Advanced Machine Learning **Fall 2020**

Lecture 7: Support Vector Machines

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes are adapted from CMU's 10-725 Course, Stanford's CS229 Course, ETH's Advanced Machine Learning Course and Bishop's "Pattern Recognition and Machine Learning" book.*

7.1 Lagrangian

Consider a general minimization problem. There is no need for it to be convex, let's keep it as general as possible.

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0, \ i = 1, \ldots, m \\ & l_j(x) = 0, \ j = 1, \ldots, r \end{aligned}$$

The Lagrangian is defined as:

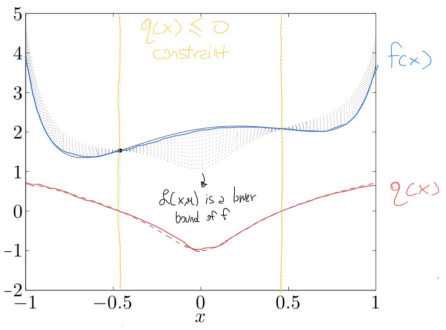
$$\mathcal{L}(x, u, v) = f(x) + \sum_{i=1}^{m} u_i \cdot h_i(x) + \sum_{j=1}^{r} v_j \cdot l_j(x) \tag{7.1}$$

Where $u \in \mathbb{R}^m \geq 0$, $v \in \mathbb{R}^r$.

(Small side-note) We define the Lagrangian $\mathcal{L}(x, u, v) = -\infty$ for values of $u < 0$.

We are going to exploit the following property of the Lagrangian:
$\forall u \geq 0$ and $\forall v : \ f(x) \geq \mathcal{L}(x, u, v)$ **at each feasible** x

Figure 7.1: Each dotted line shows $\mathcal{L}(x, u)$ for different choice of $u \geq 0$.



7.1.1 Lagrangian Dual Function

Let C denote the primal feasible set, f^* denote primal optimal value. Minimizing $\mathcal{L}(x, u, v)$ over all x gives a lower bound:

$$f^* \overset{(i)}{\geq} \min_{x \in C} \quad \mathcal{L}(x, u, v) \overset{(ii)}{\geq} \min_{x} \quad \mathcal{L}(x, u, v) := g(u, v)$$

Where (i) derives from the property we stated earlier taking the minimum of both sides and (ii) holds because the unconstrained minimum will always be less or equal the constrained one.

We call $g(u, v)$ the Lagrangian dual function and it gives a lower bound on f^*. The main takeaway here is that (i) is often not computable because of the constraints. Hence, we prefer to solve the dual problem (ii).

7.1.2 Langrange Dual Problem

Given a primal problem

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0, \ i = 1, \ldots, m \\ & l_j(x) \equiv 0, \ j = 1, \ldots, r \end{aligned}$$

We have showed that our dual function satisfies $f^* \geq g(u, v)$ for all $u \geq 0$ and v
Thus, we can get the best lower-bound estimate of f^* maximizing $g(u, v)$ over the feasible u, v, yielding the Lagrange Dual Problem:

$$\begin{aligned} \max_{u,v} \quad & g(u, v) \\ \text{subject to} \quad & u \geq 0, \end{aligned}$$

A key property is called **weak duality**:

$$f^* \geq g^*$$

Where $f^*, g*$ are the optimal values for the primal and dual problems.
Note that this property always holds, even if the primal problem is not convex. Furthermore, it is easy to prove that **the dual problems is always a convex optimization problem**, even if the primal problem is non convex.

7.1.3 Strong Duality

In some problems we will have $f^* = g^*$, this property is called Strong Duality.

Theorem 7.1 *(Slater's Condition)*
If the primal problem is a convex problem and there exists at least one strictly feasible $x \in \mathbb{R}$, then Strong Duality holds.

In other words, the condition is that exists x such that:

$$h_i(x) < 0,\ i = 1, \ldots, m$$

$$l_j(x) = 0,\ j = 1, \ldots, r$$

7.1.4 Karush-Kuhn-Tucker (KKT) Conditions

Given a general problem

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0,\ i = 1, \ldots, m \\ & l_j(x) = 0,\ j = 1, \ldots, r \end{aligned}$$

The KKT conditions are:

1. $0 \in \partial_x(\mathcal{L}(x, u, v))$ (stationarity)
2. $u_i \cdot h_i(x) = 0$ for all i (complementary slackness)
3. $h_i(x) \leq 0, l_j(x) = 0$ for all i, j (primal feasibility)
4. $u_i \geq 0$ for all i (dual feasibility)

Theorem 7.2 *For x^* and u^*, v^* to be primal and dual solutions, KKT conditions are sufficient.*

Proof: $g^* = g(u^*, v^*) = f(x^*) + \sum_{i=1}^{m} u_i^* \cdot h_i(x^*) + \sum_{j=1}^{r} v_j^* \cdot l_j(x^*) = f(x^*) = f^*$

Where the first equality holds from stationarity and the second equality holds from complementary slackness and primal feasibility. ■

Theorem 7.3 *For a problem with strong duality (e.g. assume Slater's condition holds)*
x^* and u^*, v^* are primal and dual solution $\iff$ x^* and u^*, v^* satisfy KKT conditions.

Proof:

Sufficiency: Follows from Theorem 7.2.

Necessity: Suppose x^* and u^*, v^* to be primal and dual solution, and suppose strong duality holds. Then:

$f^* = g^* = g(u^*, v^*)$ (holds by assumptions)

$= \min_x (f(x) + \sum_{i=1}^m u_i^* h_i(x) + \sum_{j=1}^r v_j^* \cdot l_j(x))$ (holds by definition)

$\leq f(x^*) + \sum_{i=1}^m u_i^* \cdot h_i(x^*) + \sum_{j=1}^r v_j^* \cdot l_j(x^*)$ (min(f) is less or equals than the value of f at any other point)

$\leq f(x^*)$ (holds by feasibility since the sums must be less or equal than 0)

The LHS equals RHS, therefore all the inequalities must be equalities. Looking at KKT Conditions:

- **Primal and dual feasibility** hold by virtue of optimality: x^*, u^*, v^* are optima $\implies$ x^*, u^*, v^* must be feasible
- **Stationarity** comes from the fact that x^* minimizes $g(u^*, v^*)$. Since x^* is the minimizer it must be a stationary point for this function.
- **Complementary Slackness** comes from the last inequality, since $\sum_{i=1}^m u_i^* \cdot h_i(x^*)$ must be equal to 0.

■

7.2 Maximum Margin Classifiers

We begin our discussion of Support Vector Machines by returning to the two-class classification problem using a linear model of the form:

$$y(x) = w^T \phi(x) + b$$

We shall assume for the moment that the training data is linearly separable. The SVMs approach this problem through the concept of margin, which is defined to be the smallest distance between decision boundary and any of the samples.

7.2.1 Finding the margin

Consider an arbitrary point x and let $x_\perp$ be its orthogonal projection onto the decision surface, so that:

$$x = x_\perp + r\frac{w}{\|w\|}$$

Multiplying both sides of this result by w^T and adding b we get:

$$w^T x + b = w^T x_\perp + r w^T \frac{w}{\|w\|} + b$$

Applying the definition $y(x) = w^T x + b$:

$$y(x) = y(x_\perp) + r\,\|w\|$$

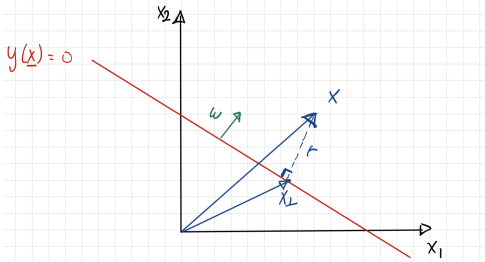
From Figure 7.2 it is clear that $x_\perp$ lies on the decision surface, hence $y(x_\perp) = 0$. Solving for r:

$$r = \frac{y(x)}{\|w\|}$$

Therefore, the perpendicular distance of a point x from a hyperplane defined by $y(x) = 0$ is given by:

$$\frac{|y(x)|}{\|w\|}$$

Figure 7.2: The decision surface shown in red is perpendicular to w.



7.2.2 SVMs Primal Problem

We wish to optimize the parameters w and b in order to maximize the minimum margin among all data points:

$$\max_{w,b} \min_{n} \frac{|y(x_n)|}{\|w\|}$$

We can take the factor $\frac{1}{\|w\|}$ outside of the optimization over n because it does not depend on n:

$$\max_{w,b} \frac{1}{\|w\|} \min_{n}[t_n(w^T\phi(x) + b)]$$

Direct solution of this optimization problem would be very complex (non-convex), and so we shall convert it into an equivalent problem that is much easier to solve.

To do this, we note that if we make the rescaling $w' = kw$ and $b' = kb$ the margin will remain unchanged:

$$r' = \frac{t_n((w')^T\phi(x_n) + b')}{\|w'\|} = \frac{kt_n(w^T\phi(x_n) + b)}{k\,\|w\|} = r$$

We can use this freedom to set the point that is closer to the decision surface:

$$t_n(w^T\phi(x_n) + b) = 1$$

In this case all data points will have to satisfy the constraints:

$$t_i(w^T\phi(x_i) + b) \geq 1 \quad i = 1, \ldots, N$$

Thus we can reduce the problem to:

$$\begin{aligned} \max_{w,b} \quad & \frac{1}{\|w\|} \\ \text{subject to} \quad & 1 - t_i(w^T\phi(x_i) + b) \leq 0,\ i = 1, \ldots, N \end{aligned}$$

Furthermore, maximizing $\|w\|^{-1}$ is equivalent to minimizing $\|w\|^2$.We include a factor $1/2$ for later convenience:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & 1 - t_i(w^T\phi(x_i) + b) \leq 0,\ i = 1, \ldots, N \end{aligned}$$

Two important observations:

- It appears that the bias parameter b has disappeared from the optimization. However, it is determined implicitly via the constraints.
- The solution to a QP problem in M variables has computational complexity that is $O(M^3)$. Thus, the primal problem is only feasible if we constrain ourselves to a fixed set of basis function (small M).

7.2.3 SVMs Dual Problem

First thing first, we should check if Slater's conditions is satisfied (remember that Slater's condition is sufficient for **strong duality**).
Let w^*, b^* be an optimal solution and $\lambda > 1$, then $\lambda w^*, \lambda b^*$ will be strictly feasible:

$$t_i(\lambda w^{*T}\phi(x_i) + \lambda b^*) = t_i\lambda(w^{*T}\phi(x_i) + b^*) > t_i(w^{*T}\phi(x_i) + b^*) \geq 1,\ i = 1, \ldots, N$$

Since strong duality holds, we can solve the dual problem:

$$\begin{aligned} &\max_{a} && \min_{w,b} \mathcal{L}(w, b, a) \\ &\text{subject to} && a_i \geq 0,\ i = 1, \ldots, N \end{aligned}$$

Where we define the Lagrangian Dual Function $\mathcal{L}(w, b, a) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{N} a_i\{1 - t_i(w^T\phi(x_i) + b)\}$

Setting the derivatives of $\mathcal{L}(w, b, a)$ with respect to w and b equal to zero, we obtain the following two conditions:

$$w = \sum_{i=1}^{N} a_i t_i \phi(x_i) \tag{7.2}$$

$$\sum_{i=1}^{N} a_i t_i = 0 \tag{7.3}$$

Now we substitute the conditions back into $\mathcal{L}(w, b, a)$:

$$\mathcal{L}(w, b, a) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{N} a_i - \overbrace{\sum_{i=1}^{N} a_i t_i b}^{=0} - \sum_{i=1}^{N} a_i t_i w^T \phi(x_i)$$

Substituting (7.2) we obtain:

$$\mathcal{L}(a) = \frac{1}{2}\,[\sum_{i=1}^{N} a_i t_i \phi(x_i)]^T [\sum_{i=1}^{N} a_i t_i \phi(x_i)] + \sum_{i=1}^{N} a_i - \sum_{i=1}^{N} a_i t_i [\sum_{i=1}^{N} a_i t_i \phi(x_i)]^T \phi(x_i)$$

Expanding the products:

$$\mathcal{L}(a) = \frac{1}{2}\ \sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j t_i t_j \phi(x_i)^T \phi(x_j) + \sum_{i=1}^{N} a_i - \sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j t_i t_j \phi(x_i)^T \phi(x_j)$$

This gives the dual representation of the SVM problem:

$$\begin{aligned} \max_{a} \quad & \sum_{i=1}^{N} a_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j t_i t_j \phi(x_i)^T \phi(x_j) \\ \text{subject to} \quad & a_i \geq 0,\ i = 1, \ldots, N \\ & \sum_{i=1}^{N} a_i t_i = 0, \end{aligned}$$

Some important observations:

- Note that now the time complexity for the QP solver is $O(N^3)$, thus it does not depend on the choice of basis function. For a fixed set of basis functions whose number M is smaller than the number N of data points, the dual problem appears disadvantageous. However, the dual problem makes feasible applying SVMs to feature spaces whose dimensionality exceed the number of data points, including infinite feature spaces.
- In order to classify new data points using the trained model, we evaluate the sign of $y(x)$:

$$y(x) = \sum_{i=1}^{N} a_i t_i \phi(x_i)^T \phi(x) + b$$

- Remember that : Strong Duality $\implies$ KKT conditions are satisfied. Hence, the following condition must hold:

$$a_i(t_i y(x_i) - 1) = 0,\ i = 1, \ldots, N$$

  Any data point for which $a_i = 0$ plays no role in making predictions for new data points. The remaining data points correspond to points that lie on the maximum margin hyperplanes in feature space and they are called support vectors. This property is central to the practical applicability of SVMs: once the model is trained, a significant proportion of the data points can be discarded.