

Support Vector Machines

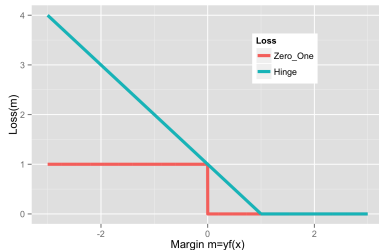
David Rosenberg

New York University

October 29, 2016

Support Vector Machine

- Hypothesis space $\mathcal{F} = \{f(x) = w^T x + b \mid w \in \mathbb{R}^d, b \in \mathbb{R}\}$.
- ℓ_2 regularization (Tikhonov style)
- Loss $\ell(m) = (1 - m)_+$
 - Margin $m = yf(x)$; “Positive part” $(x)_+ = x1(x \geq 0)$.



SVM Optimization Problem

The SVM prediction function is the solution to

$$\min_{w \in \mathbf{R}^d, b \in \mathbf{R}} \frac{1}{2} \|w\|^2 + \frac{c}{n} \sum_{i=1}^n (1 - y_i [w^T x_i + b])_+.$$

- unconstrained optimization
- not differentiable
- Can we reformulate into a differentiable problem?

SVM Optimization Problem

- The SVM optimization problem is equivalent to

$$\begin{aligned} &\text{minimize} && \frac{1}{2} \|w\|^2 + \frac{c}{n} \sum_{i=1}^n \xi_i \\ &\text{subject to} && \xi_i \geq (1 - y_i [w^T x_i + b])_+, \end{aligned}$$

- Which is equivalent to

$$\begin{aligned} &\text{minimize} && \frac{1}{2} \|w\|^2 + \frac{c}{n} \sum_{i=1}^n \xi_i \\ &\text{subject to} && \xi_i \geq 0 \text{ for } i = 1, \dots, n \\ &&& \xi_i \geq (1 - y_i [w^T x_i + b]) \text{ for } i = 1, \dots, n \end{aligned}$$

SVM as a Quadratic Program

- The SVM optimization problem is equivalent to

$$\begin{aligned}
 &\text{minimize} && \frac{1}{2} \|w\|^2 + \frac{c}{n} \sum_{i=1}^n \xi_i \\
 &\text{subject to} && \xi_i \geq 0 \text{ for } i = 1, \dots, n \\
 &&& \xi_i \geq (1 - y_i [w^T x_i + b]) \text{ for } i = 1, \dots, n
 \end{aligned}$$

- Differentiable objective function
- $n + d + 1$ unknowns and $2n$ affine constraints.
- A quadratic program that can be solved by any off-the-shelf QP solver.
- Let's learn more by examining the dual!

SVM Lagrangian

- The Lagrangian for this formulation is

$$\begin{aligned}
 & L(w, b, \xi, \alpha, \lambda) \\
 = & \frac{1}{2} \|w\|^2 + \frac{c}{n} \sum_{i=1}^n \xi_i + \sum_{i=1}^n \alpha_i (1 - y_i [w^T x_i + b] - \xi_i) - \sum_i \lambda_i \xi_i \\
 = & \frac{1}{2} w^T w + \sum_{i=1}^n \xi_i \left(\frac{c}{n} - \alpha_i - \lambda_i \right) + \sum_{i=1}^n \alpha_i (1 - y_i [w^T x_i + b]).
 \end{aligned}$$

- Primal and dual:

$$\begin{aligned}
 p^* &= \inf_{w, \xi, b} \sup_{\alpha, \lambda \succeq 0} L(w, b, \xi, \alpha, \lambda) \\
 &\geq \sup_{\alpha, \lambda \succeq 0} \inf_{w, b, \xi} L(w, b, \xi, \alpha, \lambda) = d^*
 \end{aligned}$$

- Do we have $p^* = d^*$?

Strong Duality by Slater's constraint qualification

- The SVM optimization problem is equivalent to

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} \|w\|^2 + \frac{c}{n} \sum_{i=1}^n \xi_i \\ \text{subject to} \quad & \xi_i \geq 0 \text{ for } i = 1, \dots, n \\ & \xi_i \geq (1 - y_i [w^T x_i + b]) \text{ for } i = 1, \dots, n \end{aligned}$$

- Affine constraints \implies strong duality iff problem is feasible
- Constraints are satisfied by $w = b = 0$ and $\xi_i = 1$ for $i = 1, \dots, n$,
 - so **we have strong duality** \implies

$$\begin{aligned} p^* &= \inf_{w, \xi, b} \sup_{\alpha, \lambda \geq 0} L(w, b, \xi, \alpha, \lambda) \\ &= \sup_{\alpha, \lambda \geq 0} \inf_{w, b, \xi} L(w, b, \xi, \alpha, \lambda) = d^* \end{aligned}$$

SVM Dual Function

- Lagrange dual is the inf over primal variables of the Lagrangian:

$$\begin{aligned}
 g(\alpha, \lambda) &= \inf_{w, b, \xi} L(w, b, \xi, \alpha, \lambda) \\
 &= \inf_{w, b, \xi} \left[\frac{1}{2} w^T w + \sum_{i=1}^n \xi_i \left(\frac{c}{n} - \alpha_i - \lambda_i \right) + \sum_{i=1}^n \alpha_i (1 - y_i [w^T x_i + b]) \right]
 \end{aligned}$$

- Note: $g(\alpha, \lambda) = -\infty$ when $\frac{c}{n} - \alpha_i - \lambda_i \neq 0$. (send $\xi_i \rightarrow \pm\infty$)
- Function $(w, \xi) \mapsto L(w, b, \xi, \alpha, \lambda)$ is convex and differentiable.
- Thus optimal point iff $\partial_w L = 0 \partial_b L = 0 \partial_\xi L = 0$

SVM Dual Function: First Order Conditions

- Lagrange dual function is the inf over primal variables of L :

$$g(\alpha, \lambda) = \inf_{w, b, \xi} L(w, b, \xi, \alpha, \lambda)$$

$$= \inf_{w, b, \xi} \left[\frac{1}{2} w^T w + \sum_{i=1}^n \xi_i \left(\frac{c}{n} - \alpha_i - \lambda_i \right) + \sum_{i=1}^n \alpha_i (1 - y_i [w^T x_i + b]) \right]$$

$$\partial_w L = 0 \iff w - \sum_{i=1}^n \alpha_i y_i x_i = 0 \iff \boxed{w = \sum_{i=1}^n \alpha_i y_i x_i}$$

$$\partial_b L = 0 \iff - \sum_{i=1}^n \alpha_i y_i = 0 \iff \boxed{\sum_{i=1}^n \alpha_i y_i = 0}$$

$$\partial_{\xi_i} L = 0 \iff \frac{c}{n} - \alpha_i - \lambda_i = 0 \iff \boxed{\alpha_i + \lambda_i = \frac{c}{n}}$$

SVM Dual Function

- Substituting these conditions back into L , the second term disappears.
- First and third terms become

$$\frac{1}{2} w^T w = \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$\sum_{i=1}^n \alpha_i (1 - y_i [w^T x_i + b]) = \sum_{i=1}^n \alpha_i - \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j x_j^T x_i - b \underbrace{\sum_{i=1}^n \alpha_i y_i}_{=0}.$$

- Putting it together, the dual function is

$$g(\alpha, \lambda) = \begin{cases} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j x_j^T x_i & \sum_{i=1}^n \alpha_i y_i = 0 \\ -\infty & \alpha_i + \lambda_i = \frac{c}{n}, \text{ all } i \\ & \text{otherwise.} \end{cases}$$

SVM Dual Problem

- The **dual function** is

$$g(\alpha, \lambda) = \begin{cases} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j x_j^T x_i & \begin{matrix} \sum_{i=1}^n \alpha_i y_i = 0 \\ \alpha_i + \lambda_i = \frac{c}{n}, \text{ all } i \end{matrix} \\ -\infty & \text{otherwise.} \end{cases}$$

- The **dual problem** is $\sup_{\alpha, \lambda \succeq 0} g(\alpha, \lambda)$:

$$\begin{aligned} \sup_{\alpha, \lambda} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j x_j^T x_i \\ \text{s.t.} \quad & \sum_{i=1}^n \alpha_i y_i = 0 \\ & \alpha_i + \lambda_i = \frac{c}{n} \quad \alpha_i, \lambda_i \geq 0, \quad i = 1, \dots, n \end{aligned}$$

SVM Dual Problem: Eliminating a Variable

- Can eliminate the λ variables:

$$\begin{aligned} \sup_{\alpha} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j x_j^T x_i \\ \text{s.t.} \quad & \sum_{i=1}^n \alpha_i y_i = 0 \\ & \alpha_i \in \left[0, \frac{C}{n}\right] \quad i = 1, \dots, n. \end{aligned}$$

- Quadratic objective in n unknowns and $2n$ constraints
- Constraints are **box constraints**. (Simpler than primal constraints.)

SVM Dual Problem: Connect to Primal

- Recall

$$\partial_w L = 0 \iff w = \sum_{i=1}^n \alpha_i y_i x_i$$

- If α^* is a solution to the dual problem, then

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

- Since $\alpha_i \in [0, \frac{c}{n}]$, we see that c controls the amount of weight we can put on any single example
- What's b ?

Complementary Slackness

- By strong duality, we have the following **complementary slackness** conditions
 - Lagrange multiplier is zero unless the [primal] constraint is active at the optimum: " $\lambda_i^* f_i(x^*) = 0$ "
- Our primal constraints:

$$\begin{aligned}(\alpha_i) \quad & (1 - y_i [x_i^T w + b]) - \xi_i \leq 0 \text{ for } i = 1, \dots, n \\(\lambda_i) \quad & -\xi_i \leq 0 \text{ for } i = 1, \dots, n\end{aligned}$$

- Complementary slackness is about **optimal** primal and dual variables
 - Let (w^*, b^*, ξ_i^*) be primal optimal
 - Let (α^*, λ^*) be dual optimal

The Bias Term: b

- For our SVM primal, the complementary slackness conditions are:

$$\alpha_i^* (1 - y_i [x_i^T w^* + b] - \xi_i^*) = 0 \quad (1)$$

$$\lambda_i^* \xi_i^* = \left(\frac{c}{n} - \alpha_i^* \right) \xi_i^* = 0 \quad (2)$$

- Suppose there's an i such that $\alpha_i^* \in (0, \frac{c}{n})$.
- (2) implies $\xi_i^* = 0$.
- (1) implies

$$\begin{aligned} 1 - y_i [x_i^T w^* + b^*] &= 0 \\ \iff x_i^T w^* + b^* &= y_i \text{ (use } y_i \in \{-1, 1\}) \\ \iff \boxed{b^* = y_i - x_i^T w^*} \end{aligned}$$

The Bias Term: b

- The optimal b is

$$b^* = y_i - x_i^T w^*$$

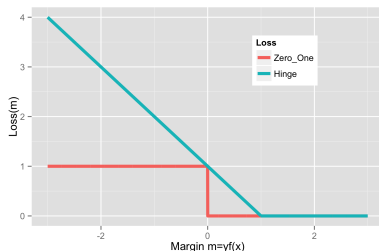
- We get the same b^* for any choice of i with $\alpha_i^* \in (0, \frac{c}{n})$
 - **With exact calculations!**
- With numerical error, more robust to average over all eligible i 's:

$$b^* = \text{mean} \left\{ y_i - x_i^T w^* \mid \alpha_i^* \in \left(0, \frac{c}{n} \right) \right\}.$$

- If there are no $\alpha_i^* \in (0, \frac{c}{n})$?
 - Then we have a **degenerate SVM training problem** ($w^* = 0$).

The Margin

- For notational convenience, define $f^*(x) = x_i^T w^* + b^*$.
- Margin $yf^*(x)$



- Incorrect classification: $yf^*(x) \leq 0$.
- Margin error: $yf^*(x) < 1$.
- “On the margin”: $yf^*(x) = 1$.
- “Good side of the margin”: $yf^*(x) > 1$.

Support Vectors and The Margin

- Recall $\xi_i^* = (1 - y_i f^*(x_i))_+$ the hinge loss on (x_i, y_i) .
- Suppose $\xi_i^* = 0$.
- Then $y_i f^*(x_i) \geq 1$
 - “on the margin” ($= 1$), or
 - “on the good side” (> 1)

Complementary Slackness Consequences

- For our SVM primal, the complementary slackness conditions are:

$$\alpha_i^* (1 - y_i f^*(x_i) - \xi_i^*) = 0$$

$$\lambda_i^* \xi_i^* = \left(\frac{c}{n} - \alpha_i^* \right) \xi_i^* = 0$$

- If $y_i f^*(x) > 1$ then the margin loss is $\xi_i^* = 0$, and we get $\alpha_i^* = 0$.
- If $y_i f^*(x_i) < 1$ then the margin loss is $\xi_i^* > 0$, so $\alpha_i^* = \frac{c}{n}$.
- If $\alpha_i^* = 0$, then $\xi_i^* = 0$, which implies no loss, so $y_i f^*(x) \geq 1$.

Complementary Slackness Results: Summary

$$\begin{aligned}\alpha_i^* = 0 &\implies y_i f^*(x_i) \geq 1 \\ \alpha_i^* \in \left(0, \frac{c}{n}\right) &\implies y_i f^*(x_i) = 1 \\ \alpha_i^* = \frac{c}{n} &\implies y_i f^*(x_i) \leq 1\end{aligned}$$

$$\begin{aligned}y_i f^*(x_i) < 1 &\implies \alpha_i^* = \frac{c}{n} \\ y_i f^*(x_i) = 1 &\implies \alpha_i^* \in \left[0, \frac{c}{n}\right] \\ y_i f^*(x_i) > 1 &\implies \alpha_i^* = 0\end{aligned}$$

Dual Problem: Dependence on x through inner products

- SVM Dual Problem:

$$\begin{aligned}
 &\sup_{\alpha} \quad \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j x_j^T x_i \\
 &\text{s.t.} \quad \sum_{i=1}^n \alpha_i y_i = 0 \\
 &\quad \alpha_i \in \left[0, \frac{c}{n}\right] \quad i = 1, \dots, n.
 \end{aligned}$$