CSC 665: Support Vector Machines

Chicheng Zhang

October 3, 2019

1 Support vector machines - the maximum margin hyperplane problem

We consider linear classification, where examples $(x_i, y_i)_{i=1}^m$ are such that $x_i \in \mathbb{R}^d$ are features, and $y_i \in \{\pm 1\}$ are binary labels.

Suppose that the training set $S = (x_i, y_i)_{i=1}^m$ is linearly separable, i.e. there exists a linear classifier $(w, b) \in \mathbb{R}^{d+1}$, such that for all i,

$$\begin{cases} \langle w, x_i \rangle + b > 0 & y_i = +1, \\ \langle w, x_i \rangle + b < 0 & y_i = -1. \end{cases} \tag{1}$$

One way to train a linear classifier would be to use the consistency algorithm, i.e. solving a linear program, that finds a (w, b) such that Equation (1) holds. However, note that not all consistent linear classifiers are created equal: some of them are closer to training examples than others. Formally, the distance of a point x in $\mathbb{R}^d$ to a hyperplane $H_{w,b} = \{x_0 : w \cdot x_0 + b = 0\}$ is defined as the shortest distance of x to any of the points in $H_{w,b}$:

$$d(x, H_{w,b}) = \min \left\{ \|x - x_0\| : w \cdot x + b = 0 \right\}. \tag{2}$$

Can we calculate this distance analytically? First, let us assume without loss of generality that $\|w\| = 1$, as any hyperplane $H_{w',b'}$ can be written as $H_{w,b}$ for $\|w\| = 1$ by letting $w = \frac{w'}{\|w'\|}$ and $b = \frac{b'}{\|w'\|}$. Now, consider a point $x_0 \in H_{w,b}$ such that $x_0 = x + \alpha w$ for some α. What is the value of α? Note that

$$\langle w, x + \alpha w \rangle + b = 0,$$

which implies that $\alpha = -(\langle w, x \rangle + b)$.

Claim 1. *For all x_1 in $H_{w,b}$,*

$$\|x_1 - x\| \geq \|x_0 - x\|. \tag{3}$$

Consequenty, $d(x, H_{w,b}) = |\langle w, x \rangle + b|$.

Proof. Note that x_0 and x_1 are both in $H_{w,b}$, $\langle w, x_0 \rangle + b = \langle w, x_1 \rangle + b = 0$. Therefore, $\langle x_1 - x_0, w \rangle = 0$. In other words,

$$\langle x_1 - x_0, x_0 - x \rangle = 0.$$

Now, by Pythagorean theorem,

$$\|x - x_1\|^2 = \|x - x_0\|^2 + \|x_0 - x_1\|^2 \geq \|x - x_0\|^2,$$

whch proves Equation (3). This implies that

$$d(x, H_{w,b}) = \|x - x_0\| = |\langle w, x \rangle + b|.$$

□

Here is a proposal:

> Find the linear classifier (w, b) that not only separates the examples but also maximizes the minimum distances to all examples.

Why is the proposal sensible? One observation is that this classifier is the most "robust". For example, if test examples happen to be just a little distance away from training examples (with the same labels), then this classifier would still classify such examples correctly.

Formally, we can describe the proposal as an optimization problem:

$$\begin{aligned} \underset{w,b,A}{\text{maximize}} \quad & A \\ \text{s.t.} \quad & A > 0, \quad \|w\| = 1, \\ & y_i(\langle w, x_i \rangle + b) > 0, && \forall i \in \{1, \dots, n\}, \\ & |\langle w, x_i \rangle + b| \geq A, && \forall i \in \{1, \dots, n\}. \end{aligned} \tag{4}$$

The above program is not a convex program, and is difficult to optimize directly. Let's make a few transformations to make it a convex program - i.e. finding a convex optimization problem whose solution is related to that of the above optimization problem.

Let's consider the following optimization problem:

$$\begin{aligned} \underset{w,b,A}{\text{maximize}} \quad & A \\ \text{s.t.} \quad & A > 0, \quad \|w\| = 1, \\ & y_i(\langle w, x_i \rangle + b) \geq A, && \forall i \in \{1, \dots, n\}, \end{aligned} \tag{5}$$

Our claim is that the above two optimization problems have the same solutions. Why? Because under $A > 0$, constraints $y_i(\langle w, x_i \rangle + b) > 0$ and $|\langle w, x_i \rangle + b| \geq A$, together, are equivalent to $y_i(\langle w, x_i \rangle + b) \geq A$, as $y_i \in \{\pm 1\}$. For every i, the quantity $y_i(\langle w, x_i \rangle + b)$ is the *margin* of halfspace $H_{w,b}$ on example (x_i, y_i). Therefore the above is also called the "maximum margin hyperplane" problem.

Now let $w' = \frac{w}{A}$, $b' = \frac{b}{A}$. Note that the above optimization problem is equivalent to

$$\begin{aligned} \underset{w',b',A}{\text{maximize}} \quad & A \\ \text{s.t.} \quad & A > 0, \quad \|w'\| = \frac{1}{A}, \\ & y_i(\langle w', x_i \rangle + b') \geq 1, && \forall i \in \{1, \dots, n\}, \end{aligned}$$

Furthermore, this is equivalent to

$$\begin{aligned} \underset{w',b'}{\text{minimize}} \quad & \|w'\| \\ \text{s.t.} \quad & y_i(\langle w', x_i \rangle + b') \geq 1, && \forall i \in \{1, \dots, n\}, \end{aligned}$$

As the function $x \mapsto \frac{1}{2}x^2$ is monotonically increasing for $x > 0$, we get that the above is equivalent to

$$\begin{aligned} \underset{w',b'}{\text{minimize}} \quad & \frac{1}{2}\|w'\|^2 \\ \text{s.t.} \quad & y_i(\langle w', x_i \rangle + b') \geq 1, && \forall i \in \{1, \dots, n\}, \end{aligned} \tag{6}$$

Optimization problem (4) is called the *support vector machine* (SVM). Note that its constraints are all linear inequalities, which defines a convex constraint set. In addition, its optimization objective is a quadratic function over optimization variables, which is a convex function. This implies that it is a *convex optimization* problem.

Recovering the optimal solution of (4)**.** Suppose we have a solution of (6), written as $(w'^\star, b'^\star)$. Note that the optimal A in (5) (thus, in (4)) is $1/\|w'^\star\|$, which is the value of the minimum margin. This implies that in (5) (thus, in (4)), $w^\star = A^\star w'^\star = \frac{w'^\star}{\|w'^\star\|}$, $b^\star = A^\star b'^\star = \frac{b'^\star}{\|w'^\star\|}$. The optimal hyperplane is simply $H_{w^\star,b^\star} = H_{w'^\star,b'^\star}$.

Optimality condition. To avoid notation clutter, let us drop the apostrophes in optimization problem (6):

$$\begin{aligned} \underset{w,b}{\text{minimize}} \quad & \frac{1}{2}\|w\|^2 \\ \text{s.t.} \quad & y_i(\langle w, x_i\rangle + b) \geq 1, \qquad \forall i \in \{1,\dots,n\}, \end{aligned} \tag{7}$$

What property does the optimal solution $(w^\star, b^\star)$ have? We will take a detour and first discuss Lagrangian duality, a fundamental concept in constrained optimization. Let us first write (7) as an unconstrained optimization problem over a slightly more complicated objective:

$$\min_{w,b,\xi} \max_{\alpha \geq 0} L(w, b, \xi; \alpha), \tag{8}$$

where $L(w, b, \xi; \alpha) = \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i(1 - y_i(\langle w, x_i\rangle + b))$.

Define $F(w, b, \xi) := \max_{\alpha \geq 0} L(w, b, \xi; \alpha)$. Observe that:

$$F(x, b, \xi) = \begin{cases} -\infty, & \exists i, 1 - y_i(\langle w, x_i\rangle + b) < 0 \\ \frac{\lambda}{2}\|w\|^2, & \forall i, 1 - y_i(\langle w, x_i\rangle + b) \geq 0 \end{cases}$$

Therefore, optimization problem (8) is equivalent to (7). Now consider switching the orders of min and max in (8):

$$\max_{\alpha \geq 0} \min_{w,b,\xi} L(w, b, \xi; \alpha).$$

This is called the dual problem of (8) ((8) is therefore called the primal problem). Let's call the optimal primal objective $p^\star$ and the optimal dual objective $d^\star$ It is well known that for general function L, $p^\star \geq d^\star$. However, under fairly general assumptions, it can also be shown that $p^\star = d^\star$.

We state the following result from numerical optimization. Consider a constrained convex optimization problem that has both equality and inequality constraints:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0, \qquad \forall i \in \{1,\dots,n\}, h_i(x) = 0, \qquad \forall i \in \{1,\dots,m\}, \end{aligned}$$

Similar as before, we can define Lagrange function $L(x, \alpha, \beta) = f(x) + \sum_{i=1}^n \alpha_i g_i(x) + \sum_{i=1}^m \beta_i h_i(x)$. Define $p^\star = \min_x \max_{\alpha \geq 0, \beta} L(x, \alpha, \beta)$, and $d^\star = \max_{\alpha \geq 0, \beta} \min_x L(x, \alpha, \beta)$, we have the following result.

Theorem 1. *Under mild assumptions* [1]*, we have that there exists $x^\star$, $\alpha^\star$, and β, such that*

$$p^\star = L(x, \alpha, \beta) = d^\star.$$

Moreover, the following set of equations (called Karush-Kuhn-Tucker (KKT) condition) is true:

$$\begin{aligned} \nabla_x L(x^\star, \alpha^\star, \beta^\star) &= 0 \\ g_i(x^\star) &\leq 0 \\ h_i(x^\star) &\leq 0 \\ \alpha_i &\geq 0 \\ \alpha_i g_i(x^\star) &= 0. \end{aligned}$$

[1] in particular, the Slater condition, that is, there exists w, b, ξ such that all inequality constraints in are stictly satisfied

Applying the theorem to SVM optimization, we immediately have that the dual optimization problem has the same optimal value as the primal optimization problem. In addition, in SVM, we can also recover the optimal solution from dual solution by invoking the KKT condition! To see why, note that in SVM,

$$\nabla_w L(w, b, \xi; \alpha) = \lambda w - \sum_{i=1}^{n} \alpha_i y_i x_i,$$

which implies that

$$w^\star - \sum_{i=1}^{n} \alpha_i^\star y_i x_i.$$

1.1 The soft-margin SVM

Can we still train SVM if the data is not linearly separable? Note that optimization problem (6) will not find a solution, as now the constraint set become infeasible.

We introduce slack variables $\xi_i \geq 0$ for every example i, to allow some example to be misclassified. In addition, we introduce a regularization parameter $\lambda > 0$ that trades off misclassification and margin on correct examples:

$$\begin{aligned} \underset{w,b,\xi}{\text{minimize}} \quad & \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{n} \xi_i \\ \text{s. t.} \quad & y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i, \qquad \forall i \in \{1, \ldots, n\}, \xi_i \geq 0, \qquad \forall i \in \{1, \ldots, n\}, \end{aligned} \tag{9}$$

Intuitively, when λ is larger, it focuses more on enforcing large margin on correct examples; when λ is smaller, it forces more on reducing misclassification. Notice that the last two lines can be summarized by: $\forall i \in \{1, \ldots, n\}$, $\xi_i \geq \max(0, 1 - y_i(\langle w, x_i\rangle + b))$. Therefore, the optimal choice of ξ_i equals $\max(0, 1 - y_i(\langle w, x_i\rangle + b))$. Let $\phi(z) = \max(0, 1 - z)$ and $R(w) = \frac{\lambda}{2}\|w\|^2$. We thus can rewrite optimization problem (9) as:

$$\underset{w,b}{\text{minimize}} \quad \lambda R(w) + \sum_{i=1}^{n} \phi(y_i(\langle w, x_i\rangle + b)). \tag{10}$$

As a convention, we call $\phi(y(\langle w, x\rangle + b))$ the *hinge loss* of linear classifier (w, b) on example (x, y), written as $\ell_{\text{hinge}}((w, b), (x, y))$. When the margin $y(\langle w, x\rangle + b)$ is larger, the hinge loss is smaller. The above form is also called a *regularized loss minimzation* formulation, which captures a wide range of optimization problems in machine learning (by changing loss function ϕ and regularizer R), such as logistic regression, ridge regression, lasso, etc.

2 The dual of SVM

Let's consider writing the constrained optimization problem (9) in the following alternative way:

$$\underset{w,b,\xi}{\text{minimize}} \max_{\alpha \geq 0, \beta \geq 0} \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i(1 - \xi_i - y_i(\langle w, x_i\rangle + b)) + \sum_{i=1}^{n} \beta_i(-\xi_i). \tag{11}$$

The α_i, β_i's are called Lagrangian multipliers. Why is the above equivalent to the original soft margin SVM?

Define $F(w, b, \xi) = \max_{\alpha \geq 0, \beta \geq 0} \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i(1 - \xi_i - y_i(\langle w, x_i\rangle + b)) + \sum_{i=1}^{n} \beta_i(-\xi_i)$. Observe that

$$F(w, b, \xi) = \begin{cases} -\infty & \exists i, y_i(\langle w, x_i\rangle + b)) < 1 - \xi_i \\ -\infty & \exists i, \xi_i < 0 \\ \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{n} \xi_i & \text{otherwise.} \end{cases}$$

Every constrained convex optimization problem has a dual optimization problem. Sometimes looking at the dual problem will yield unexpected insights about the original (primal) problem! Indeed, SVM is a canonical example for this claim.

We consider the soft-margin SVM formulation (9). To derive its dual problem, we introduce dual variables (aka Lagrange multipliers) to incorporate all constraints to the objective function:

$$\min_{w,b,\xi\geq 0}\max_{\alpha\geq 0,\beta\geq 0}\left(\frac{\lambda}{2}\|w\|^2+\sum_{i=1}^{n}\xi_i+\sum_{i=1}^{n}\alpha_i\cdot(1-\xi_i-y_i(\langle w,x_i\rangle+b))-\sum_{i=1}^{n}\beta_i\xi_i\right)$$

It can be shown that the above is equivalent to

$$\max_{\alpha\geq 0,\beta\geq 0}\min_{w,b,\xi\geq 0}\left(\frac{\lambda}{2}\|w\|^2+\sum_{i=1}^{n}\xi_i+\sum_{i=1}^{n}\alpha_i\cdot(1-\xi_i-y_i(\langle w,x_i\rangle+b))-\sum_{i=1}^{n}\beta_i\xi_i\right)$$

Note that the inner minimization can be written as follows:

$$\sum_{i=1}^{n}\alpha_i+\min_{w}\left(\frac{\lambda}{2}\|w\|^2-\left\langle w,\sum_{i=1}^{n}\alpha_i y_i x_i\right\rangle\right)+\min_{b}(-b\cdot\sum_{i=1}^{n}y_i\alpha_i)+\sum_{i=1}^{n}\min_{\xi_i\geq 0}\xi_i(1-\alpha_i-\beta_i)$$

Let us denote by $h(z)=\begin{cases}-\infty & z<0\\ 0 & z\geq 0\end{cases}$ and $g(z)=\begin{cases}-\infty & z=0\\ 0 & z\neq 0\end{cases}$. The optimal value of the inner optimization problem is

$$\sum_{i=1}^{n}\alpha_i-\frac{1}{2\lambda}\|\sum_{i=1}^{n}\alpha_i y_i x_i\|^2+g(\sum_{i=1}^{n}y_i\alpha_i)+\sum_{i=1}^{n}h(1-\alpha_i-\beta_i).$$

Therefore, the dual problem is equivalent to:

$$\begin{aligned}\underset{\alpha,\beta}{\text{maximize}}\quad & \sum_{i=1}^{n}\alpha_i-\frac{1}{2\lambda}\|\sum_{i=1}^{n}\alpha_i y_i x_i\|^2 && && (12)\\ \text{s.t.}\quad & \alpha_i+\beta_i\leq 1, && \forall i\in\{1,\ldots,n\},\\ & \sum_{i=1}^{n}y_i\alpha_i=0, && \forall i\in\{1,\ldots,n\},\end{aligned}$$

3 The kernel trick