Support Vector Machines

Somak Aditya Sudeshna Sarkar
CSE Department, IIT Kharagpur
Aug 30, 2024

These slides were assembled by Byron Boots, with only minor modifications from Eric Eaton's slides and grateful acknowledgement to the many others who made their course materials freely available online. Feel free to reuse or adapt these slides for your own academic purposes, provided that you include proper attribution.

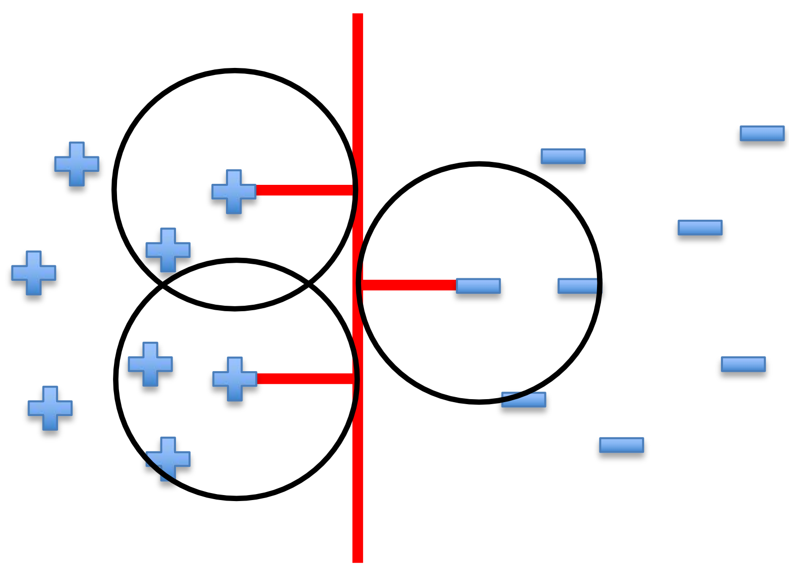
Adapted from slides by Tim Oates

Last Time: SVMs, Maximizing Margin

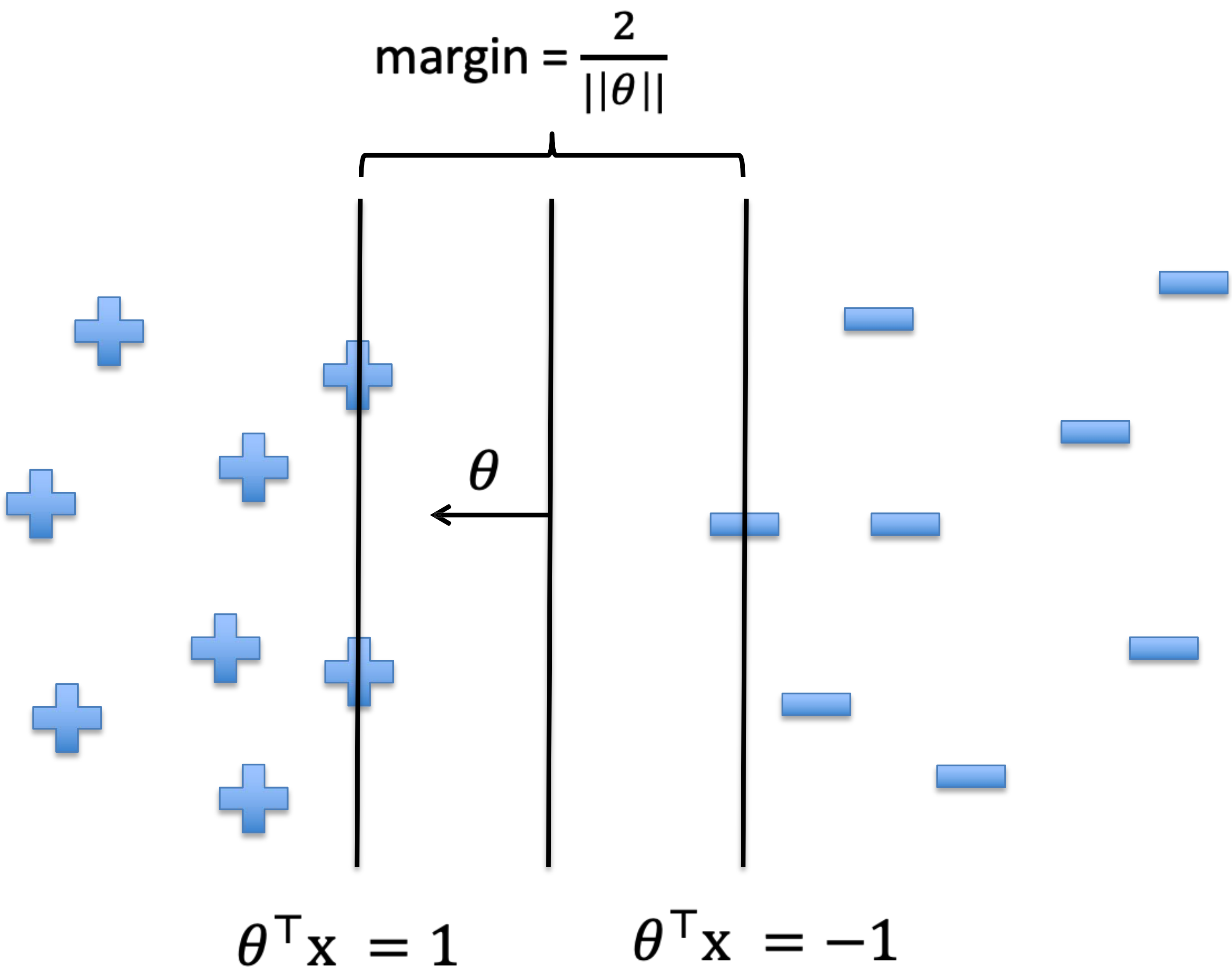
The SVM problem (assuming data is linearly separable):

$$\min_{\theta} \frac{1}{2} \sum_{j=1}^{d} \theta_j^2$$

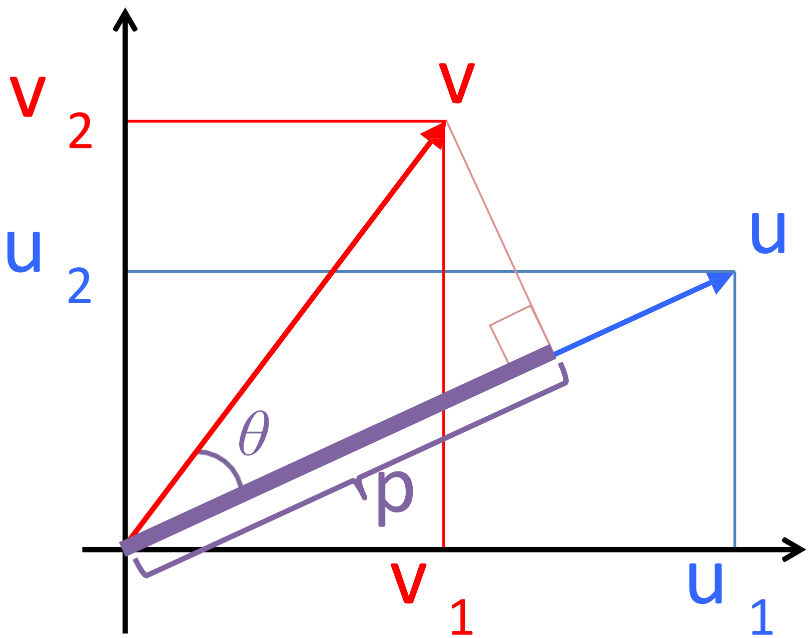
$$\text{s.t. } y_i(\theta^\top x_i) \geq 1 \forall i$$



Maximum Margin Hyperplane



Vector Inner Product



$$u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \qquad v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

$$\|\mathbf{u}\|_2 = \text{length}(\mathbf{u}) \in \mathbb{R}$$
$$= \sqrt{u_1^2 + u_2^2}$$

$$\mathbf{u}^\mathsf{T}\mathbf{v} = \mathbf{v}^\mathsf{T}\mathbf{u}$$
$$= u_1 v_1 + u_2 v_2$$
$$= \|\mathbf{u}\|_2 \; \|\mathbf{v}\|_2 \cos\theta$$
$$= p\|\mathbf{u}\|_2 \quad \text{where } p = \|\mathbf{v}\|_2 \cos\theta$$

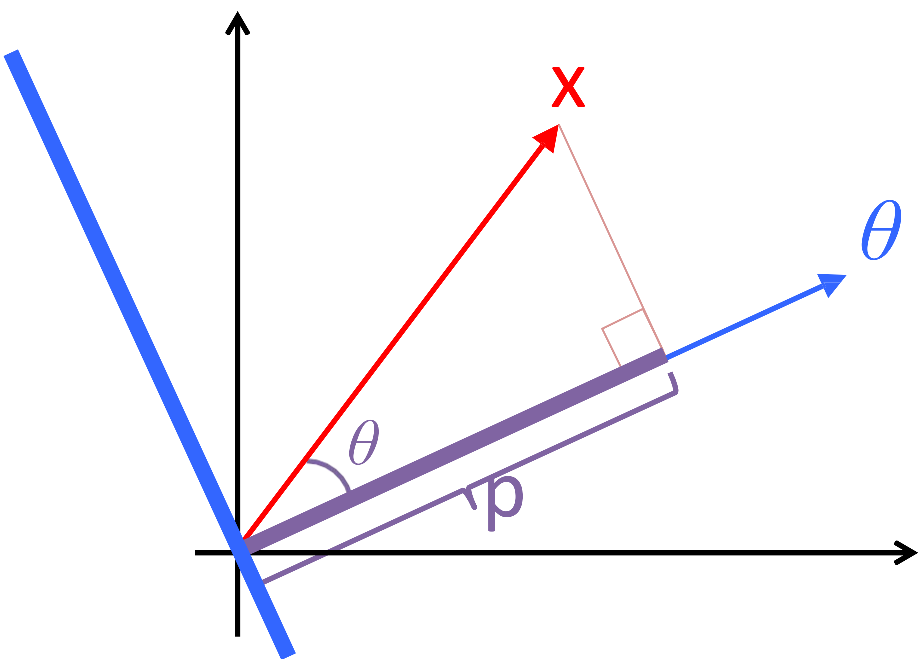
Understanding the Hyperplane

$$\min_{\theta} \frac{1}{2}\sum_{j=1}^{d} \theta_j^2$$

$$\text{s.t. } \theta^T x_i \geq 1, \quad \text{if } y_i = 1$$

$$\text{s.t. } \theta^T x_i \leq -1, \quad \text{if } y_i = -1$$

Assume $\theta_0 = 0$ so that the hyperplane is centered at the origin, and that d = 2



$$\boldsymbol{\theta}^{\top}\mathbf{x} = \left|\left|\boldsymbol{\theta}\right|\right|_2 \left|\left|\mathbf{x}\right|\right|_2 \cos(\boldsymbol{\theta})$$
$$= \boldsymbol{p} \left|\left|\boldsymbol{\theta}\right|\right|_2$$

Based on example by Andrew Ng

Maximizing the Margin

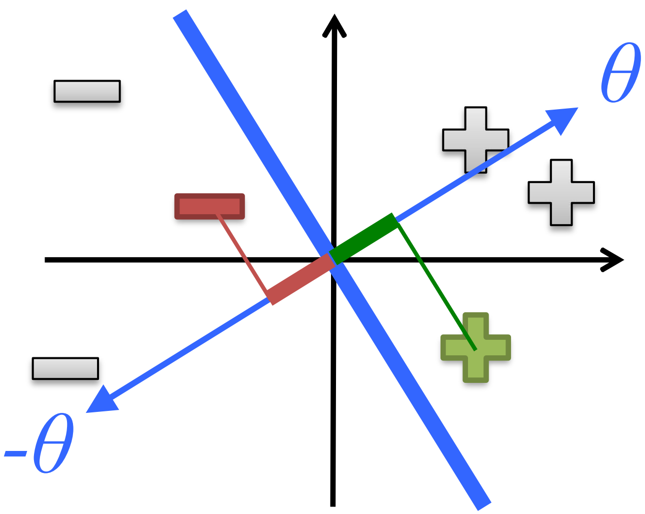
$$\min_{\theta} \frac{1}{2} \sum_{j=1}^{d} \theta_j^2$$

$$\text{s.t. } \theta^T x_i \geq 1, \quad \text{if } y_i = 1$$

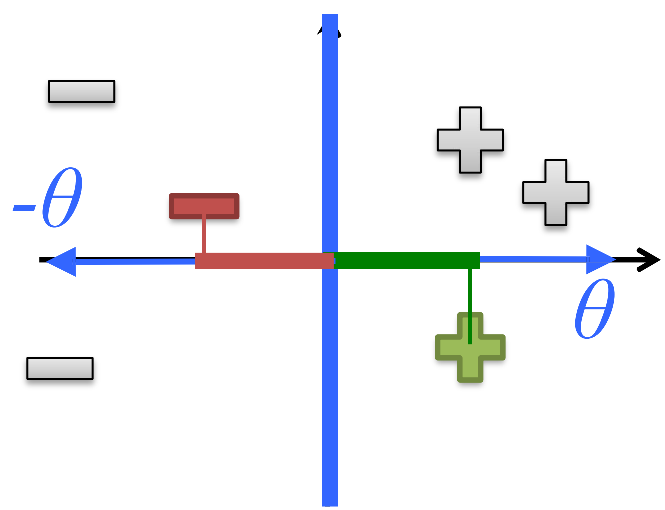
$$\text{s.t. } \theta^T x_i \leq -1, \quad \text{if } y_i = -1$$

Assume $\theta_0 = 0$ so that the hyperplane is centered at the origin, and that d = 2

Let p_i be the projection of x_i onto the vector θ

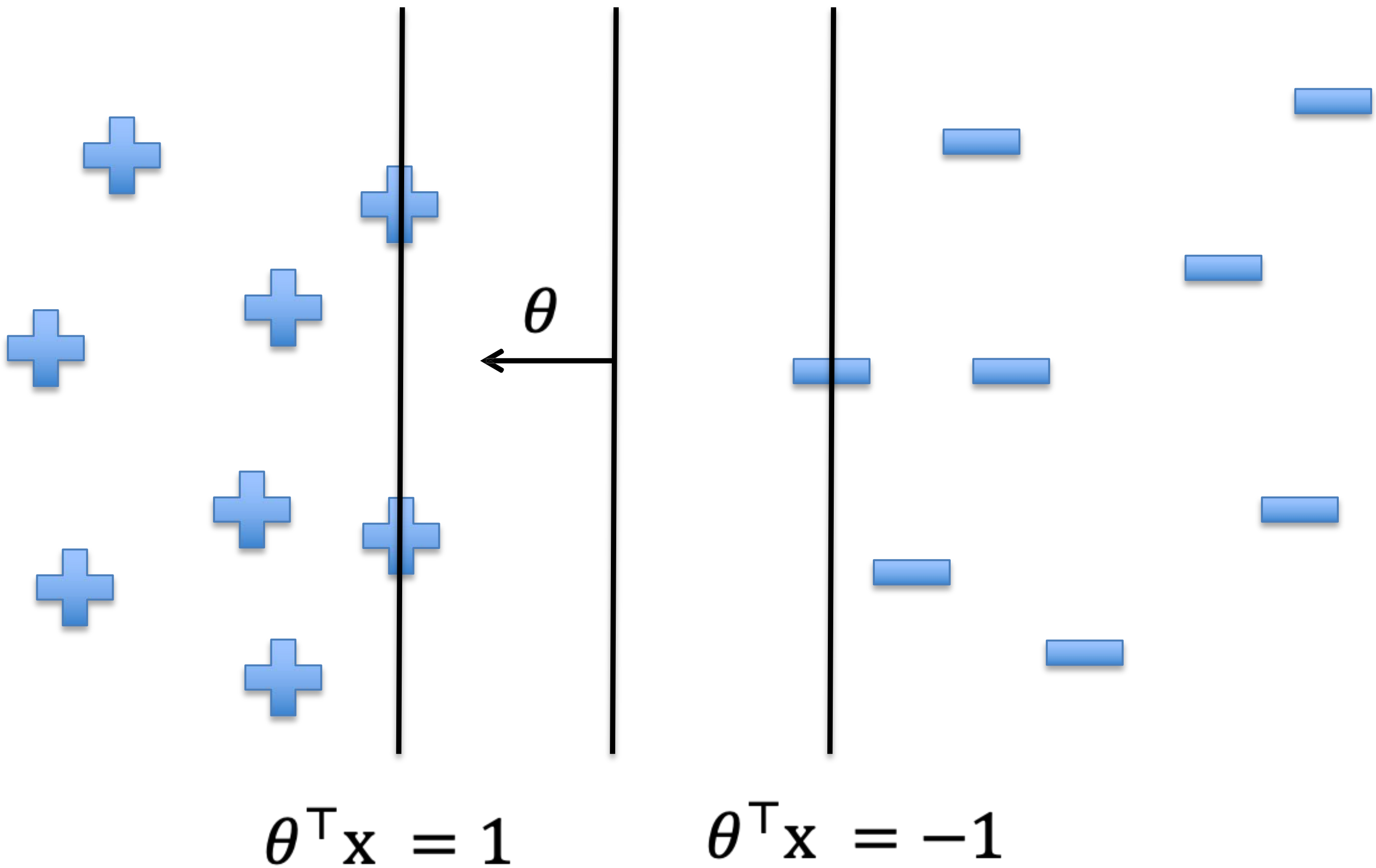


Since p is small, therefore $||\theta||_2$ must be large to have $p||\theta||_2 \geq 1$ (or ≤ -1)



Since p is larger, $||\theta||_2$ can be smaller and still satisfy $p||\theta||_2 \geq 1$ (or ≤ -1)

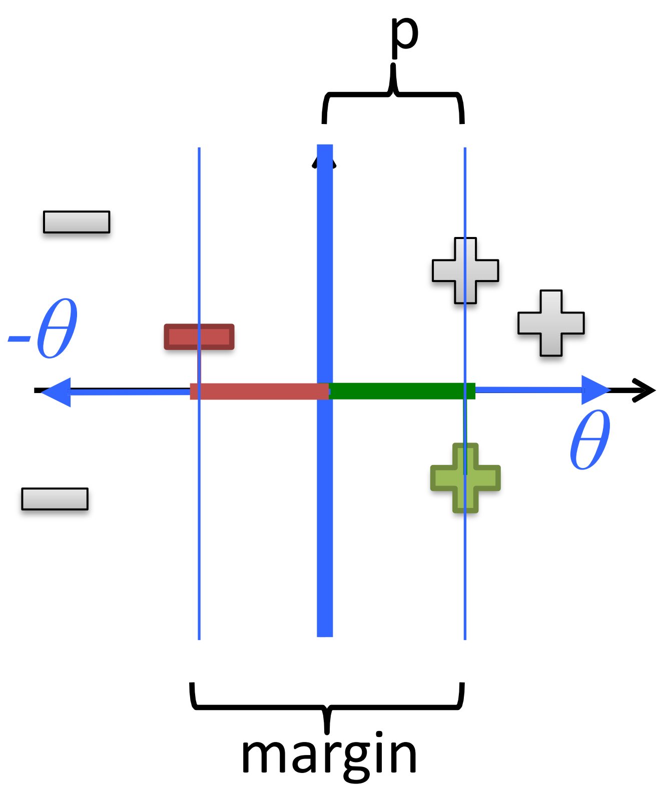
Support Vectors



Size of the Margin

For the support vectors, we have. $p||\boldsymbol{\theta}||_2 = \pm 1$

p is the length of the projection of the SVs onto $\boldsymbol{\theta}$



Therefore,

$$p = \frac{1}{||\boldsymbol{\theta}||_2}$$

$$\text{Margin} = 2p = \frac{2}{||\boldsymbol{\theta}||_2}$$

The SVM Dual Problem

The primal SVM problem was given as

$$J(\boldsymbol{\theta}) = \min_{\theta} \frac{1}{2} \sum_{j=1}^{d} \theta_j^2$$

$$\text{s.t. } y_i(\boldsymbol{\theta}^T x_i) \geq 1, \quad \text{if } \forall i$$

Can solve it more efficiently by taking the Lagrangian dual

- Duality is a common idea in optimization
- Transforms into a simpler optimization
- Key idea: introduce slack variables α_i for each constraint
  - α_i indicates how important a particular constraint is to the solution

The SVM Dual Problem

- The Lagrangian is given by

$$L(\boldsymbol{\theta}, \boldsymbol{\alpha}) = \frac{1}{2}\sum_{j=1}^{d} \theta_j^2 - \sum_{i=1}^{n} \alpha_i (y_i(\theta^T x_i) - 1)$$

$$\text{s.t. } \alpha_i \geq 0. \quad \forall i$$

- By definition this new formulation

$$\min_{\theta} \max_{\alpha} L(\boldsymbol{\theta}, \boldsymbol{\alpha}) \equiv \min_{\theta} J(\boldsymbol{\theta})$$

- We must minimize over $\boldsymbol{\theta}$ and maximize over $\boldsymbol{\alpha}$
- At optimal solution, partials w.r.t $\boldsymbol{\theta}$'s are 0

Solve by a bunch of algebra and calculus ... and we obtain ...

Solving the Optimization Problem (Primal to Dual)

$$\min_{\theta} \frac{1}{2}\sum_{j=1}^{d} \theta_j^2$$
$$\text{s.t.}\ \forall i\ y_i(\theta^T x_i) \geq 1$$

Quadratic programming with linear constraints

Minimize

$$L(\boldsymbol{\theta}, \boldsymbol{\alpha}) = \frac{1}{2}\sum_{j=1}^{d} \theta_j^2 - \sum_{i=1}^{n} \alpha_i (y_i(\theta^T x_i) - 1)$$
$$\text{s.t.}\ \forall i\ \ \alpha_i \geq 0$$

Lagrangian Function

Solving the Optimization Problem

Minimize

$$L(\boldsymbol{\theta}, \boldsymbol{\alpha}) = \frac{1}{2}\sum_{j=1}^{d} \theta_j^2 - \sum_{i=1}^{n} \alpha_i (y_i(\theta^T x_i) - 1)$$

$$\text{s.t.} \, \forall i \;\; \alpha_i \geq 0$$

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = 0 \Rightarrow \boldsymbol{\theta} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0$$

$$\Rightarrow \boldsymbol{\theta} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

The representer theorem: $\boldsymbol{\theta}$ as linear combination of training data

$$\frac{\partial L}{\partial \theta_0} = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$$

Where does θ_0 come from?

Solving the Optimization Problem

$$L = \frac{1}{2}\sum_{j=1}^{d}\theta_j^2 - \sum_{i=1}^{n}\alpha_i(y_i(\theta^T x_i) - 1)$$

If we substitute $\boldsymbol{\theta} = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i$ to L, we have

Details may be skipped

$$L = \frac{1}{2}\sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i^T \sum_{j=1}^{n}\alpha_j y_j \mathbf{x}_j - \sum_{i=1}^{n}\alpha_i(y_i(\theta^T x_i) - 1)$$

$$L = \frac{1}{2}\sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i^T \sum_{j=1}^{n}\alpha_j y_j \mathbf{x}_j - \sum_{i=1}^{n}\alpha_i\left(y_i\left(\sum_{j=1}^{n}\alpha_j y_j \mathbf{x}_j^T \mathbf{x}_i\right) - 1\right)$$

$$L = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i y_i \sum_{j=1}^{n}\alpha_j y_j \mathbf{x}_j^T \mathbf{x}_i$$

$$L = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j + \sum_{i=1}^{n}\alpha_i$$

This is a function of α_i

The Dual Problem

The objective function is in terms of α_i only.

- It is known as the dual problem: if we know $\boldsymbol{\theta}$, we know all α_i; if we know all α_i, we know $\boldsymbol{\theta}$
- The original problem = primal problem
- The objective function of the dual problem needs to be maximized (comes out from the KKT theory)
- *Learn* d *parameters for primal.* N *parameters for dual. Efficient if* $N \ll d$

The dual problem is:

$$\max \quad W(\alpha) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j + \sum_{i=1}^{n}\alpha_i$$

Subject to $\alpha_i \geq 0$, $\sum_{i=1}^{n}\alpha_i y_i = 0$

Properties of α_i when we introduce the Lagrange multipliers

The result when we differentiate the original Lagrangian w.r.t. θ_0

This is a quadratic programming (QP) problem. A global maximum of α_i can always be found.
$\boldsymbol{\theta}$ can be recovered by $\boldsymbol{\theta} = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i$

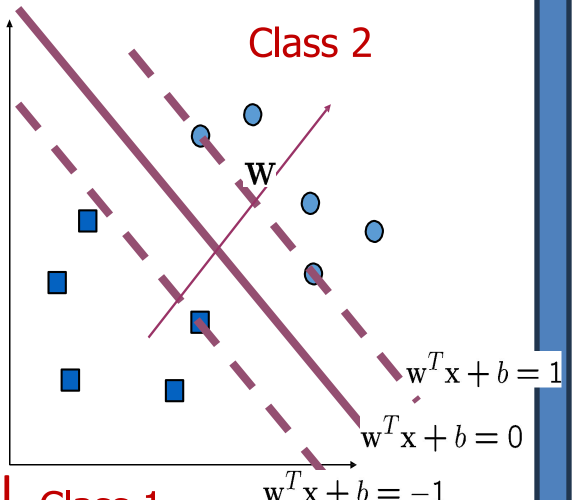
Rememeber the classifier becomes = $\boldsymbol{\theta}^\top\mathbf{x} = \left(\sum_{i=1}^{n}\alpha_i y_i \mathbf{x_i}\right)^\top\mathbf{x}$

One Slide Summary

Minimize $\frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta}$

subject to $y_n(\boldsymbol{\theta}^\top x_n) \geq 1$, for $n = 1,2, \ldots N$

Maximize Margin
Learn θ (hyperplane)



Minimize **Primal**

$$L(\boldsymbol{\theta}, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta} - \sum_{i=1}^{n} \alpha_i (y_i(\boldsymbol{\theta}^T x_i) - 1)$$

$$\text{s.t.} \forall i \ \ \alpha_i \geq 0$$

Dual

Maximize $J(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$

$\text{s.t.}\ \alpha_i \geq 0\ \forall i$ (comes from Lagrangian Assumptions)

$\sum_i \alpha_i y_i = 0$ (comes from differentiating w.r.t θ_0)

Maximize Margin
Learn α (weight
of support vectors)

SVM Dual

$$\text{Maximize } J(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

$$\text{s. t. } \alpha_i \geq 0 \; \forall i$$

$$\sum_i \alpha_i y_i = 0$$

The decision function is given by

$$h(\mathbf{x}) = \operatorname{sign}\left(\sum_{i \in \mathcal{SV}} \alpha_i y_i \langle \mathbf{x}, \mathbf{x}_i \rangle + b \right)$$

$$\text{where} \quad b = \frac{1}{|\mathcal{SV}|} \sum_{i \in \mathcal{SV}} \left(y_i - \sum_{j \in \mathcal{SV}} \alpha_j y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \right)$$

Interesting Twist:
Many α_i's are zero.
Only SVs have non-zero α_i's

Understanding the Dual

Maximize $J(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$

$\boxed{\text{s.t. } \alpha_i \geq 0 \ \forall i}$

$\boxed{\sum_i \alpha_i y_i = 0}$

Balances between the weight of constraints for different classes

Constraint weights (α_i's) cannot be negative

Understanding the Dual

Maximize $J(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boxed{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}$

$$\text{s.t. } \alpha_i \geq 0 \; \forall i$$

$$\sum_i \alpha_i y_i = 0$$

Points with different labels increase the sum

Points with same label decrease the sum

Measures the similarity between points

Intuitively, we should be more careful around points near the margin

Understanding the Dual

Maximize $J(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boxed{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}$

$$\text{s.t. } \alpha_i \geq 0 \; \forall i$$

$$\sum_i \alpha_i y_i = 0$$

In the solution, either:

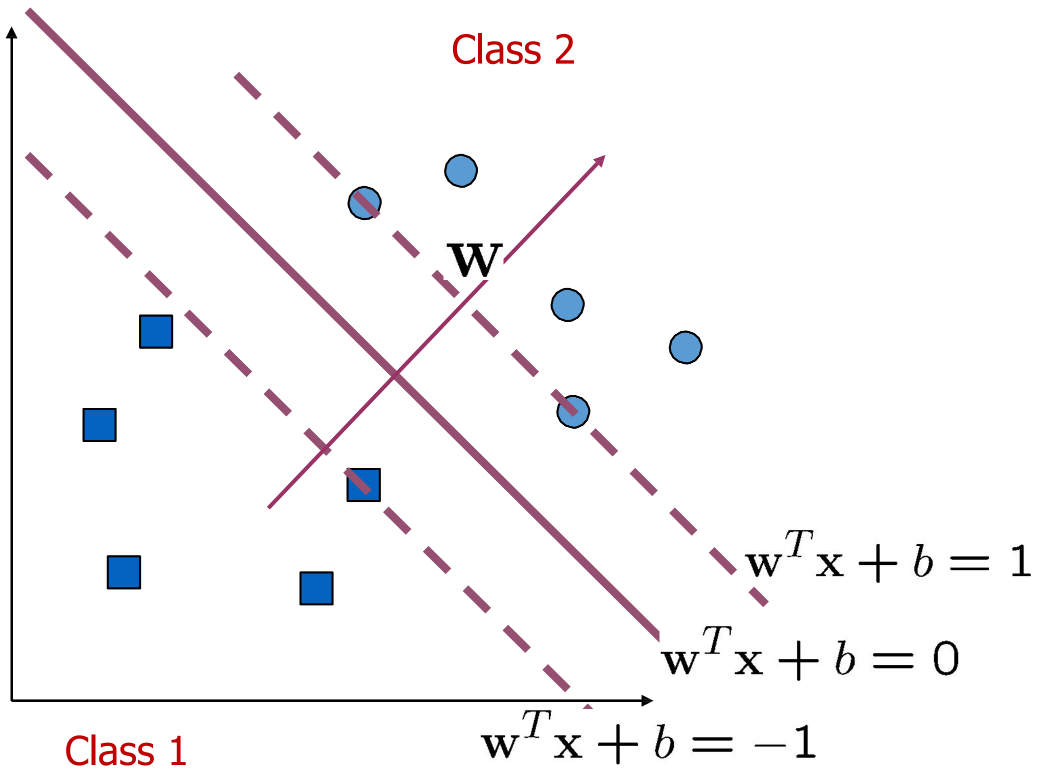
- $\alpha_i > 0$ and the constraint is tight $\left(y_i(\boldsymbol{\theta}^\top x_i) = 1\right)$
  - ➢ point is a support vector
- $\alpha_i = 0$
  - ➢ point is not a support vector

Deploying the Solution

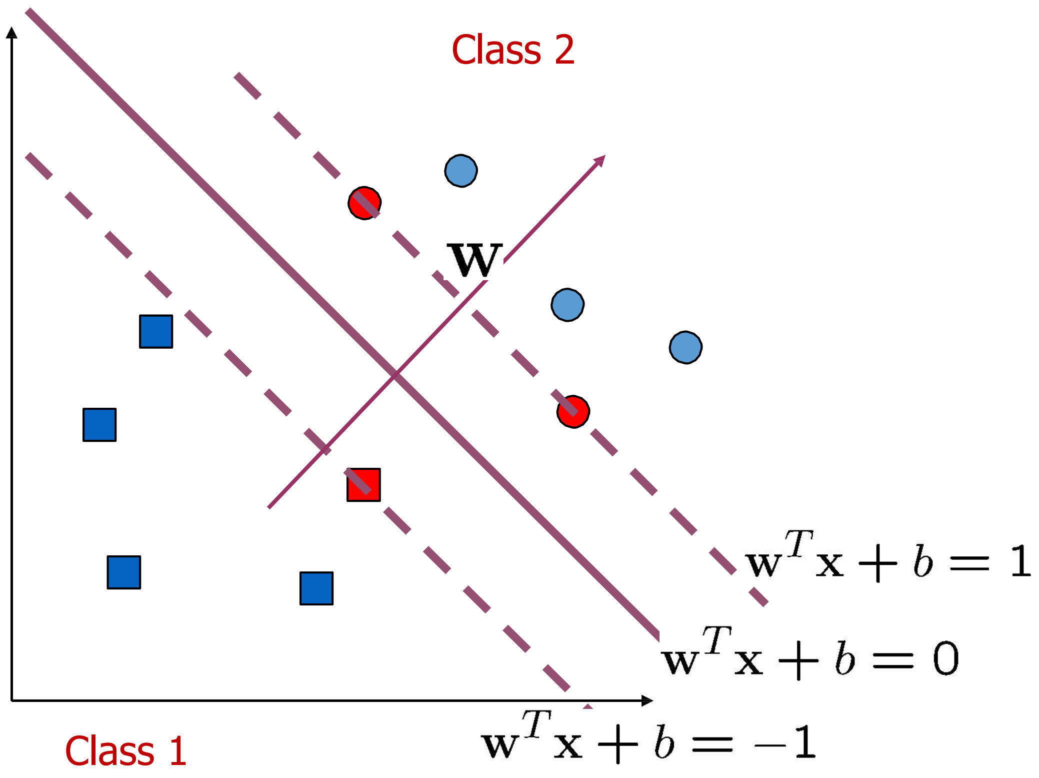
Given the optimal solution $\boldsymbol{\alpha}*$, optimal weights are

$$\boldsymbol{\theta}^* = \sum_{i \in \mathrm{SVs}} \alpha_i^* y_i \mathbf{x}_i$$

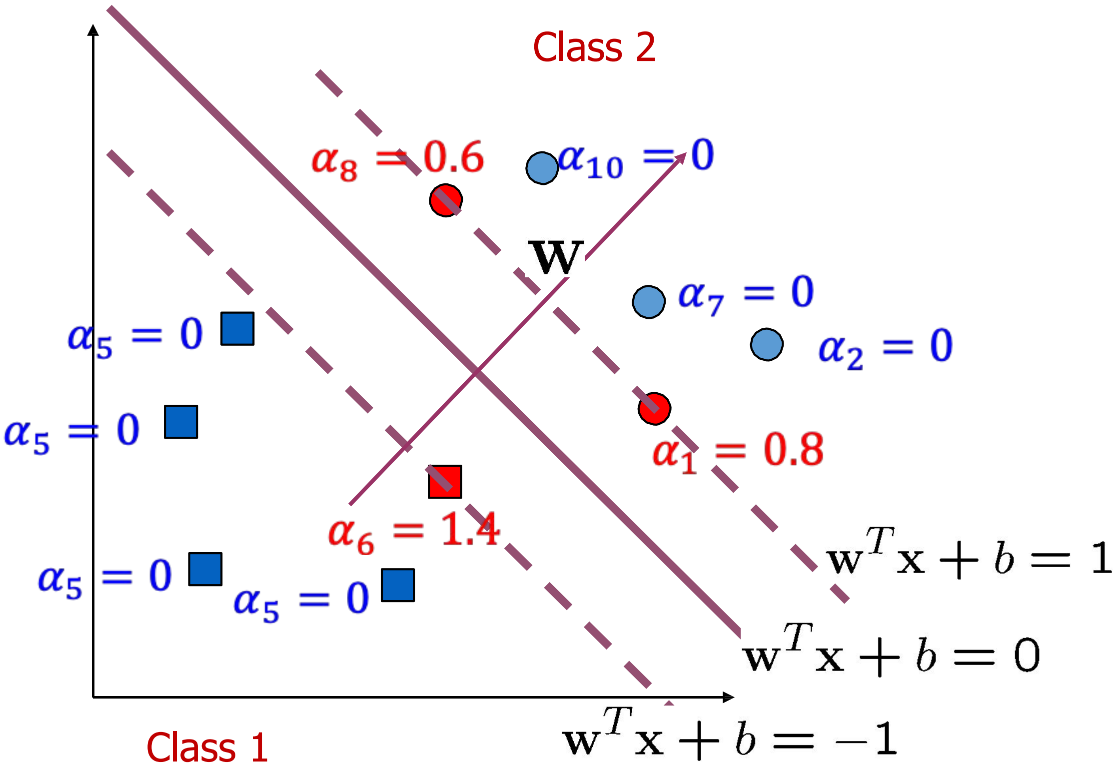
A Geometrical Interpretation



A Geometrical Interpretation



A Geometrical Interpretation



Characteristics of the Solution

For testing with a new data $\mathbf{z}$
Compute

$$\boldsymbol{\theta}^T \mathbf{z} = \sum_{j=1}^{s} \alpha_{t_j} y_{t_j} \left(\mathbf{x}_{t_j}^T \mathbf{z} \right)$$

Classify $\mathbf{z}$ as class 1 if the sum is positive, and class 2 otherwise. Note $\boldsymbol{\theta}$ need not be formed explicitly.

Given the optimal solution $\boldsymbol{\alpha}^*$, optimal weights are

$$\boldsymbol{\theta}^* = \sum_{i \in \mathrm{SVs}} \alpha_i^* y_i \mathbf{x}_i$$

Note: The computation relies on a dot product between the test point and the support vectors

What if Data Are Not Linearly Separable?

Cannot find $\boldsymbol{\theta}$ that satisfies. $y_i(\theta^T x_i) \geq 1 \; \forall i$

Introduce slack variables ξ_i

$$y_i(\theta^T x_i) \geq 1 - \xi_i \; \forall i$$

New Problem

$$\min_{\theta} \frac{1}{2} \sum_{j=1}^{d} \theta_j^2 + C \sum_{i} \xi_i$$

$$\text{s.t.}\; y_i(\theta^T x_i) \geq 1 - \xi_i, \quad \text{if}\; \forall i$$

Strengths of SVMs

- Good generalization in theory
- Good generalization in practice
- Work well with few training instances
- Find globally best model
- Efficient algorithms
- Amenable to the kernel trick …