Support Vector Machine: Lagrangian Dual Formulation.

primal problem: $\min f(w,b) = \frac{1}{2}\|w\|^2$

(P) $\quad$ s.t. $y_i(w^T x_i + b) - 1 \geq 0$

Lagrangian dual: $\max g(\lambda) = \inf_{w,b} L(w,b,\lambda) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \lambda_i (y_i(w^T x_i + b) - 1)$

(D) $\quad$ s.t. $\lambda_i \geq 0$

$y_i(w^T x_i + b) - 1 \geq 0.$

$g(\lambda)$ is the minimum attainable function value of L on the space (w, b)

For most convex optimization problems, the primal problem reaches its minimal when the dual problem reaches its maximal. This is the so called "strict duality". See Chap. 5 of "Convex Optimization" for proof details.

Strict duality implies that assume we found optimal w^* and b^* for (P), and optimal λ^* for (D). we have $g(\lambda^*) = f(w^*, b^*) = L(w^*, b^*, \lambda^*)$

Therefore, $\sum_{i=1}^{n} \lambda_i^* (y_i (w^{*T} x_i + b^*) - 1) = 0$

Due to the nonnegativity, $\lambda_i (y_i (w^T x_i + b) - 1) = 0 \quad \forall i = 1, \cdots, n$ holds for optimal λ^*, w^*, b^*. This is the so called "complementary slackness". Note that this is very important for deriving b^*!

Now let's sum up the conditions you need to meet for optimality:

$y_i(w^T x_i + b) - 1 \geq 0 \quad \forall i = 1, \cdots, n$ → original constraints

$\lambda_i \geq 0. \quad \forall i = 1, \cdots, n$ → Lagrangian assumption

$\lambda_i (y_i (w^T x_i + b) - 1) = 0 \quad \forall i = 1, \cdots, n$ → complementary slackness

$$\left.\begin{aligned} \frac{\partial L(w,b,\lambda)}{\partial w} = 0 &\Rightarrow w = \sum_{i=1}^{n} \lambda_i y_i x_i \\ \frac{\partial L(w,b,\lambda)}{\partial b} = 0 &\Rightarrow \sum_{i=1}^{n} \lambda_i y_i = 0 \end{aligned}\right\} \rightarrow \text{optimality assumption.}$$

The 5 conditions above form the well-known KKT conditions.

Now, let's solve the λ. By making (D) reach its maximal we get (P) reaching its minimal. Inserting $w = \sum_{i=1}^{n} \lambda_i y_i x_i$ and $\sum_{i=1}^{n} \lambda_i y_i = 0$ back to $g(\lambda)$ we formulate (D) as:

$$\max g(\lambda) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i^T x_j$$

$$\text{s.t.} \quad \lambda_i \geq 0 \quad \forall i = 1, \cdots, n$$

$$\sum_{i=1}^{n} \lambda_i y_i = 0$$

This is a Quadratic programming problem with constraints. Many modern solvers can be used to solve it (e.g., Gurobi).

Let's say we get optimal λ^*. From optimality condition we can get

$$w^* = \sum_{i=1}^{n} \lambda_i^* y_i x_i$$

From complementary slackness, we can find out the data samples x_i that has $\lambda_i^* > 0$, and derive b^* by:

$$b^* = y_i - w^{*T} x_i \quad (\text{for } x_i \text{ with } \lambda_i > 0)$$

Some follow-up notations:

1°. How do we use w^* and b^*?

We can use $w^{*T} x + b^*$ for inference. Given a new sample x with unknown label, we use $y = w^{*T} x + b^*$, if $y > 0$ x belongs to class $+1$, vice versa.

2°. kernel trick.

Both the optimization objective and the inference contain the dot product $x_i^T x_j$, that's why we only care about the dot product of two feature vectors and why we can directly define the transformation outcome as kernel functions.

3°. what are the support vectors in soft margin SVM?

The Lagrangian deriviation of soft Margin SVM is very similar to hard margin SVM. I highly recommend to do it yourself if you are interested.

When you obtain KKT conditions for soft margin SVM, complementary slackness would tell you:

$$\lambda_i (y_i (w^T x_i + b) - 1 + \xi_i) = 0. \quad \text{where } \xi_i \text{ is the slack variable.}$$

It means if $\lambda_i > 0$, $y_i (w^T x_i + b) = 1 - \xi_i$ equality holds. Therefore, those data samples that are both on the margin and misclassified would have influence on w^*. This means the final decision boundary is determined by both margin data samples and wrongly classified samples.