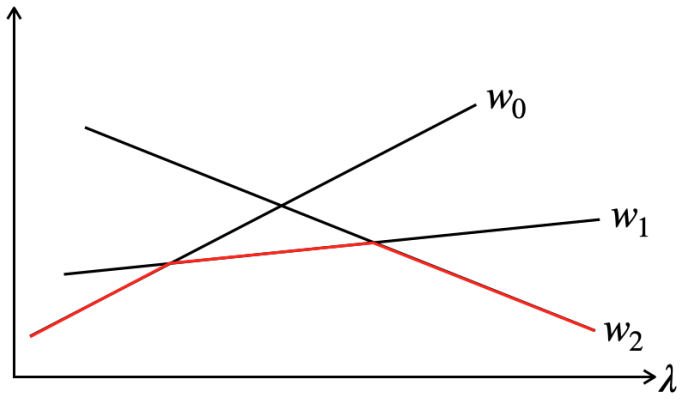

Figure 1: Fixed w_i

2. The Dual solution is a lower bound on the Primal solution. Pick any λ_0 and take w^* as the optimal value of the Primal problem,

$$\begin{aligned} G(\lambda_0) &= \min_w F(w) + \sum_{j=1}^{J} \lambda_{0,j} h_j(w) \\ &\leq F(w^*) + \sum_{j=1}^{J} \lambda_{0,j} h_j(w^*) \\ &= L(w^*, \lambda_0) \end{aligned} \tag{17}$$

When we plug in the optimal solution λ^* for Dual problem, the inequality still holds.

$$\begin{aligned} G(\lambda^*) &\leq L(w^*, \lambda^*) \\ \Rightarrow \underbrace{\max_{\lambda_j \geq 0} \min_w L(w; \lambda_j)}_{\text{Dual}} &\leq \underbrace{\min_w \max_{\lambda_j \geq 0} L(w; \lambda_j)}_{\text{Primal}} \end{aligned} \tag{18}$$

In summary,

$$\begin{aligned} \max_{\lambda_j \geq 0} \min_w L(w; \lambda_j) &= G(\lambda^*) \\ &\leq L(w^*, \lambda^*) \\ &= F(w^*) + \sum_{j=1}^{J} \lambda_j h_j(w^*) \\ &\leq \max_{\lambda_j \geq 0} F(w^*) + \sum_{j=1}^{J} \lambda_j h_j(w^*) \\ &= \min_w \max_{\lambda_j \geq 0} L(w; \lambda_j) \end{aligned} \tag{19}$$

Basically, we can know Dual problem is the lower envelope of the Primal problem. This is called weak duality, which hold for arbitrary function F and h_j.

2.5 Strong Duality

In SVM, we have strong duality. Specially, when functions F and h_j are convex and assume some other mild conditions, we can get:

$$\begin{aligned} &\max_{\lambda_j \geq 0} \min_{w} L(w; \lambda_j) = \min_{w} \max_{\lambda_j \geq 0} L(w; \lambda_j) \\ \Rightarrow &L(w^*, \lambda^*) = F(w^*) + \sum_{j=1}^{J} \lambda_j^* h_j^*(w^*) = F(w^*) \end{aligned} \tag{20}$$

When strong duality holds, it implies the following three conditions:

- Complementary slackness:
$$\forall j, \lambda_j^* h_j(w^*) = 0 \tag{21}$$
- Stationarity:
$$\bigtriangledown_w L(w^*, \lambda_j^*) = \bigtriangledown F(w^*) + \sum_{j=1}^{J} \lambda_j^* \bigtriangledown h_j(w^*) = 0 \tag{22}$$
- Feasibility:
$$h_j(w^*) \leq 0, \lambda_j^* \geq 0 \tag{23}$$

In summary, The equation 21,22 and 23 make up the KKT conditions. Sufficiency and necessity of KKT conditions for characterizing optimal solutions are as follows:

- Necessity: If w^* and λ^* are optimal solutions for primal and dual with zero duality gap (e.g., convex problem and there exists w strictly satisfying non-affine inequality constraints), then w^* and λ^* satisfy the KKT conditions.
- Sufficiency: If w^* and λ^* satisfy the KKT conditions, then w^* and λ^* are optimal solutions for primal and dual problem.