Support Vector Machines

David S. Rosenberg

Bloomberg ML EDU

October 11, 2017

The SVM as a Quadratic Program

The Margin

Definition

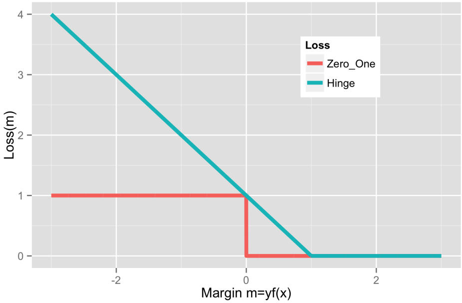
The **margin** (or **functional margin**) for predicted score $\hat{y}$ and true class $y \in \{-1, 1\}$ is $y\hat{y}$.

- The margin often looks like $yf(x)$, where $f(x)$ is our score function.
- The margin is a measure of how **correct** we are.
- We want to **maximize the margin**.
- Most classification losses depend only on the margin.

(This is distinct from but related to **geometric margin**.)

Hinge Loss

- SVM/Hinge loss: $\ell_{\text{Hinge}} = \max\{1-m, 0\} = (1-m)_+$
- Margin $m = yf(x)$; "Positive part" $(x)_+ = x\mathbb{1}(x \geqslant 0)$.



Hinge is a **convex**, **upper bound** on $0-1$ loss. Not differentiable at $m = 1$.
We have a **"margin error"** when $m < 1$.

Support Vector Machine

- Hypothesis space $\mathcal{F} = \left\{ f(x) = w^T x + b \mid w \in \mathbf{R}^d, b \in \mathbf{R} \right\}$.
- ℓ_2 regularization (Tikhonov style)
- Loss $\ell(m) = \max\{1-m, 0\} = (1-m)_+$
- The SVM prediction function is the solution to

$$\min_{w \in \mathbf{R}^d, b \in \mathbf{R}} \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i\left[w^T x_i + b\right]\right).$$

SVM Optimization Problem (Tikhonov Version)

The SVM prediction function is the solution to

$$\min_{w \in \mathbf{R}^d, b \in \mathbf{R}} \frac{1}{2}\|w\|^2 + \frac{c}{n} \sum_{i=1}^{n} \max\left(0, 1 - y_i \left[w^T x_i + b\right]\right).$$

- unconstrained optimization
- not differentiable because of the max (right at the border of a margin error)
- Can we reformulate into a differentiable problem?

SVM Optimization Problem

- The SVM optimization problem is equivalent to

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} & \xi_i \geqslant \max\left(0, 1 - y_i\left[w^T x_i + b\right]\right). \end{array}$$

- Which is equivalent to

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

SVM as a Quadratic Program

- The SVM optimization problem is equivalent to

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

- 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.

The SVM Dual Problem

SVM Lagrange Multipliers

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

Lagrange Multiplier	Constraint
λ_i	$-\xi_i \leqslant 0$
α_i	$\left(1 - y_i\left[w^T x_i + b\right]\right) - \xi_i \leqslant 0$

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

SVM Lagrangian

- The Lagrangian for this formulation is

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

- Primal and dual:

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

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

Strong Duality by Slater's constraint qualification

- The SVM optimization problem:

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

- Convex problem + affine constraints $\Longrightarrow$ strong duality iff problem is feasible
- Constraints are satisfied by $w = b = 0$ and $\xi_i = 1$ for $i = 1,\ldots,n$,
  - so **we have strong duality** $\Longrightarrow$

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

SVM Dual Function

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

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

- Taking inf of convex and differentiable function of w, b, ξ.
  - Quadratic in w and linear in ξ and b.
- 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:

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

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

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

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

The SVM Dual Problem

- Using 1st order conditions, and some massaging, the SVM dual problem is:

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

- Given solution α^* to dual, primal solution is $w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$.
- w^* is "in the span of the data" – i.e. a linear combination of $x_1, \ldots, x_n$.
- Note $\alpha_i^* \in [0, \frac{c}{n}]$. So c controls max weight on each example. (**Robustness**!)

SVM Dual Problem

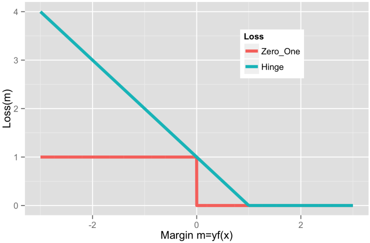
$$
\begin{aligned}
\sup_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_j^T x_i \\
\text{s.t.} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0 \\
& \alpha_i \in \left[0, \frac{c}{n}\right] \quad i = 1, \ldots, n.
\end{aligned}
$$

- Quadratic objective in n unknowns and $n+1$ constraints
- Efficient minimization algorithm: SMO (sequential minimal optimization)
- What other insights can we get from the dual formulation?

Insights From Complementary Slackness: Margin and Support Vectors

The Margin and Some Terminology

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



- Incorrect classification: $yf^*(x) \leqslant 0$.
- 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 "**slack variable**" $\xi_i^* = \max(0, 1 - y_i f^*(x_i))$ is the hinge loss on (x_i, y_i).
- Suppose $\xi_i^* = 0$.
- Then $y_i f^*(x_i) \geqslant 1$
  - "on the margin" $(= 1)$, or
  - "on the good side" (> 1)

Complementary Slackness Conditions

- Recall our primal constraints and Lagrange multipliers:

Lagrange Multiplier	Constraint
λ_i	$-\xi_i \leqslant 0$
α_i	$(1 - y_i f(x_i)) - \xi_i \leqslant 0$

- Recall first order condition $\nabla_{\xi_i} L = 0$ gave us $\lambda_i^* = \frac{c}{n} - \alpha_i^*$.
- By strong duality, we must have **complementary slackness**:

$$\alpha_i^* \left(1 - y_i f^*(x_i) - \xi_i^*\right) = 0$$
$$\lambda_i^* \xi_i^* = \left(\frac{c}{n} - \alpha_i^*\right) \xi_i^* = 0$$

Consequences of Complementary Slackness

- By strong duality, we must have **complementary slackness**:

$$
\begin{aligned}
\alpha_i^*\left(1-y_i f^*(x_i)-\xi_i^*\right) &= 0 \\
\left(\frac{c}{n}-\alpha_i^*\right)\xi_i^* &= 0
\end{aligned}
$$

- If $y_i f^*(x_i) > 1$ 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 $y_i f^*(x_i) \geqslant 1$.
- If $\alpha_i^* \in \left(0, \frac{c}{n}\right)$, then $\xi_i^* = 0$, which implies $1 - y_i f^*(x_i) = 0$.

Complementary Slackness Results: Summary

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

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

Support Vectors

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

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

  with $\alpha_i^* \in [0, \frac{c}{n}]$.
- The x_i's corresponding to $\alpha_i^* > 0$ are called **support vectors.**
- Few margin errors or "on the margin" examples $\Longrightarrow$ **sparsity in input examples**.

Complementary Slackness To Get b^*

The Bias Term: b

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

$$\alpha_i^* \left(1 - y_i \left[x_i^T w^* + b\right] - \xi_i^*\right) = 0 \tag{1}$$

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

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

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

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 $\alpha_i^* \in \left(0, \frac{c}{n}\right)$
  - **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 $\alpha_i^* \in \left(0, \frac{c}{n}\right)$?
  - Then we have a **degenerate SVM training problem**[1] ($w^* = 0$).

[1] See Rifkin et al.'s "A Note on Support Vector Machine Degeneracy", an MIT AI Lab Technical Report.

Teaser for Kernelization

Dual Problem: Dependence on x through inner products

- SVM Dual Problem:

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

- Note that all dependence on inputs x_i and x_j is through their inner product: $\langle x_j, x_i \rangle = x_j^T x_i$.
- We can replace $x_j^T x_i$ by any other inner product...
- This is a "kernelized" objective function.