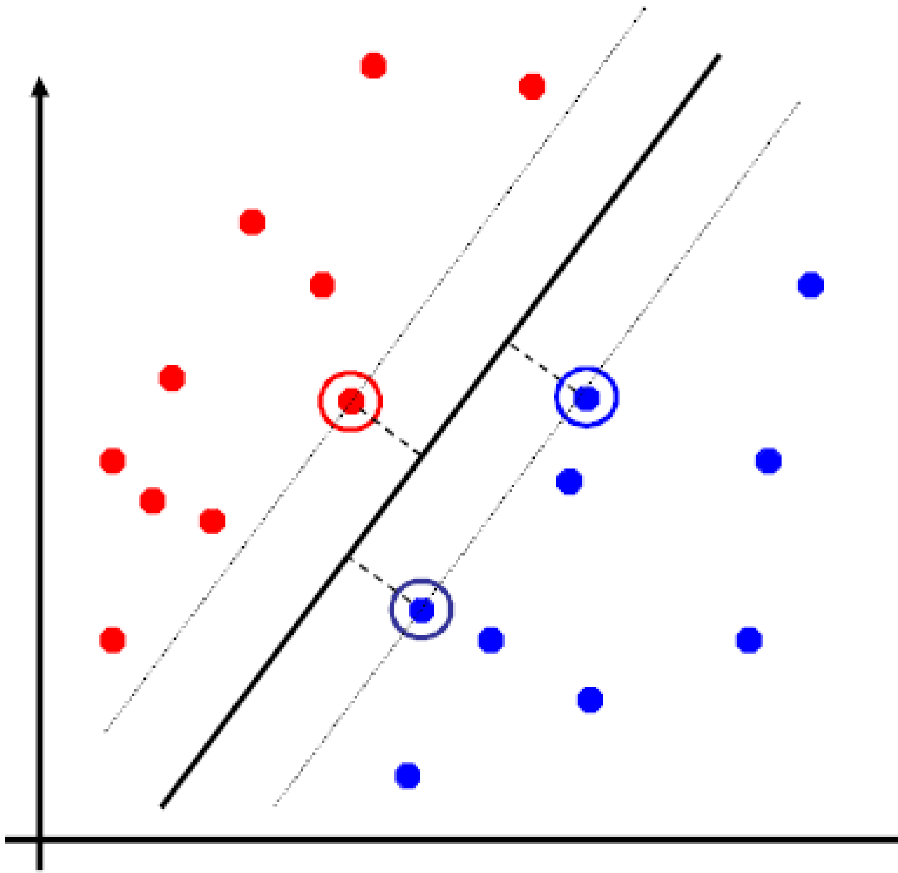
Support Vector Machines: What is the margin?

Mining of Massive Datasets
Leskovec, Rajaraman, and Ullman
Stanford University

Support Vector Machines

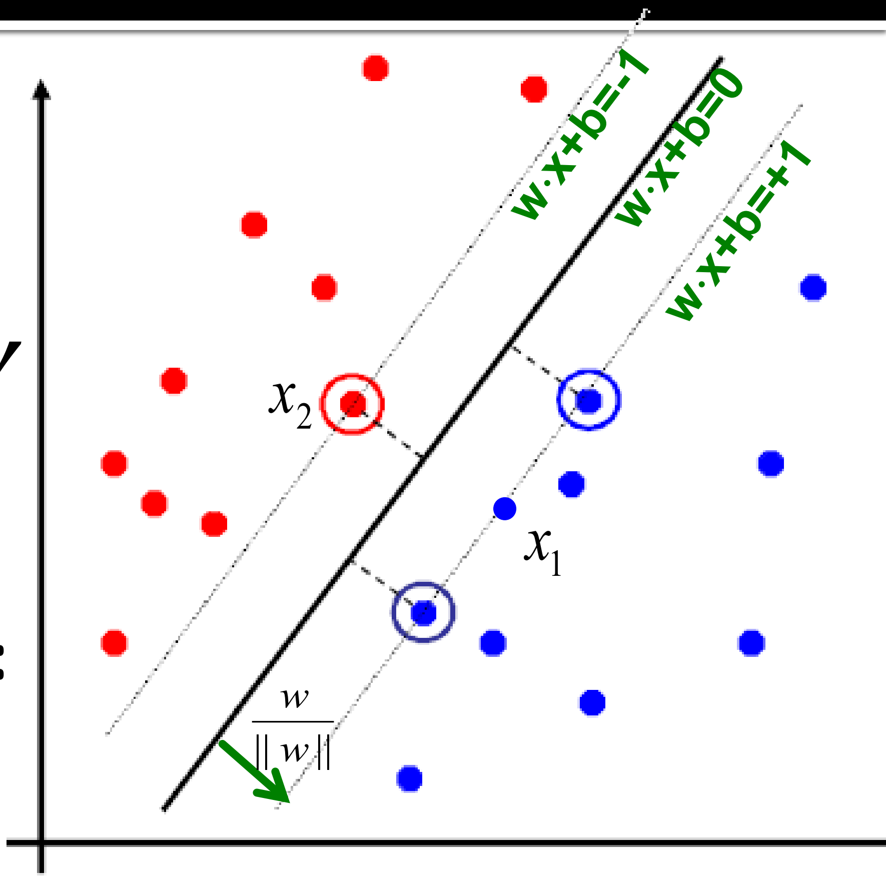
- **Separating hyperplane is defined by the support vectors**
  - Points on **+/-** planes from the solution
  - If you knew these points, you could ignore the rest
  - If no degeneracies, $d+1$ support vectors (for d dim. data)



Canonical Hyperplane: Problem

- **Problem:**
  - Let $(\boldsymbol{w} \cdot \boldsymbol{x} + \boldsymbol{b})\boldsymbol{y} = \boldsymbol{\gamma}$
    then $(\mathbf{2}\boldsymbol{w} \cdot \boldsymbol{x} + \mathbf{2}\boldsymbol{b})\boldsymbol{y} = \mathbf{2}\boldsymbol{\gamma}$
    - Scaling ***w*** increases margin!
- **Solution:**
  - **Work with normalized *w*:**

$$\boldsymbol{\gamma} = \left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \cdot \boldsymbol{x} + \boldsymbol{b}\right) \boldsymbol{y}$$

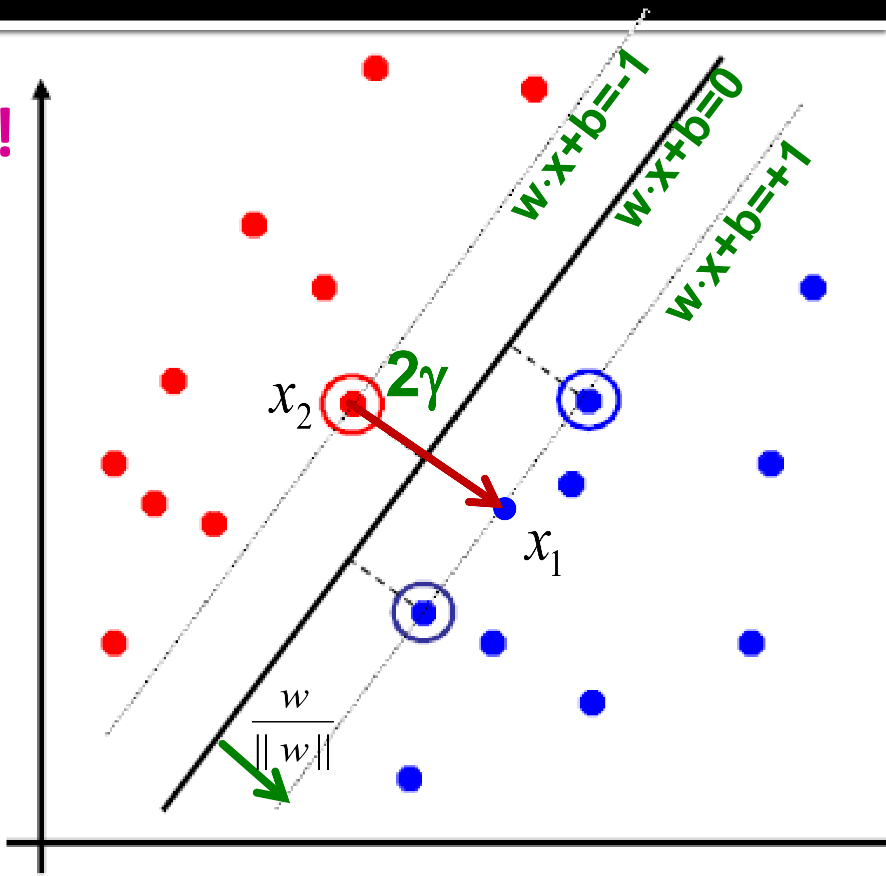


  - Let's also require **support vectors $\boldsymbol{x_j}$** to be on the plane defined by:
    $\boldsymbol{w} \cdot \boldsymbol{x_j} + \boldsymbol{b} = \pm \mathbf{1}$

$$\|w\| = \sqrt{\sum_{j=1}^{d} (w^{(j)})^2}$$

Canonical Hyperplane: Solution

- **Want to maximize margin γ!**
- **What is the relation between x_1 and x_2?**
  - $x_1 = x_2 + 2\gamma \frac{w}{||w||}$
  - **We also know:**
    - $w \cdot x_1 + b = +1$
    - $w \cdot x_2 + b = -1$
- **So:**
  - $w \cdot x_1 + b = +1$
  - $w\left(x_2 + 2\gamma \frac{w}{||w||}\right) + b = +1$
  - $\underbrace{w \cdot x_2 + b}_{-1} + 2\gamma \frac{w \cdot w}{||w||} = +1$



$$\boxed{\Rightarrow \gamma = \frac{\|w\|}{w \cdot w} = \frac{1}{\|w\|}}$$

Note: $w \cdot w = \|w\|^2$

Maximizing the Margin

- **We started with**

$$\max_{w,\gamma} \gamma$$

$$s.t. \forall i, y_i (w \cdot x_i + b) \geq \gamma$$

But *w* can be arbitrarily large!

- **We normalized and...**

$$\max \gamma \approx \max \frac{1}{\|w\|} \approx \min \|w\| \approx \min \tfrac{1}{2} \|w\|^2$$

- **Then:**

$$\min_w \tfrac{1}{2} \| w \|^2$$

$$s.t. \forall i, y_i (w \cdot x_i + b) \geq 1$$

This is called SVM with “hard” constraints