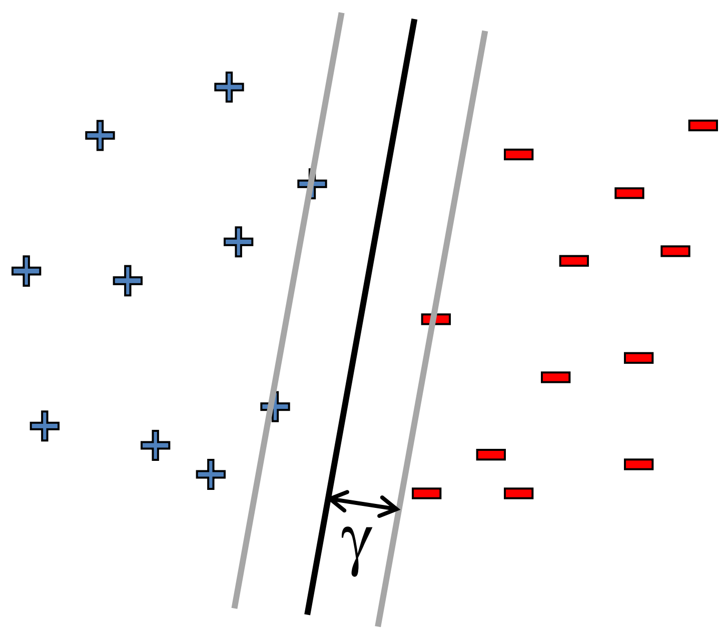
Support Vector Machines (SVMs) contd…

Aarti Singh

Machine Learning 10-315
Mar 23, 2022

Hard-margin SVM

Data perfectly separable by a linear decision boundary



Hard margin approach

$$\min_{\mathbf{w},b} \ \mathbf{w}.\mathbf{w}$$

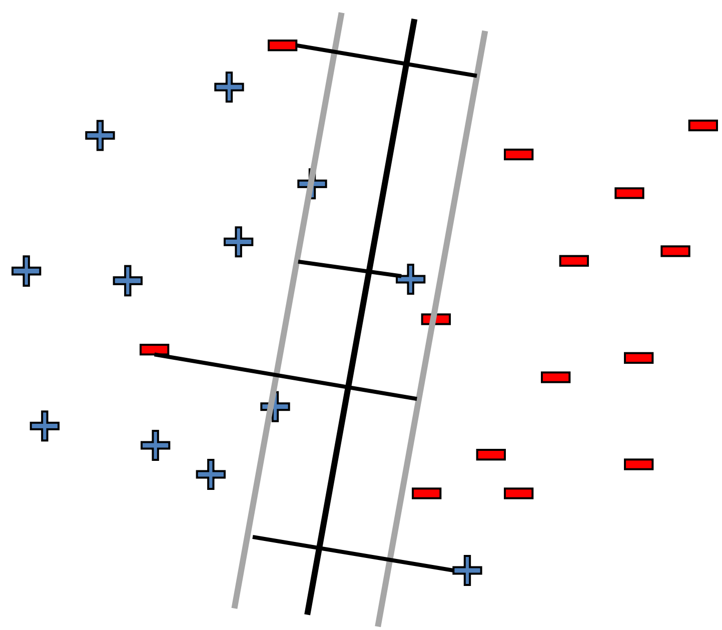
$$\text{s.t. } (\mathbf{w}.\mathbf{x}_j + b)\, y_j \geq 1 \quad \forall j$$

Solve using Quadratic Programming (QP)

$$\boxed{\text{Margin, } \gamma \quad \alpha \quad 1/\|w\|}$$

Soft-margin SVM

Allow "error" in classification



Soft margin approach

$$\min_{\mathbf{w},b,\{\xi_j\}} \mathbf{w}.\mathbf{w} + C \sum_j \xi_j$$

$$\text{s.t. } (\mathbf{w}.\mathbf{x}_j + b)\, y_j \geq 1 - \xi_j \quad \forall j$$

$$\xi_j \geq 0 \quad \forall j$$

ξ_j - "slack" variables

= (>1 if x_j misclassifed)

pay linear penalty if mistake

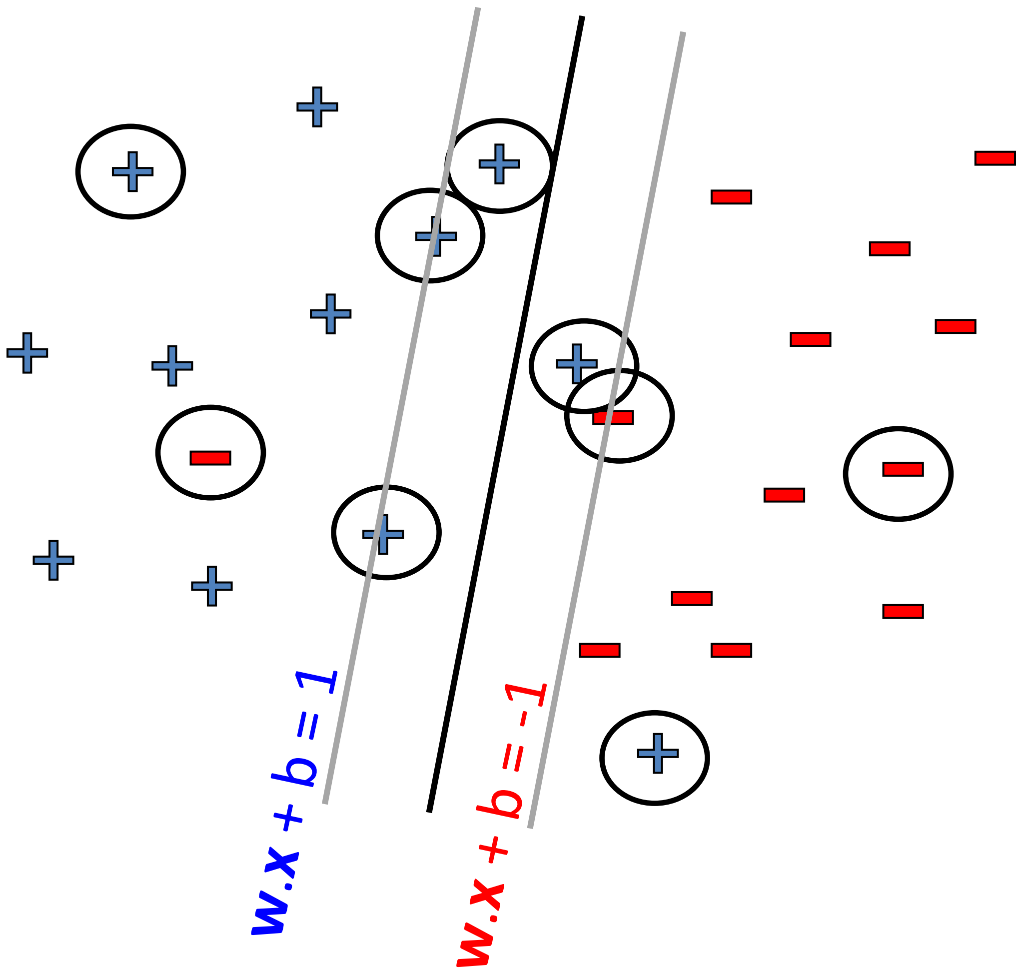
C - tradeoff parameter (chosen by cross-validation)

Still QP ☺

$$\min_{\mathbf{w},b,\{\xi_j\}} \quad \mathbf{w}.\mathbf{w} + C \sum \xi_j$$
$$\text{s.t. } (\mathbf{w}.\mathbf{x}_j + b)\, y_j \geq 1 - \xi_j \quad \forall j$$
$$\xi_j \geq 0 \quad \forall j$$

Slack variables

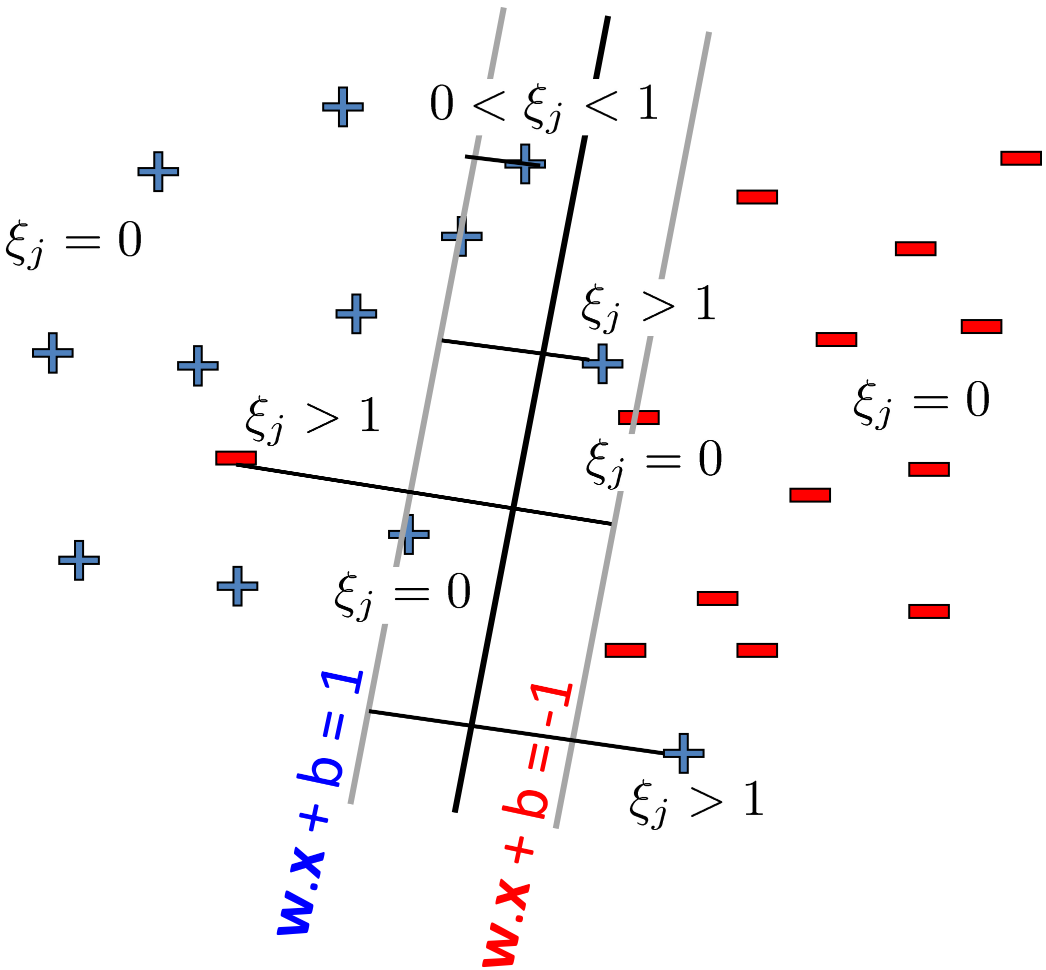
$$(\mathbf{w}.\mathbf{x}_j + b)\, y_j \geq 1 - \xi_j \quad \forall j$$



What is the slack ξ_j for the following points?

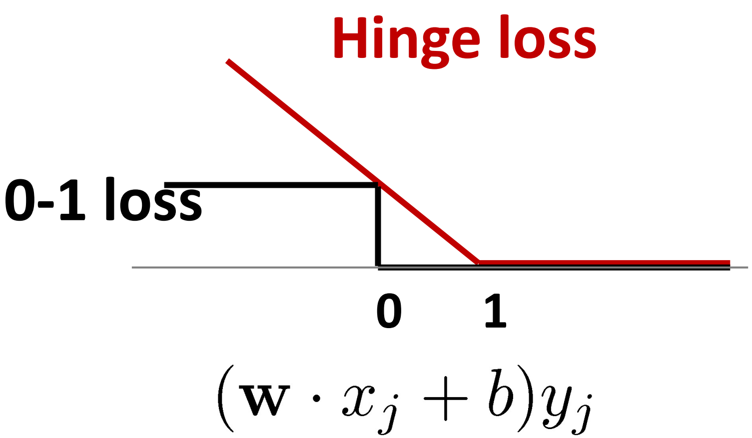
Confidence | Slack

Slack variables – Hinge loss



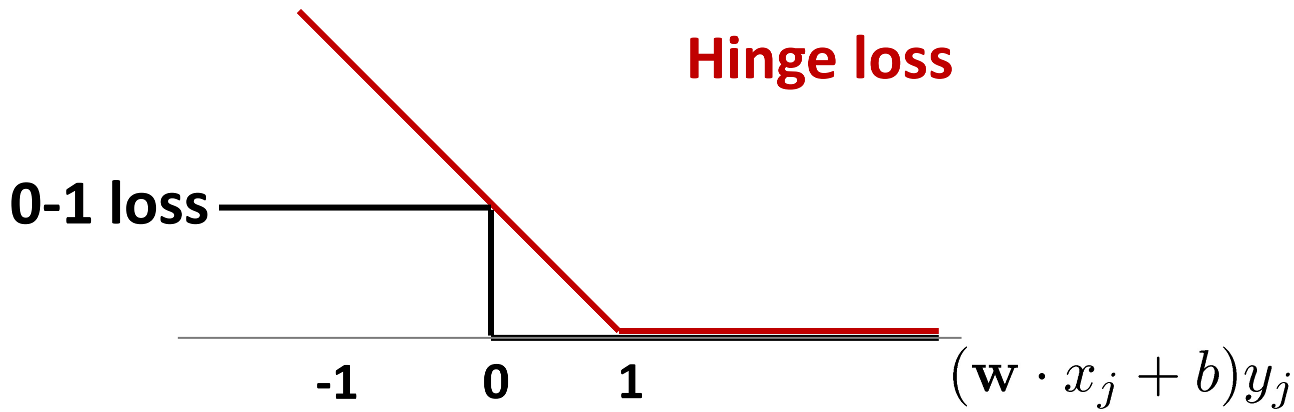
Notice that

$$\xi_j = (1 - (\mathbf{w} \cdot x_j + b)y_j))_+$$



Slack variables – Hinge loss

$$\xi_j = (1 - (\mathbf{w} \cdot x_j + b) y_j))_+$$



$$\min_{\mathbf{w}, b, \{\xi_j\}} \ \mathbf{w}.\mathbf{w} + C \sum_j \xi_j$$

$$\text{s.t. } (\mathbf{w}.\mathbf{x}_j + b)\, y_j \geq 1 - \xi_j \quad \forall j$$

$$\xi_j \geq 0 \quad \forall j$$

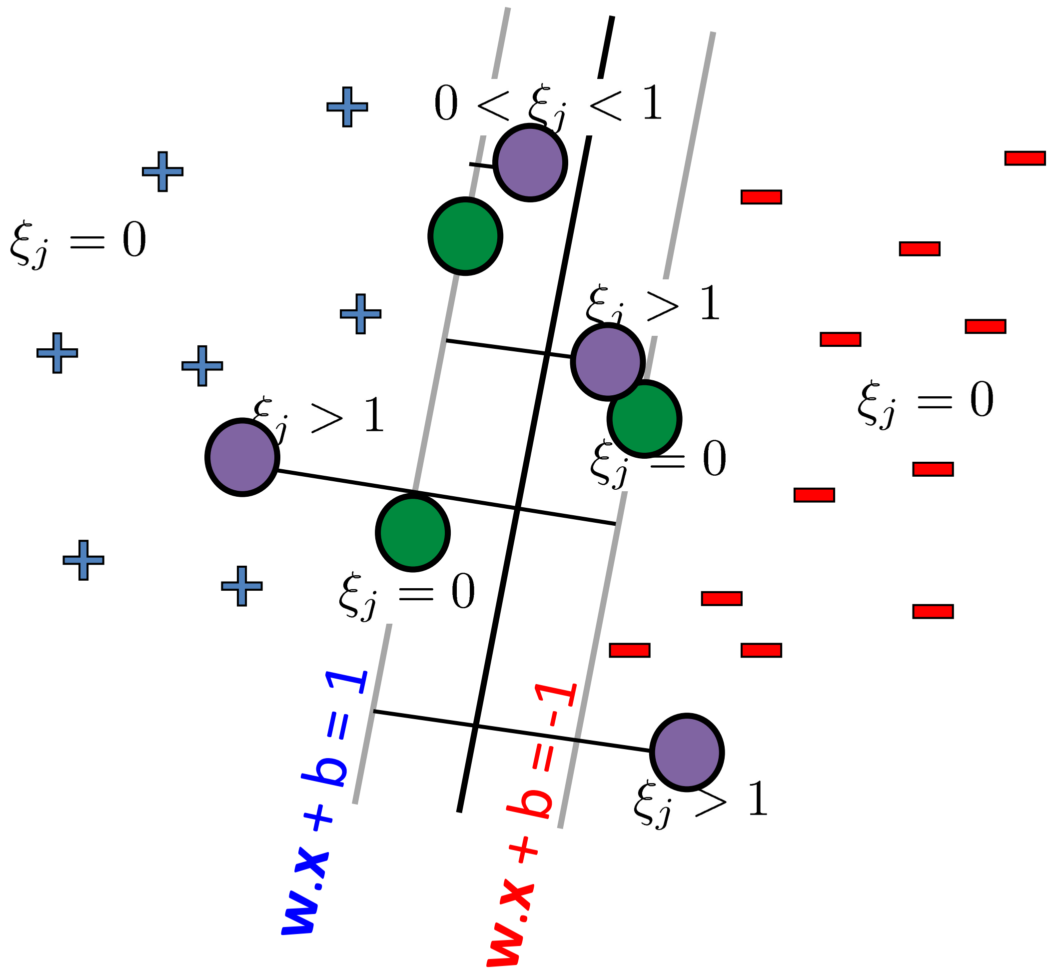
$\Leftrightarrow$

Regularized hinge loss

$$\min_{\mathbf{w}, b} \ \mathbf{w}.\mathbf{w} + C \sum_j (1 - (\mathbf{w}.x_j + b) y_j)_+$$

$$\min_{\mathbf{w},b,\{\xi_j\}} \quad \mathbf{w}.\mathbf{w} + C \sum \xi_j$$
$$\text{s.t. } (\mathbf{w}.\mathbf{x}_j+b)\, y_j \geq 1-\xi_j \quad \forall j$$
$$\xi_j \geq 0 \quad \forall j$$

Support Vectors



Margin support vectors

$\xi_j = 0$, $(\mathbf{w}.\mathbf{x}_j+b)\, y_j = 1$

(don't contribute to objective but enforce constraints on solution)

Correctly classified but on margin

Non-margin support vectors

$\xi_j > 0$

(contribute to both objective and constraints)

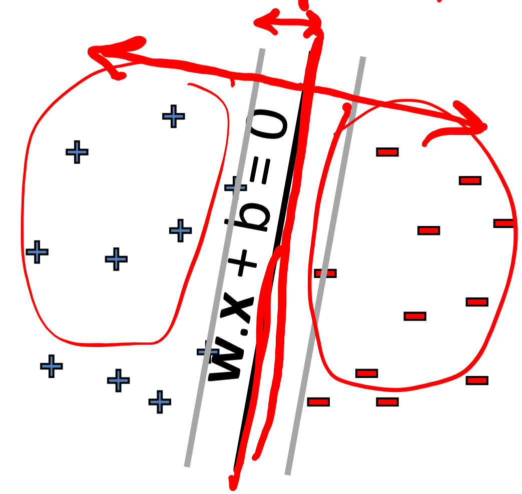
$1 > \xi_j > 0$ Correctly classified but inside margin

$\xi_j > 1$ Incorrectly classified

SVM – linearly separable case

$\gamma \propto \frac{1}{\|w\|}$

- n training points $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$
- d features $\mathbf{x}_j$ is a d-dimensional vector



- Primal problem: $$\text{minimize}_{\boxed{\mathbf{w},b}} \quad \frac{1}{2}\mathbf{w}.\mathbf{w}$$ $$\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j \geq 1, \ \forall j$$

w - weights on features (d-dim problem)

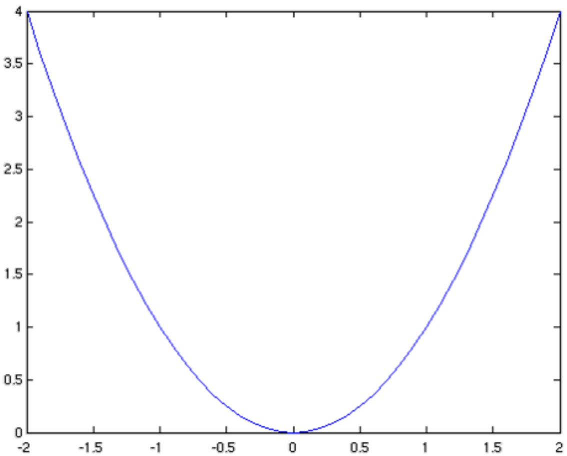
- Convex quadratic program – quadratic objective, linear constraints
- But expensive to solve if d is very large
- Often solved in dual form (n-dim problem)

Detour - Constrained Optimization

quadratic objective →
$$\min_x \; x^2$$
linear constraint →
$$\text{s.t.} \quad x \geq b$$

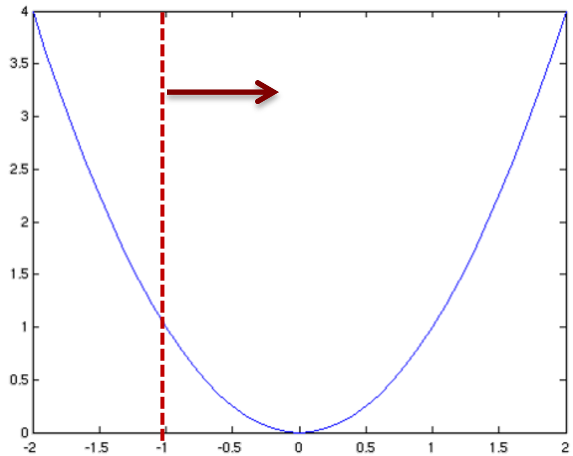
$$x^* = \max(b, 0)$$

$$\min_x \; x^2$$



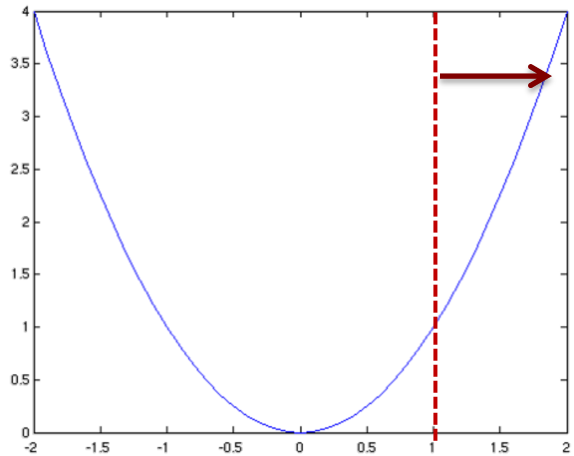
$$x^* = 0$$

$$\min_x \; x^2$$
$$\text{s.t.} \quad x \geq -1$$



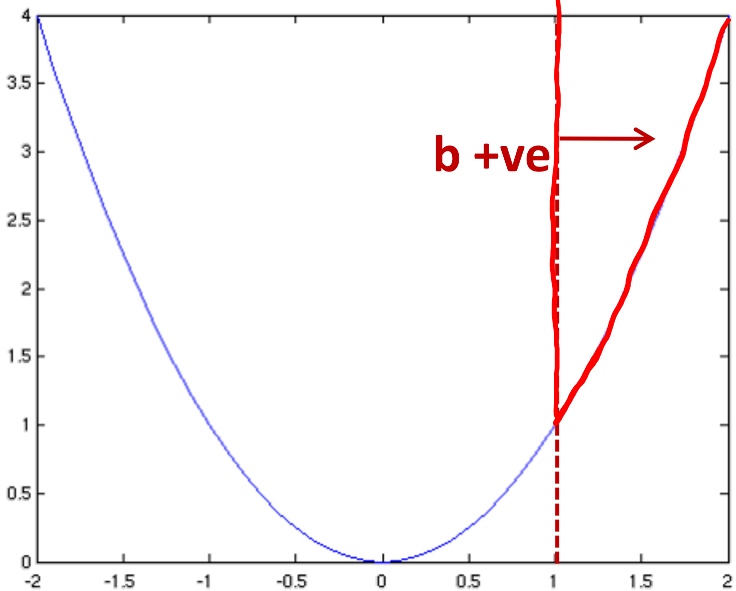
$$x^* = 0$$
Constraint inactive

$$\min_x \; x^2$$
$$\text{s.t.} \quad x \geq 1$$

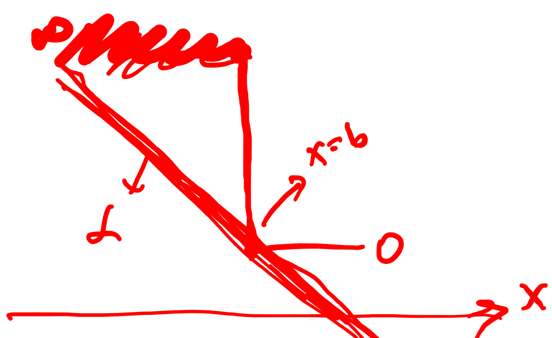


$$x^* = 1$$
Constraint active (tight)

Constrained Optimization



$x^* = b$



$\min_x \ x^2$

s.t. $\ x \geq b$

Equivalent unconstrained optimization:

$\min_x \ x^2 + I(x-b)$

$I(x-b) = \begin{cases} \infty & x < b \\ 0 & x \geq b \end{cases}$

Replace with lower bound ($\alpha >= 0$)

$\underbrace{x^2 + I(x-b)}_{\text{LHS}} \ >= \ \underbrace{x^2 - \alpha(x-b)}_{L(x,\alpha)}$

$\max_{\alpha \geq 0} L(x,\alpha)$

Primal and Dual Problems

Primal problem: p* = $\min_x \; x^2$
s.t. $x \geq b$

$\leftrightarrow \alpha \geq 0$

Dual problem: d* = $\max_\alpha \boxed{d(\alpha)}$
s.t. $\alpha \geq 0$

= $\min_{x} \max_{\alpha \geq 0} L(x, \alpha)$ $\longleftrightarrow$ = $\max_\alpha \boxed{\min_x L(x, \alpha)}$
s.t. $\alpha \geq 0$

where Lagrangian $L(x, \alpha) = x^2 - \alpha(x - b)$

$L(x, \alpha) = x^2 - \alpha(x-b)$

How to form the Lagrangian?

$\alpha \geq 0$

For each constraint, introduce a positive Lagrange multiplier

Fold constraints into objective

$x^2 - \alpha(x-b)$

$\min_{x_1, x_2} x_1^2 + x_2^2$ s.t. $x_1 \geq b_1$, $x_2 \geq b_2$ $\quad \alpha_1 \geq 0$, $\alpha_2 \geq 0$

$x_1^2 + x_2^2 - \alpha_1(x_1 - b_1) - \alpha_2(x_2 - b_2)$

Why solve the Dual problem?

w, b (d+1) dim

Primal problem: p* = $\min_x \ x^2$
s.t. $x \geq b$

Dual problem: d* = $\max_\alpha \ \boxed{d(\alpha)}$
s.t. $\alpha \geq 0$

= $\underbrace{\min_x \max_{\alpha \geq 0}} L(x, \alpha)$

= $\max_\alpha \min_x L(x, \alpha)$
s.t. $\alpha \geq 0$

- **Dual problem (maximization) is always concave even if primal is not convex**

  Why? Pointwise infimum of concave functions is concave.
  [Pointwise supremum of convex functions is convex.]

$$L(x, \alpha) = x^2 - \alpha(x - b)$$

- **As many dual variables α as constraints, helpful if fewer constraints than dimension of primal variable x**

Connection between Primal and Dual

Primal problem: p* = $\min_x \; x^2$ s.t. $x \geq b$

Dual problem: d* = $\max_\alpha \; d(\alpha)$ s.t. $\alpha \geq 0$

- **Weak duality:** The dual solution d* lower bounds the primal solution p* i.e. d* ≤ p*

To see this, recall $L(x, \alpha) = x^2 - \alpha(x - b)$

For every feasible x' (i.e. x' ≥ b) and feasible α' (i.e. α' ≥ 0) , notice that

$$d(\alpha) = \min_x L(x, \alpha) \leq \underbrace{x'^2 - \alpha'(x'-b)}_{L(x',\alpha')} \leq x'^2$$

Since above holds true for every feasible x', we have d(α) ≤ x^{*2} = p*

Connection between Primal and Dual

x^* α^*

Primal problem: p* = $\min_x \; x^2$
s.t. $x \geq b$

Dual problem: d* = $\max_\alpha \; d(\alpha)$
s.t. $\alpha \geq 0$

- **Weak duality:** The dual solution d* lower bounds the primal solution p* i.e. d* ≤ p*

- **Strong duality:** d* = p* holds often for many problems of interest e.g. if the primal is a feasible convex objective with linear constraints

SVM

primal variables w^*, b^* ⟷ dual variables $\alpha_1^*, \ldots, \alpha_n^*$

Connection between Primal and Dual

What does strong duality say about α^* (the α that achieved optimal value of dual) and x^* (the x that achieves optimal value of primal problem)?

$\min_x \max_\alpha L(x,\alpha)$ $\max_\alpha \min_x L(x,\alpha)$

Whenever strong duality holds, the following conditions (known as KKT conditions) are true for α^* and x^*:

- 1. $\nabla L(x^*, \alpha^*) = 0$ i.e. Gradient of Lagrangian at x^* and α^* is zero.
- 2. $x^* \geq b$ i.e. x^* is primal feasible
- 3. $\alpha^* \geq 0$ i.e. α^* is dual feasible
- 4. $\alpha^*(x^* - b) = 0$ (called as complementary slackness)

dual const primal constraint

$\alpha \geq 0$

$x \geq b$

$\max \|w\|^2$

s.t. $(wx_i + b)y_i \geq 1$

We use the first one to relate x^* and α^*. We use the last one (complimentary slackness) to argue that $\alpha^* = 0$ if constraint is inactive and $\alpha^* > 0$ if constraint is active and tight.

Primal and Dual Problems

Primal problem: p* = $\min_x \; x^2$
s.t. $x \geq b$

= $\min_x \max_{\alpha \geq 0} L(x, \alpha)$

Dual problem: d* = $\max_\alpha \; d(\alpha)$
s.t. $\alpha \geq 0$

= $\max_\alpha \underbrace{\min_x L(x, \alpha)}^{d(\alpha)}$
s.t. $\alpha \geq 0$

where Lagrangian $L(x, \alpha) = x^2 - \alpha(x - b)$

How to form the Lagrangian?

For each constraint, introduce a positive Lagrange multiplier

Fold constraints into objective

Dual SVM – linearly separable case

n training points, d features $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ where x_i is a d-dimensional vector

- <u>Primal problem</u>:

$$\text{minimize}_{\mathbf{w},b} \quad \frac{1}{2}\mathbf{w}.\mathbf{w}$$
$$\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j \geq 1, \ \forall j = 1, \ldots, n$$

$\alpha_1 \ldots \alpha_n \geq 0$

w - weights on features (d-dim problem)

- <u>Dual problem</u> (derivation):

$\max_{\alpha_1 \ldots \alpha_n \geq 0} d(\alpha) = \min_{w,b} L(w, b, \alpha)$

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\mathbf{w}.\mathbf{w} - \sum_j \alpha_j \left[\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j - 1\right]$$
$$\alpha_j \geq 0, \ \forall j$$

α - weights on training pts (n-dim problem)

Dual SVM – linearly separable case

- Dual problem:

$$\max_\alpha \underbrace{\min_{\mathbf{w},b} L(\mathbf{w},b,\alpha)}_{d(\alpha)} = \frac{1}{2}\mathbf{w}.\mathbf{w} - \sum_j \alpha_j \left[\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j - 1\right]$$

$$\alpha_j \geq 0, \ \forall j$$

$\frac{\partial L}{\partial w} = w - \sum_j \alpha_j x_j y_j = 0$

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \qquad \Rightarrow \mathbf{w} = \sum_j \alpha_j y_j \mathbf{x}_j$$

If we can solve for αs (dual problem), then we have a solution for **w** (primal problem)

$$\frac{\partial L}{\partial b} = 0 \qquad \Rightarrow \sum_j \alpha_j y_j = 0$$

$\frac{\partial L}{\partial b} = \sum_j \alpha_j y_j = 0$

Dual SVM – linearly separable case

- Dual problem:

$$\underbrace{\max_\alpha \min_{\mathbf{w},b} L(\mathbf{w},b,\alpha)}_{d(\alpha)} = \frac{1}{2}\mathbf{w}.\mathbf{w} - \sum_j \alpha_j \left[\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j - 1\right]$$

$\alpha_j \geq 0, \ \forall j$

$$\Rightarrow \mathbf{w} = \sum_j \alpha_j y_j \mathbf{x}_j \qquad \Rightarrow \sum_j \alpha_j y_j = 0$$

$$L(w^*, b^*, \alpha) = \frac{1}{2} \sum_j \alpha_j y_j x_j \cdot \sum_i \alpha_i y_i x_i - \sum_j \alpha_j \left[\left(\underbrace{\sum_i \alpha_i y_i x_i}\cdot x_j + b\right) y_j - 1\right]$$

$$= \underbrace{\frac{1}{2} \sum_j \alpha_j y_j x_j \cdot \sum_i \alpha_i y_i x_i}_{} - \underbrace{\sum_i \alpha_i x_i y_i \cdot \sum_j \alpha_j y_j x_j}_{} - b \overbrace{\sum_j \alpha_j y_j}^{0} + \sum_j \alpha_j$$

$$= -\frac{1}{2} \sum_i \alpha_i y_i x_i \cdot \sum_j \alpha_j y_j x_j + \sum_j \alpha_j \quad = \underline{d(\alpha)}$$

Dual SVM – linearly separable case

maximize$_{\alpha}$ $\quad \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i . \mathbf{x}_j$

$\alpha_1 \ldots \alpha_n$

$$\sum_i \alpha_i y_i = 0$$
$$\alpha_i \geq 0$$

n-dim

Dual problem is also QP

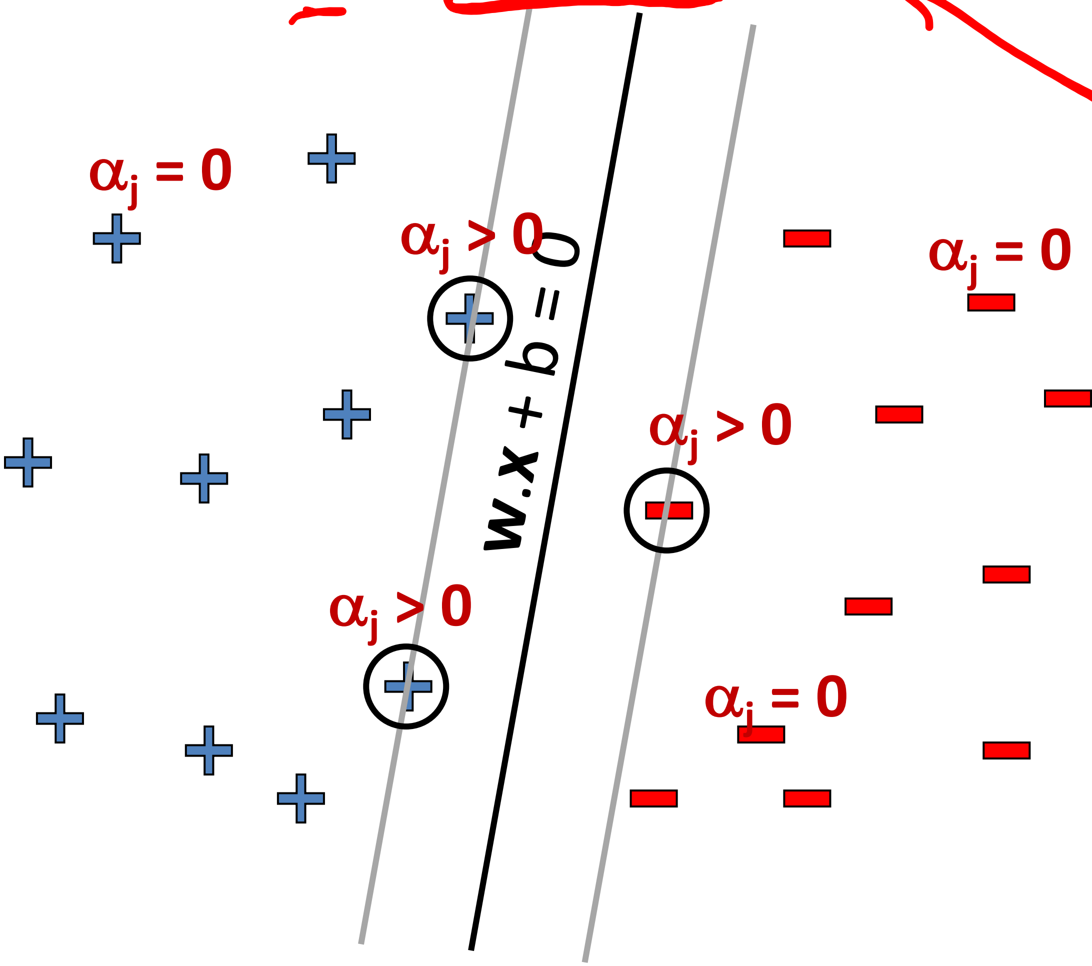
Solution gives α_js

$$\boxed{\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i}$$

What about b?

Dual SVM: Sparsity of dual solution

$\alpha_j \left[(w \cdot x_j + b) y_j - 1 \right] = 0$



$$\mathbf{w} = \sum_j \alpha_j y_j \mathbf{x}_j$$

Complementary slackness implies
Only few α_js can be non-zero : where constraint is active and tight

$(\mathbf{w}.\mathbf{x}_j + b)y_j = 1$

$(w \cdot x_j + b) y_j > 1$

Support vectors – training points j whose α_js are non-zero