

# SVM and Complementary Slackness

David Rosenberg

New York University

February 21, 2017

# SVM Review: Primal and Dual Formulations

# Support Vector Machine

- Hypothesis space  $\mathcal{F} = \{f(x) = w^T x + b \mid w \in \mathbf{R}^d, b \in \mathbf{R}\}$ .
- $\ell_2$  regularization (Tikhonov style)
- Loss  $\ell(m) = \max\{1 - m, 0\} = (1 - m)_+$
- The SVM prediction function is the solution to

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

# SVM as a Quadratic Program

- The SVM optimization problem is equivalent to

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

- Differentiable objective function
- $2n$  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 Lagrange Multipliers

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

Lagrange Multiplier	Constraint
$\lambda_i$	$-\xi_i \leq 0$
$\alpha_i$	$(1 - y_i [w^T x_i + b]) - \xi_i \leq 0$

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

# SVM Primal and Dual

- Lagrangian:

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

- Primal and dual formulations:

$$p^* = \underbrace{\inf_{w, \xi, b} \sup_{\alpha, \lambda \geq 0} L(w, b, \xi, \alpha, \lambda)}_{\text{primal objective}} \geq \sup_{\alpha, \lambda \geq 0} \underbrace{\inf_{w, b, \xi} L(w, b, \xi, \alpha, \lambda)}_{\text{dual objective } g(\alpha, \lambda)} = d^*$$

- Constraints are satisfied by  $w = b = 0$  and  $\xi_i = 1$  for  $i = 1, \dots, n$ .
- So we have strong duality by Slater's conditions. That is:  $p^* = d^*$ .

## First Order Conditions (KKT)

- Suppose  $(w^*, b^*)$  are primal optimal and  $(\xi^*, \alpha^*, \lambda^*)$  are dual optimal.
- By strong duality,  $p^* = d^*$ , and by “sandwich proof”,

$$p^* = d^* = L(w^*, b^*, \xi^*, \alpha^*, \lambda^*).$$

- Since  $L$  is differentiable, we must have first order conditions:

$$\begin{aligned} \nabla_w L(w^*, b^*, \xi^*, \alpha^*, \lambda^*) &= 0 \\ \text{(shown last week)} \iff w^* &= \sum_{i=1}^n \alpha_i^* y_i x_i \end{aligned}$$

- Conclude that  $w^*$  is “in the span of the data” – i.e. a linear combination of  $x_1, \dots, x_n$ .

# The SVM Dual Problem

- We found the SVM dual problem can be written as:

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

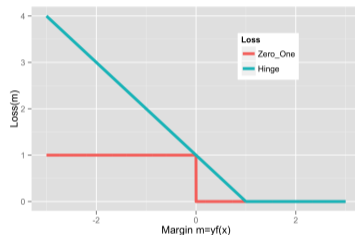
- Given solution  $\alpha^*$  to dual, primal solution is  $w^* = \sum_{i=1}^n \alpha_i^* y_i x_i$ .
- Note  $\alpha_i^* \in [0, \frac{c}{n}]$ . So  $c$  controls max weight on each example. (**Robustness!**)



# 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:  $yf^*(x) \leq 0$ .
- 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 “**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  $y_i f^*(x_i) \geq 1$ 
  - “on the margin” ( $= 1$ ), or
  - “on the good side” ( $> 1$ )

# Complementary Slackness Conditions

- Recall our primal constraints and Lagrange multipliers:

Lagrange Multiplier	Constraint
$\lambda_i$	$-\xi_i \leq 0$
$\alpha_i$	$(1 - y_i f(x_i)) - \xi_i \leq 0$

- 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**:

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

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

## Consequences of Complementary Slackness

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

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

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

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

## Complementary Slackness Results: Summary

$$\alpha_i^* = 0 \implies y_i f^*(x_i) \geq 1$$

$$\alpha_i^* \in \left(0, \frac{c}{n}\right) \implies y_i f^*(x_i) = 1$$

$$\alpha_i^* = \frac{c}{n} \implies y_i f^*(x_i) \leq 1$$

$$y_i f^*(x_i) < 1 \implies \alpha_i^* = \frac{c}{n}$$

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

$$y_i f^*(x_i) > 1 \implies \alpha_i^* = 0$$

# Support Vectors

- If  $\alpha^*$  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  $\implies$  **sparsity in input examples**.

## Complementary Slackness To Get $b^*$



## 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

$$y_i [x_i^T w^* + b^*] = 1$$

$$\iff x_i^T w^* + b^* = y_i \text{ (use } y_i \in \{-1, 1\})$$

$$\iff \boxed{b^* = y_i - x_i^T w^*}$$

## 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**<sup>1</sup> ( $w^* = 0$ ).

<sup>1</sup>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:

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

$$\text{s.t.} \quad \sum_{i=1}^n \alpha_i y_i = 0$$

$$\alpha_i \in \left[0, \frac{C}{n}\right] \quad i = 1, \dots, n.$$

- 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 product...
- This is a “kernelized” objective function.