Support Vector Machines & Kernels

Doing *really* well with linear decision surfaces

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.

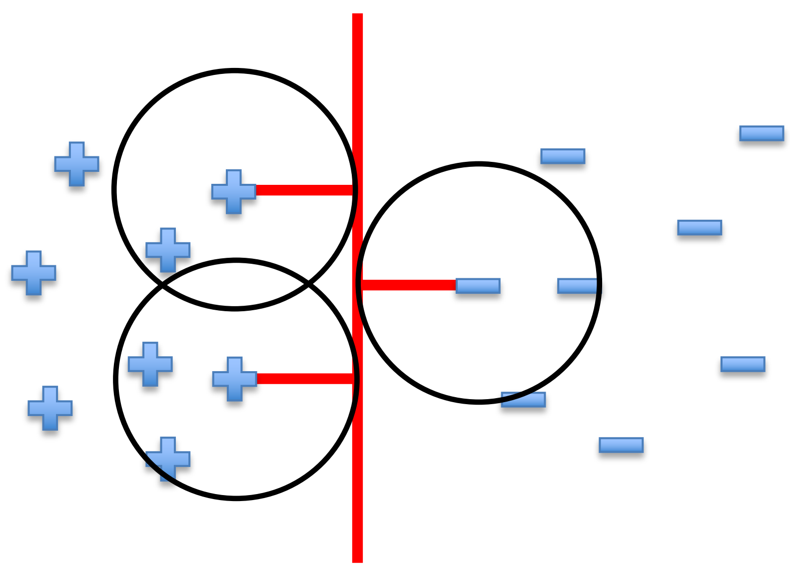
Adapted from slides by Tim Oates

Last Time: SVMs, Maximizing Margin

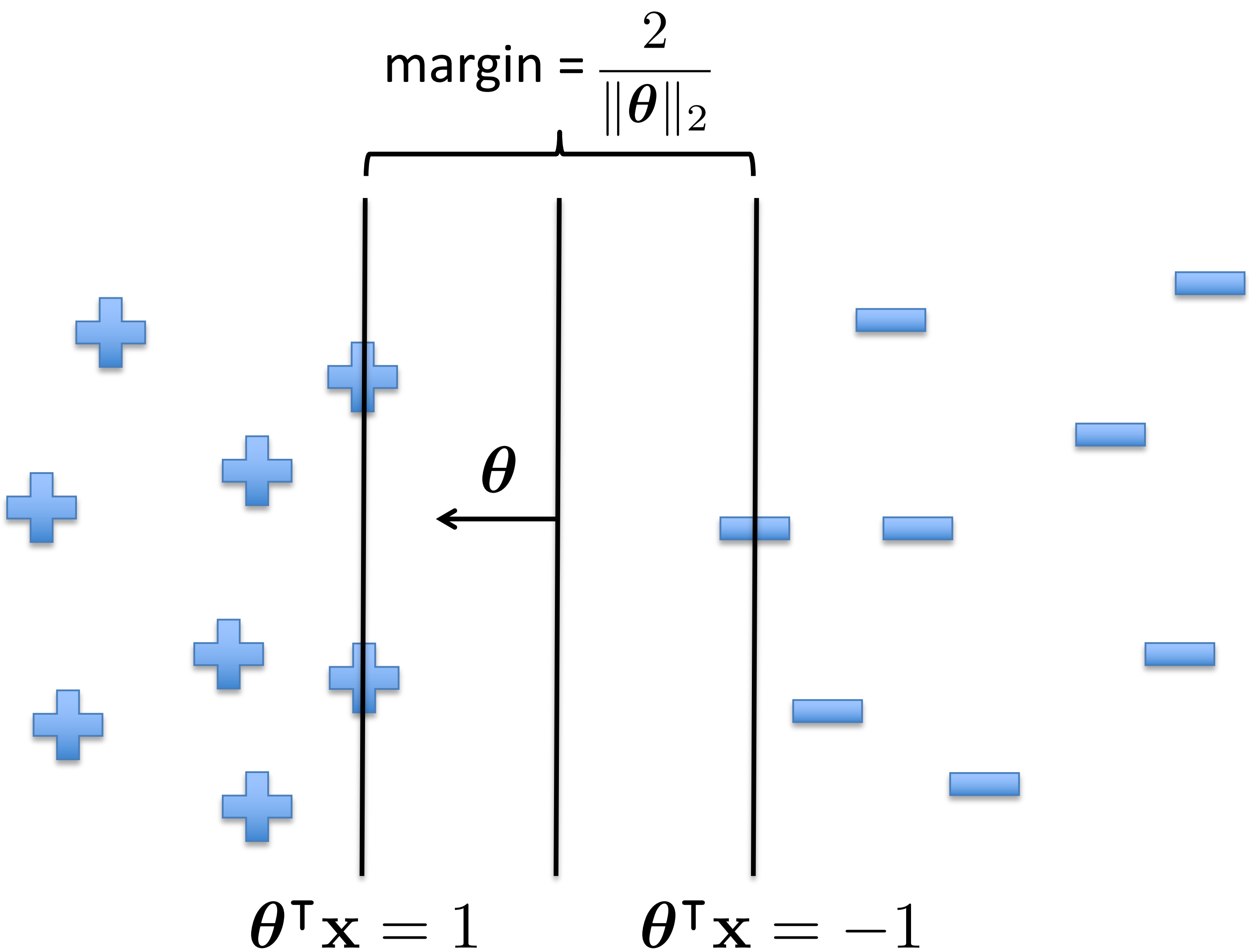
The SVM problem (assuming data is linearly separable):

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

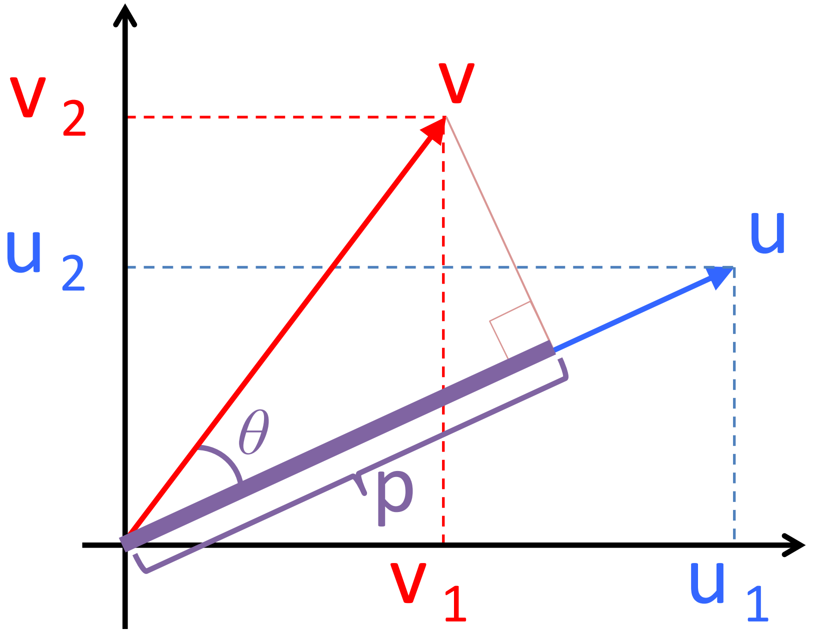
$$\text{s.t. } y_i(\boldsymbol{\theta}^\intercal \mathbf{x}_i) \geq 1 \quad \forall i$$



Maximum Margin Hyperplane



Vector Inner Product



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

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

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

Based on example by Andrew Ng

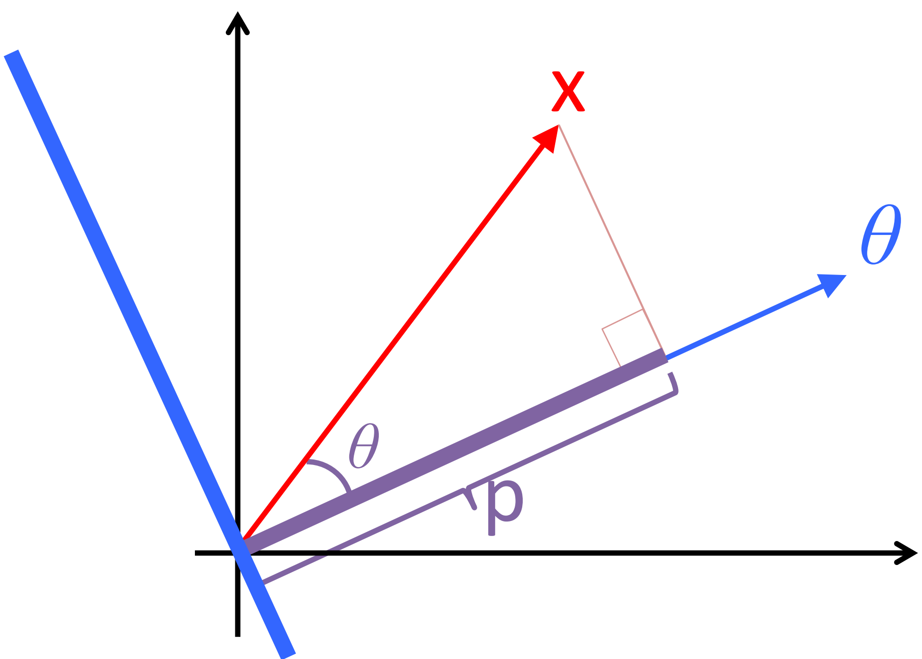
Understanding the Hyperplane

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

$$\text{s.t. } \boldsymbol{\theta}^\intercal \mathbf{x}_i \geq 1 \quad \text{if } y_i = 1$$

$$\boldsymbol{\theta}^\intercal \mathbf{x}_i \leq -1 \quad \text{if } y_i = -1$$

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



$$\boldsymbol{\theta}^\intercal \mathbf{x} = \|\boldsymbol{\theta}\|_2 \underbrace{\|\mathbf{x}\|_2 \cos\theta}_{p}$$

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

Based on example by Andrew Ng

Based on example by Andrew Ng

Maximizing the Margin

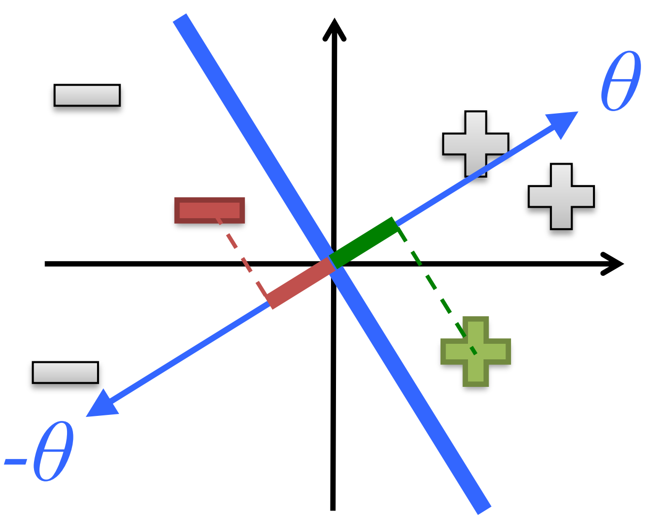
$$\min_{\boldsymbol{\theta}} \frac{1}{2} \sum_{j=1}^{d} \theta_j^2$$

$$\text{s.t.} \quad \boldsymbol{\theta}^\intercal \mathbf{x}_i \geq 1 \quad \text{if } y_i = 1$$

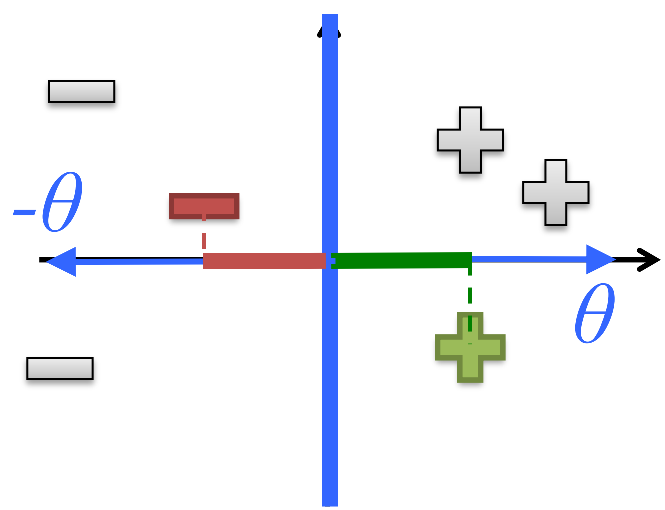
$$\boldsymbol{\theta}^\intercal \mathbf{x}_i \leq -1 \quad \text{if } y_i = -1$$

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 $\boldsymbol{\theta}$

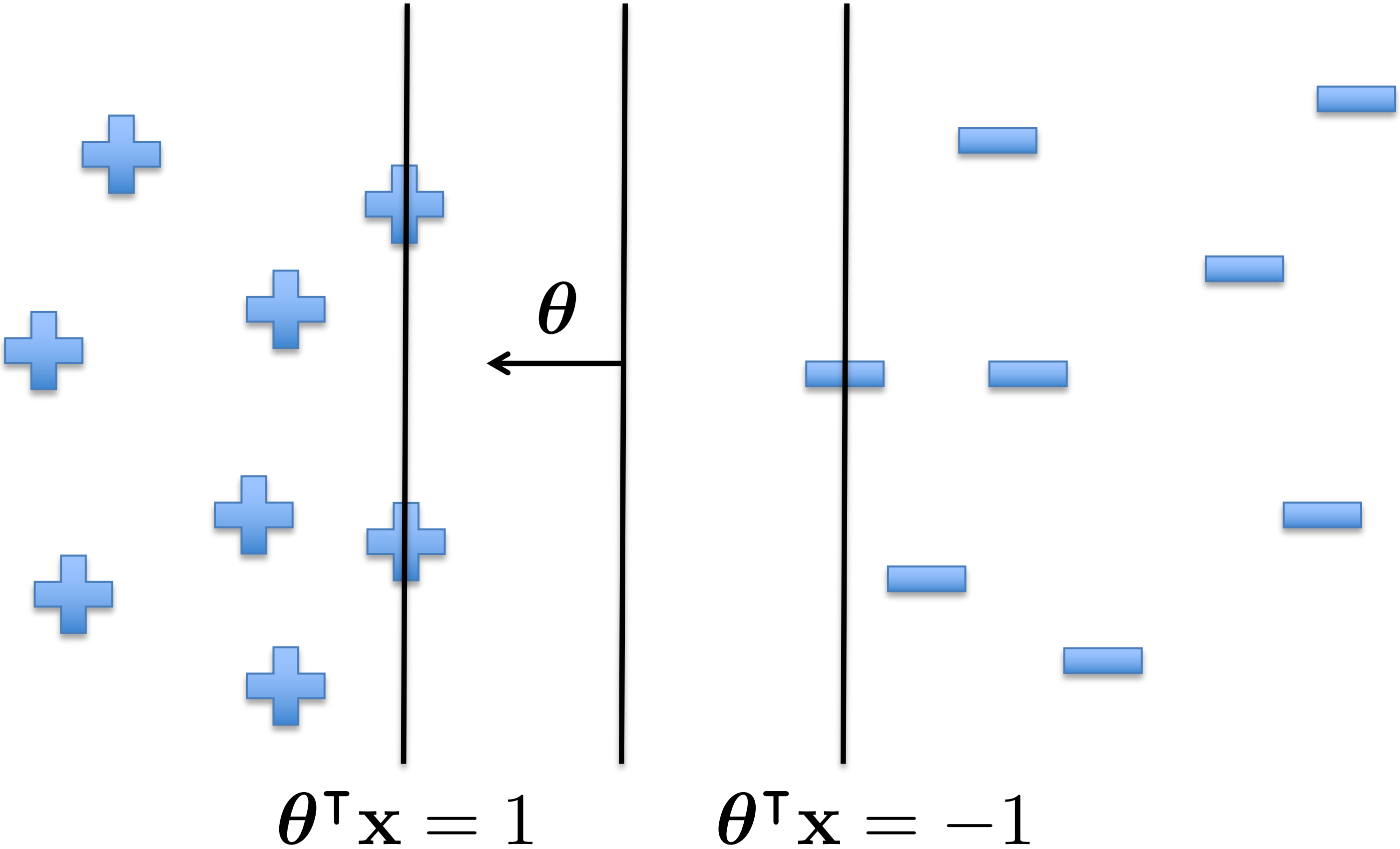


Since p is small, therefore $\|\boldsymbol{\theta}\|_2$ must be large to have $p\|\boldsymbol{\theta}\|_2 \geq 1$ (or ≤ -1)



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

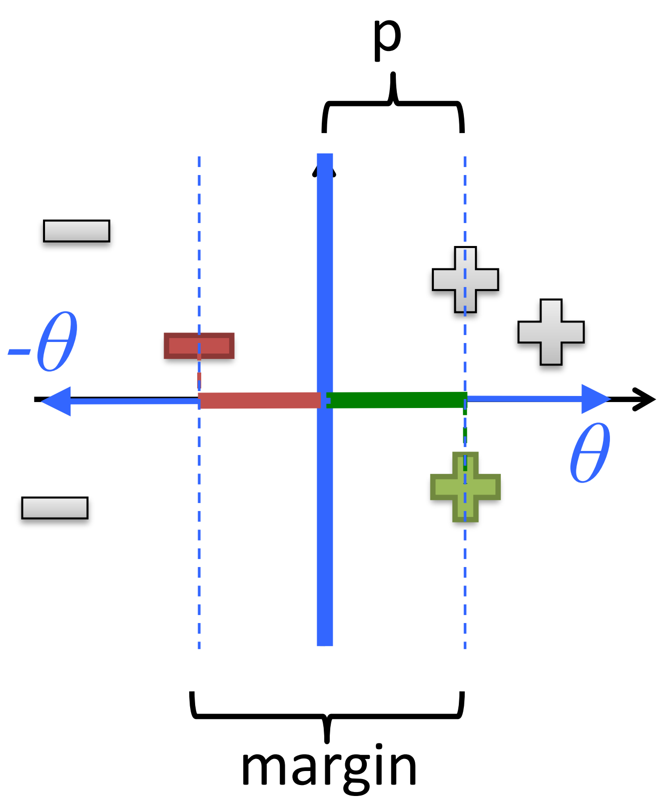
Support Vectors



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

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

$$\text{s.t. } y_i(\boldsymbol{\theta}^\intercal \mathbf{x}_i) \geq 1 \quad \forall i$$

Can solve it more efficiently by taking the Lagrangian dual

- Duality is a common idea in optimization
- It transforms a difficult optimization problem into a simpler one
- 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

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

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

- 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 ...

SVM Dual Representation

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

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

$$\sum_{i} \alpha_i y_i = 0$$

The decision function is given by

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

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

Understanding the Dual

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

s.t. $\alpha_i \geq 0 \quad \forall i$

$\sum_i \alpha_i y_i = 0$

Balances between the weight of constraints for different classes

Constraint weights (α_i's) cannot be negative

Understanding the Dual

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

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

$$\sum \alpha_i y_i = 0$$

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

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

$$\sum_i \alpha_i y_i = 0$$

In the solution, either:

- $\alpha_i > 0$ and the constraint is tight ($y_i(\boldsymbol{\theta}^\intercal \mathbf{x}_i) = 1$)
  - ➢ point is a support vector
- $\alpha_i = 0$
  - ➢ point is not a support vector

Deploying the Solution

Given the optimal solution **α***, optimal weights are

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

What if Data Are Not Linearly Separable?

- Cannot find $\boldsymbol{\theta}$ that satisfies $y_i(\boldsymbol{\theta}^\intercal \mathbf{x}_i) \geq 1 \quad \forall i$

- Introduce slack variables ξ_i

$$y_i(\boldsymbol{\theta}^\intercal \mathbf{x}_i) \geq 1 - \xi_i \quad \forall i$$

- New problem:

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

$$\text{s.t. } y_i(\boldsymbol{\theta}^\intercal \mathbf{x}_i) \geq 1 - \xi_i \quad \forall i$$

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 …