The Maximum Margin Principle

Announcements

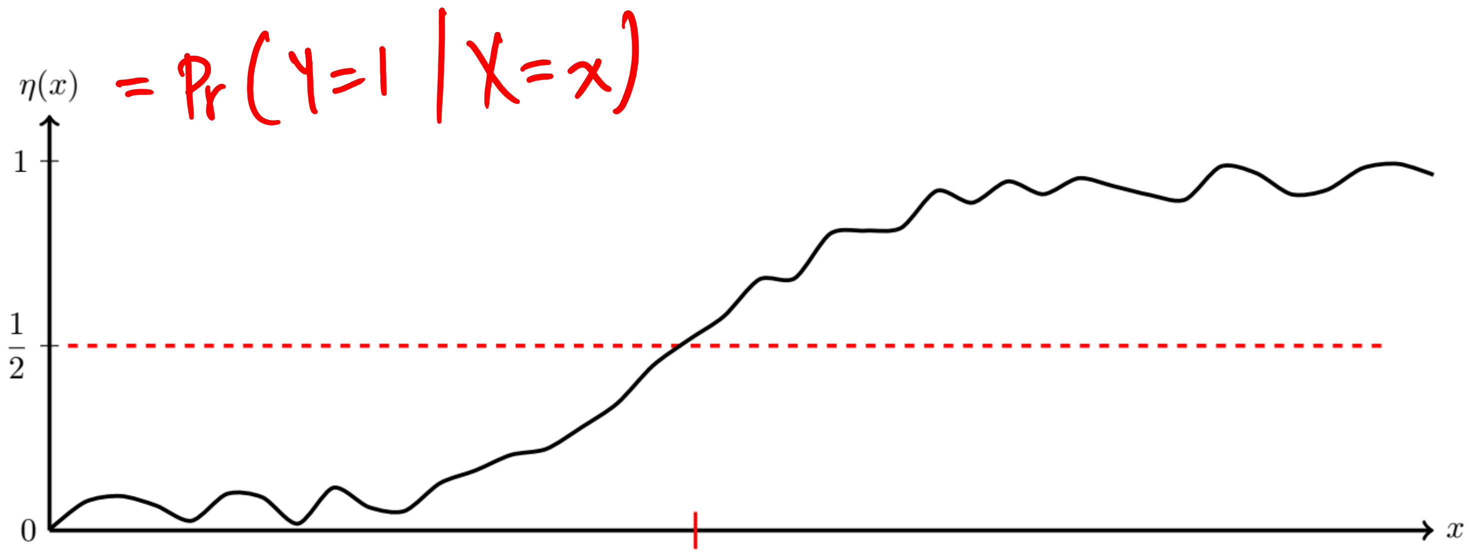
- None

Outline

- Hyperplanes
- Max-margin hyperplanes
- Optimal soft-margin hyperplanes

Drawback of Plug-in Classifiers

- Plug-in methods require estimation of (conditional) densities or mass functions, which can be more difficult than estimating a decision boundary



$\eta(x)$ is quite complicated but the decision regions are simple and η is smooth near $1/2$

Linear Classifiers

- Binary classification
- Training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$.
- Assume the labels are -1 and 1
- Recall a linear classifier has the form

$x_i \in \mathbb{R}^d$

$w \in \mathbb{R}^d$

$b \in \mathbb{R}$

$$f(\boldsymbol{x}) = \operatorname{sign}(w^T x + b)$$

$$\operatorname{sign}(t) = \begin{cases} 1 & t \geq 0 \\ -1 & t < 0 \end{cases}$$

- How can we use the training data to directly optimize for $\boldsymbol{w}$ and b?

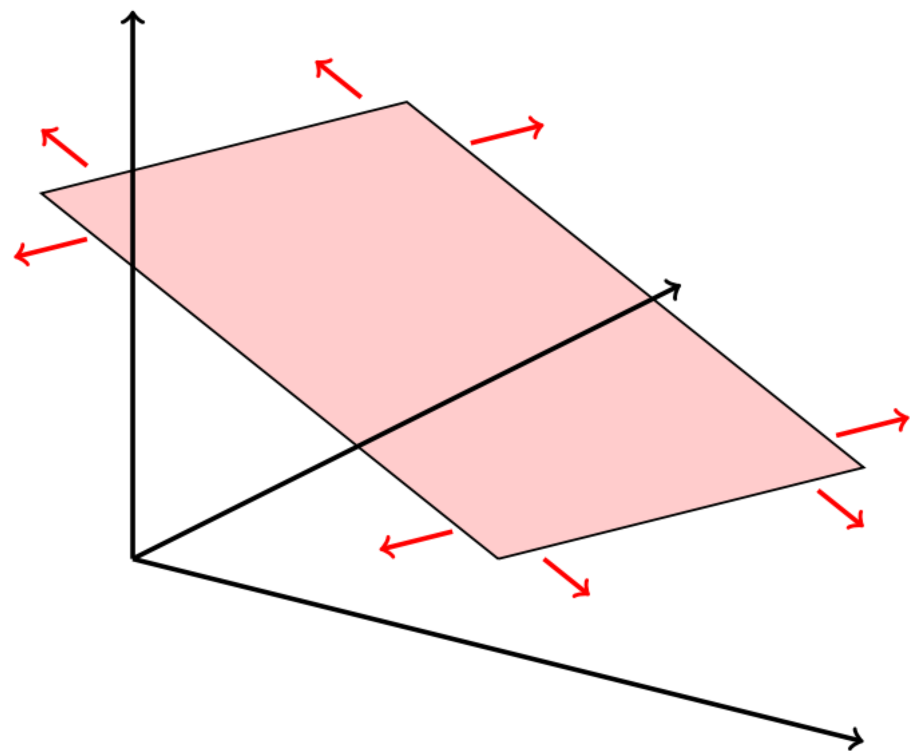
Hyperplanes

- A *hyperplane* is a subset of $\mathbb{R}^d$ of the form

$$H = \{x : w^T x + b = 0\}$$

for some $\boldsymbol{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$.

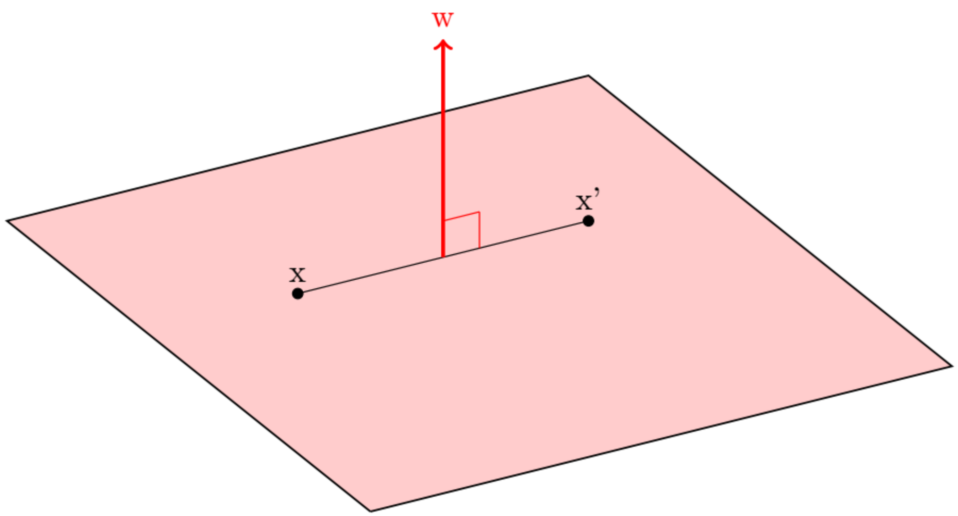
- In general, a hyperplane is an *affine subspace* of dimension $d - 1$



Normal Vectors

- The vector $\boldsymbol{w}$ is orthogonal to the hyperplane, and for this reason is called a *normal vector*.
- To say that $\boldsymbol{w}$ is ~~orthogonal to a hyperplane means~~ a normal vector that it is orthogonal to every vector that lies in the hyperplane. Every such vector can be written as the difference of two points $\boldsymbol{x}$ and $\boldsymbol{x}'$ in the hyperplane.

$$\left.\begin{array}{l} w^T x + b = 0 \\ w^T x' + b = 0 \end{array}\right\} \Rightarrow w^T(x - x') = 0 \Rightarrow w \perp x - x'$$



Distance from Point to Hyperplane

- Let $\mathcal{H} = \{\boldsymbol{x} \mid \boldsymbol{w}^T \boldsymbol{x} + b = 0\}$. The distance from $\boldsymbol{z} \notin \mathcal{H}$ to $\mathcal{H}$ is

$$\frac{|w^T z + b|}{\|w\|}$$

- Let $\mathcal{H} = \{\boldsymbol{x} \mid x_1 - 5x_2 + 5x_3 - 7 = 0\} \subseteq \mathbb{R}^3$. The distance from $\mathcal{H}$ to

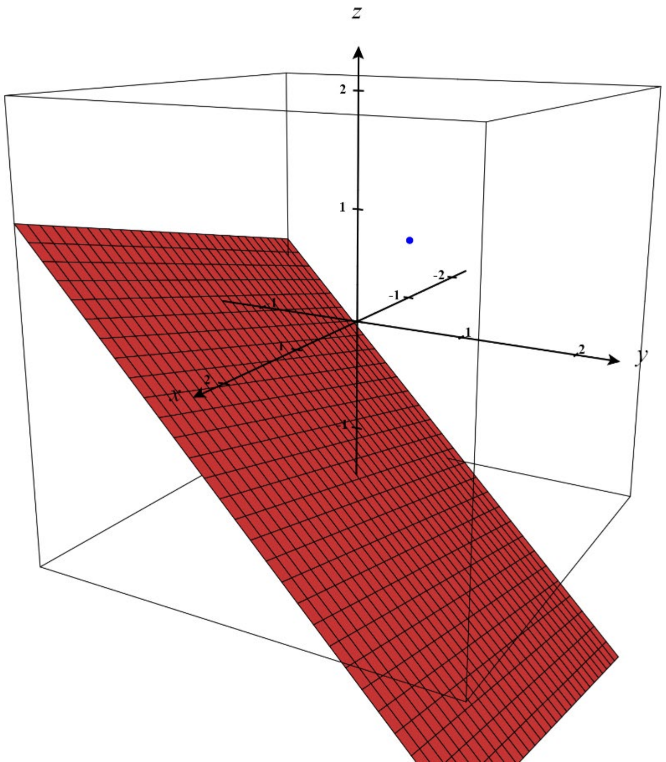
$$\boldsymbol{z} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \qquad w = \begin{bmatrix} 1 \\ -5 \\ 5 \end{bmatrix}, \; b = -7$$

is

$$|w^T z + b| = \big|1 - 5 + 5 - 7\big| = 6$$

$$\|w\| = \sqrt{1^2 + (-5)^2 + 5^2} = \sqrt{51}$$

distance : $\frac{6}{\sqrt{51}}$



Separating Hyperplanes

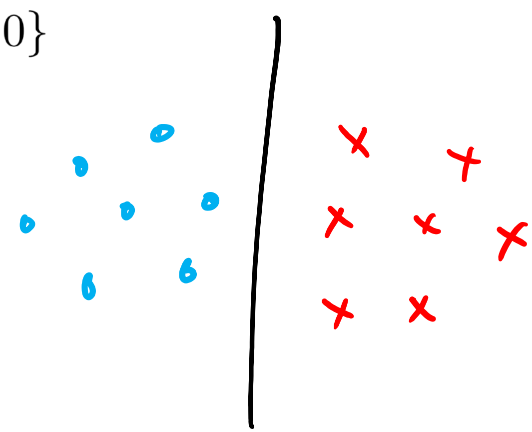
- Let $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ be training data for a binary classification problem
- Assume $y_i \in \{-1, 1\}$.
- We say the training data are *linearly separable* if there exist $\boldsymbol{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$ such that

$$\forall i \qquad y_i \big(\underbrace{w^T x_i + b}_{\text{take sign to get the predicted label}}\big) > 0$$

- In this case we refer to

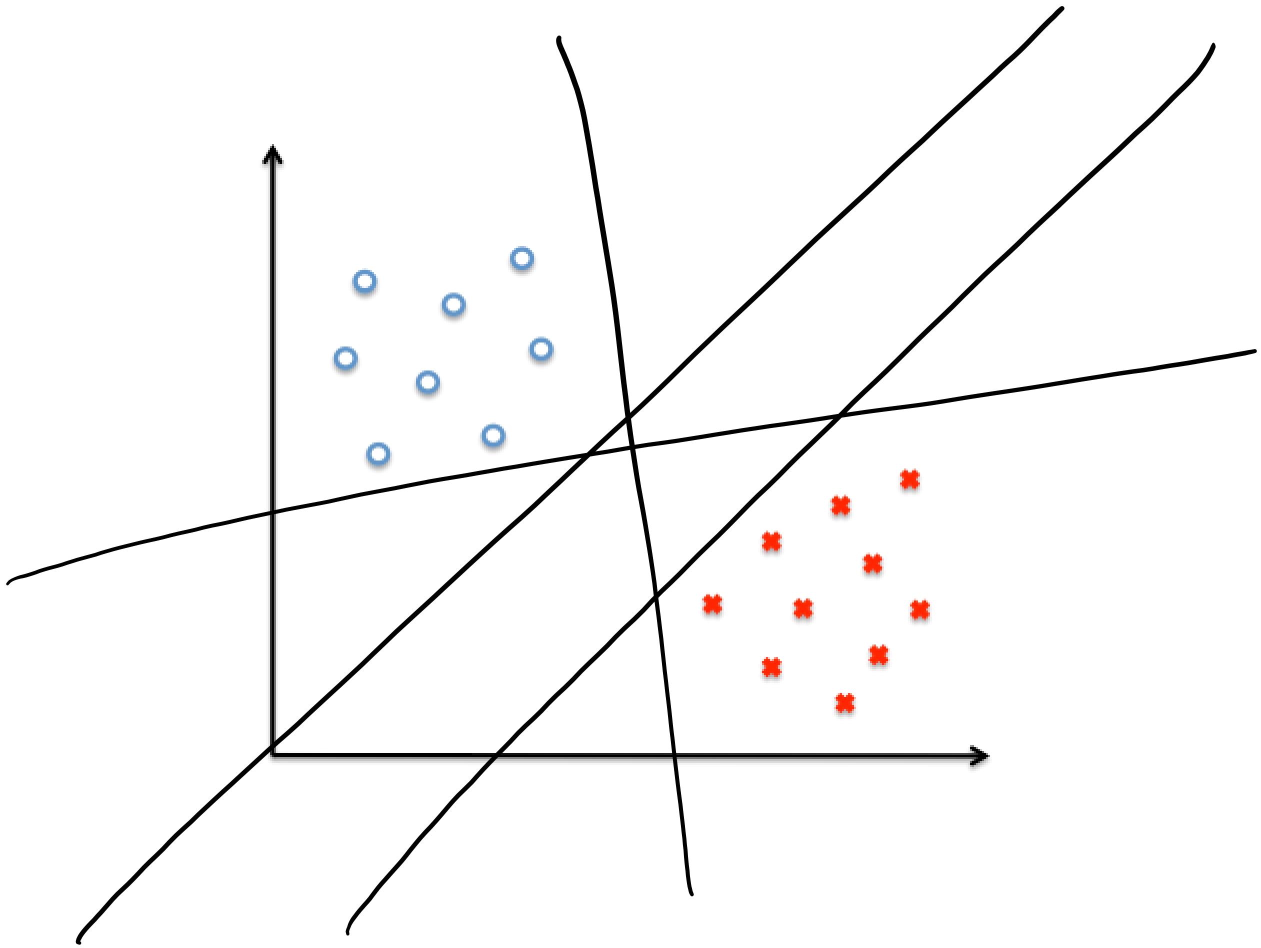
$$\mathcal{H} = \{\boldsymbol{x} : \boldsymbol{w}^T \boldsymbol{x} + b = 0\}$$

as a *separating hyperplane.*



Separating Hyperplanes

- Are all separating hyperplanes equally good?



Poll: Margin of a Hyperplane

- Let $\mathcal{H} = \{\boldsymbol{x} : \boldsymbol{w}^T\boldsymbol{x} + b = 0\}$ be a separating hyperplane.
- The *margin* ρ of a $\mathcal{H}$ is the distance from $\mathcal{H}$ to the nearest training point $\boldsymbol{x}_i$.
- **Poll:** A formula for ρ is

  (A) $\rho(\boldsymbol{w}, b) = \min_{\boldsymbol{z}\in\mathbb{R}^d} \frac{|\boldsymbol{w}^T\boldsymbol{z}+b|}{\|\boldsymbol{w}\|}$

  (B) $\rho(\boldsymbol{w}, b) = \max_{\boldsymbol{z}\in\mathbb{R}^d} \frac{|\boldsymbol{w}^T\boldsymbol{z}+b|}{\|\boldsymbol{w}\|}$

  (C) $\rho(\boldsymbol{w}, b) = \min_{i=1,\ldots,n} \frac{|\boldsymbol{w}^T\boldsymbol{x}_i+b|}{\|\boldsymbol{w}\|}$ ✓

  (D) $\rho(\boldsymbol{w}, b) = \max_{i=1,\ldots,n} \frac{|\boldsymbol{w}^T\boldsymbol{x}_i+b|}{\|\boldsymbol{w}\|}$

Max-Margin Hyperplane

- The *margin* ρ of a separating hyperplane is the distance from the hyperplane to the nearest training point:

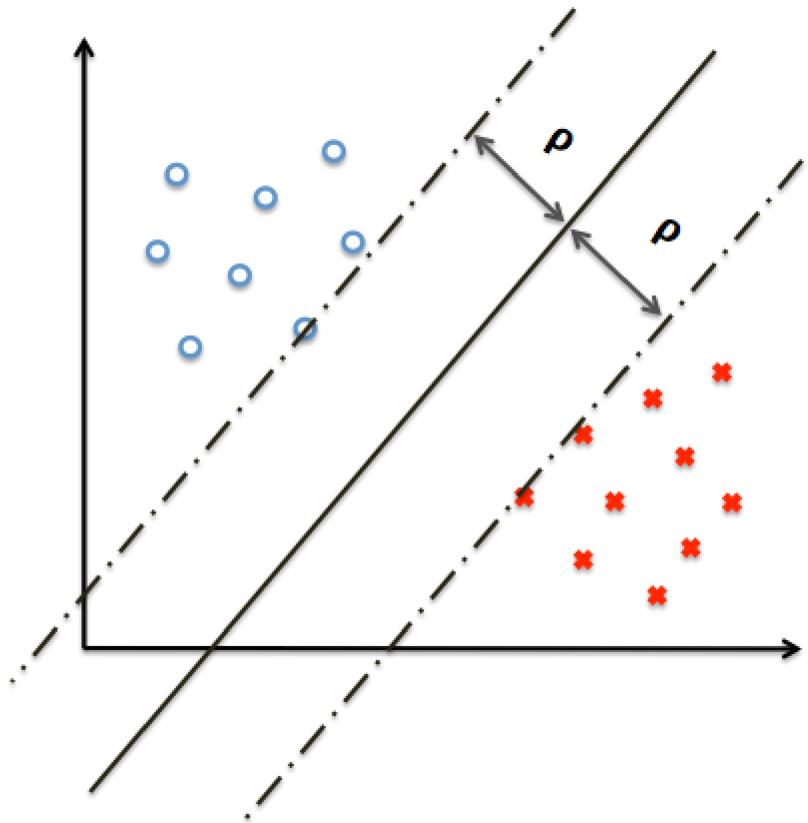
$$\rho(\boldsymbol{w}, b) := \min_{i=1,\dots,n} \frac{|w^T x_i + b|}{\|w\|}$$

- The *maximum margin* or *optimal* separating hyperplane is the solution of

$$\max_{w,b} \left(\min_{i=1,\dots,n} \frac{|w^T x_i + b|}{\|w\|} \right)$$

$$\text{s.t.} \quad y_i (w^T x_i + b) > 0 \quad \forall i = 1,\dots,n$$

Max-Margin Hyperplane



Canonical Form

- A separating hyperplane is said to be in *canonical form* if $\boldsymbol{w}$ and b are such that

$$
\begin{aligned}
y_i(w^T x_i + b) &\geq 1 \quad \forall i \\
y_i(w^T x_i + b) &= 1 \quad \text{for some } i
\end{aligned}
$$

- Every separating hyperplane can be represented in canonical form. If $\mathcal{H} = \{\boldsymbol{x} : \boldsymbol{w}^T\boldsymbol{x} + b = 0\}$ and $\alpha > 0$, then

$$\mathcal{H} = \{x : (\alpha w)^T x + (\alpha b) = 0\}$$

and

$$\forall x \quad \operatorname{sign}\{(\alpha w)^T x + (\alpha b)\} = \operatorname{sign}\{w^T x + b\}$$

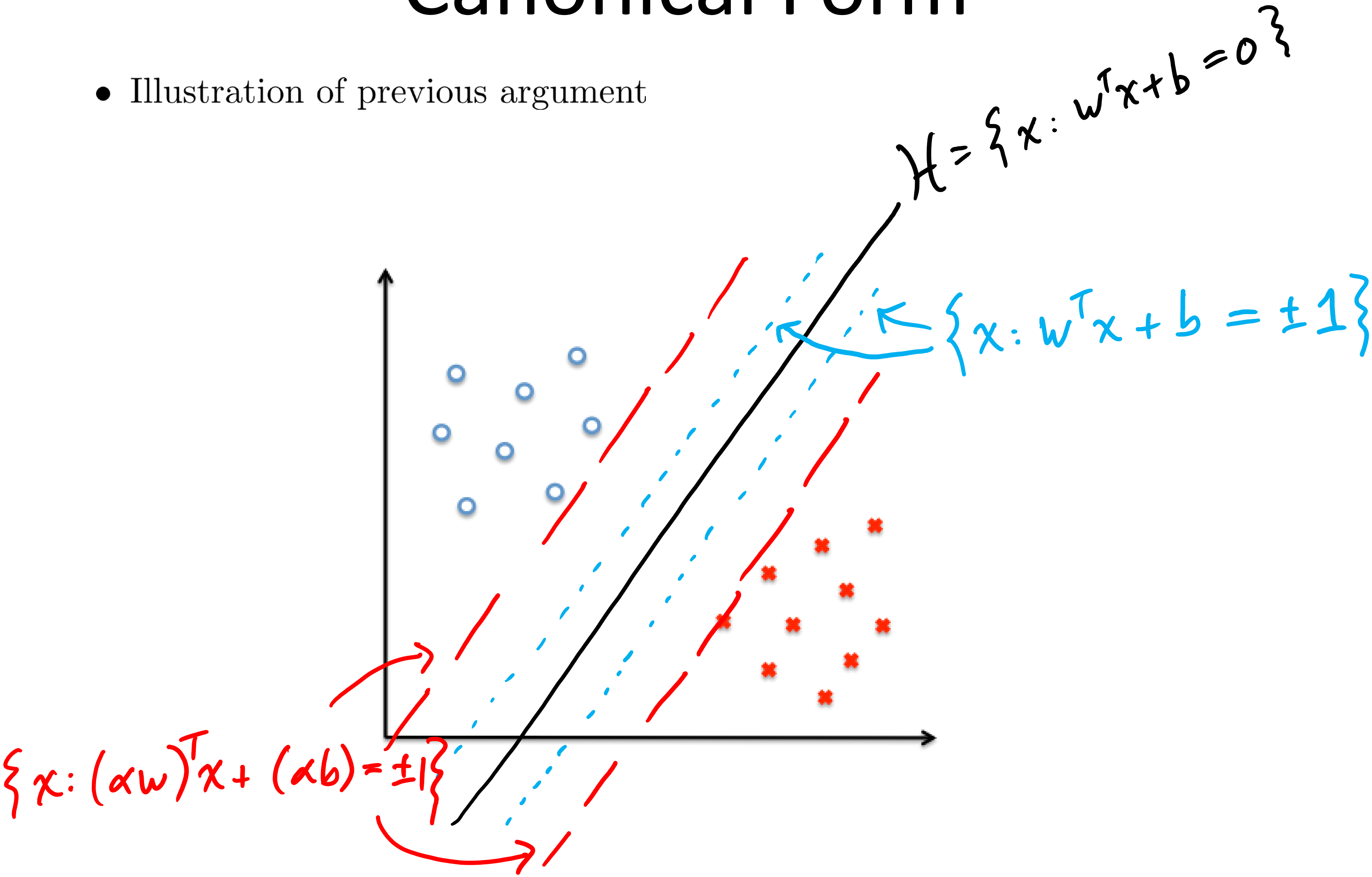
- Thus we can always scale $\boldsymbol{w}$ and b such that the smallest value of

$$y_i((\alpha w)^T x_i + (\alpha b))$$

is 1

Canonical Form

- Illustration of previous argument



Max-Margin Hyperplane

- This allows us to write the max-margin hyperplane

$$\max_{\boldsymbol{w},b} \left(\min_{i=1,\dots,n} \frac{|\boldsymbol{w}^T \boldsymbol{x}_i + b|}{\|\boldsymbol{w}\|} \right)$$
$$\text{s.t. } \forall i \quad y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1$$
$$\exists i \quad y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) = 1.$$

- Equivalently,

$$\max_{w,b} \frac{1}{\|w\|}$$
$$\text{s.t. } \forall i \quad y_i(w^T x_i + b) \geq 1$$
$$\exists i \quad y_i(w^T x_i + b) = 1$$

$$\Longrightarrow$$

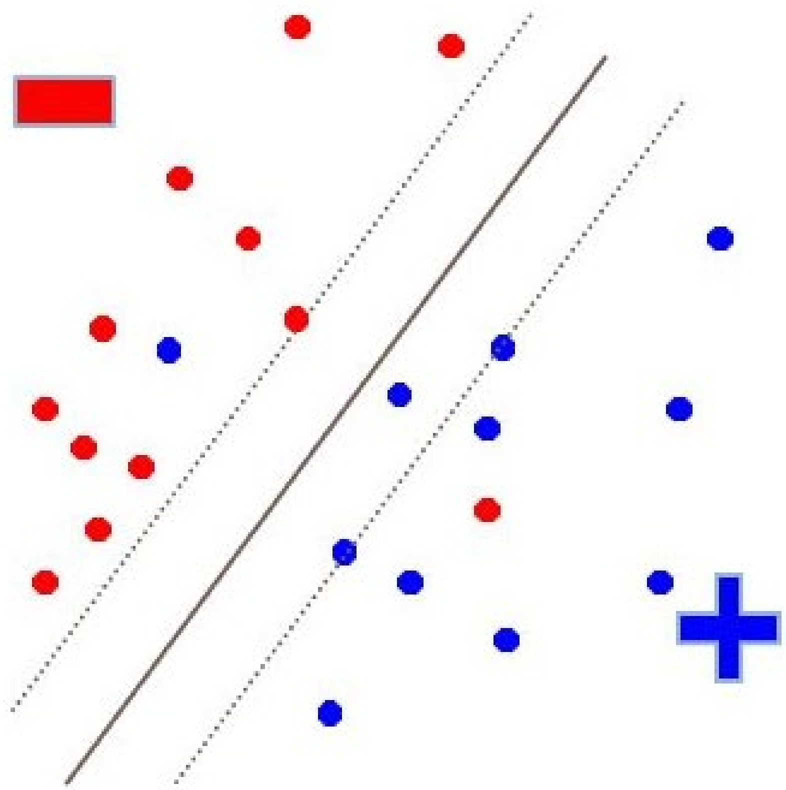
$$\min_{w,b} \frac{1}{2}\|w\|^2$$
$$\text{s.t. } \forall i \quad y_i(w^T x_i + b) \geq 1$$
$$\cancel{\exists i \quad y_i(w^T x_i + b) = 1}$$

satisfied even if omitted

- This is an example of a quadratic program

Non-Separable Data

- What if the training data are not linearly separable?



ksee

Optimal Soft-Margin Hyperplane

$$\xi = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix}$$

- Introduce *slack variables* $\xi_1, \ldots, \xi_n \geq 0$.
- The *optimal soft-margin hyperplane* is the solution of

$$\min_{\boldsymbol{w}, b, \boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n} \xi_i$$

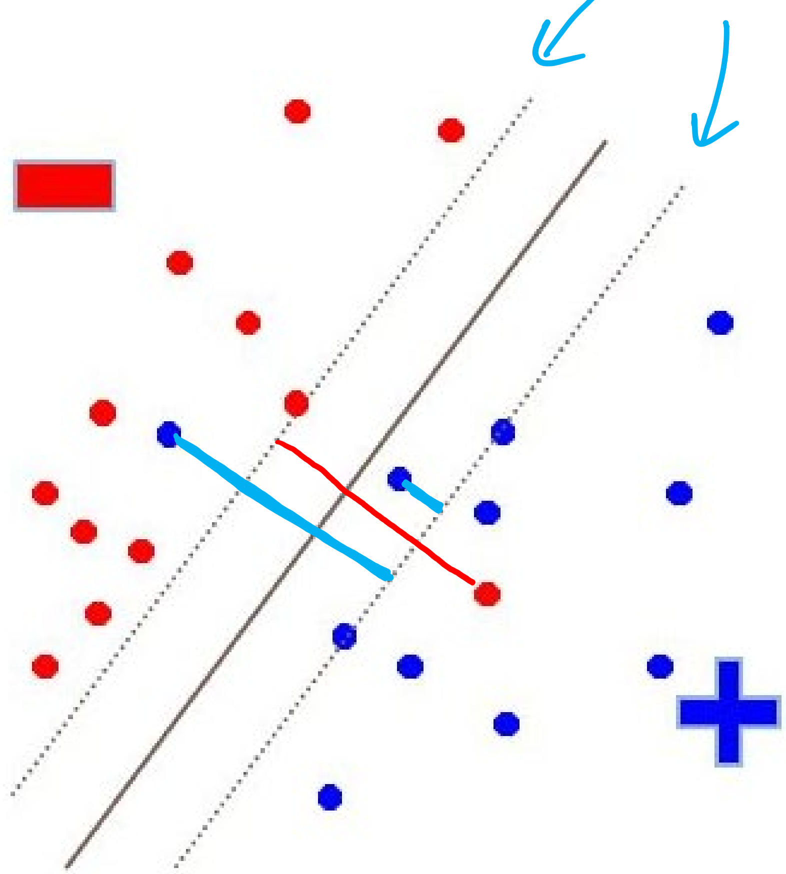
$$\text{s.t. } y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 - \xi_i \qquad \forall i = 1, \ldots, n$$

$$\xi_i \geq 0 \qquad \forall i = 1, \ldots, n$$

- C is a user-defined parameter
- This is another quadratic program

Optimal Soft-Margin Hyperplane

$\{x : w^T x + b = \pm 1\}$



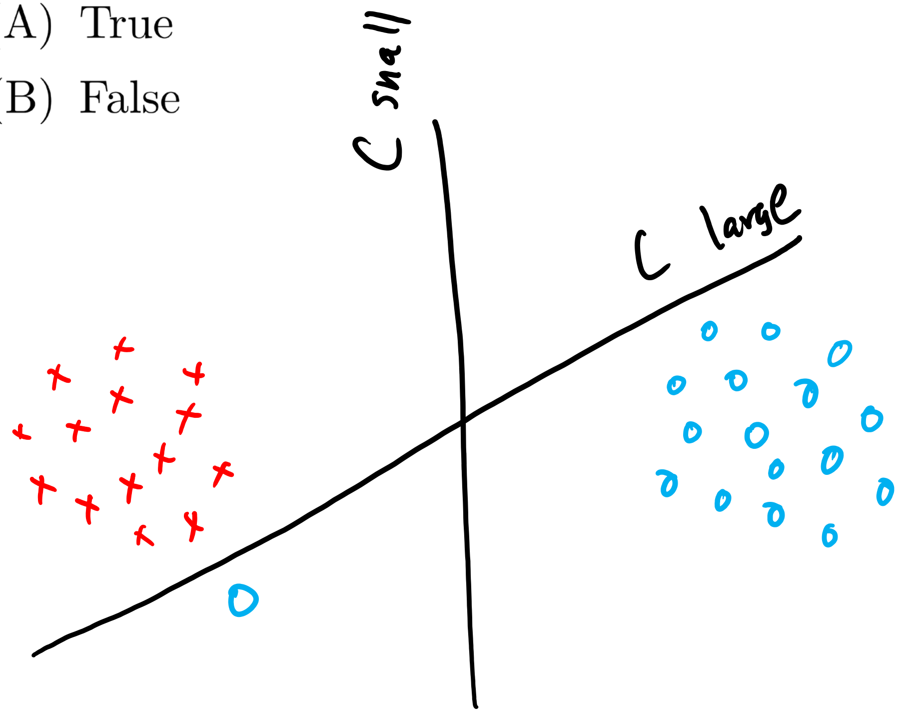
Poll 2

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i$$
$$\text{s.t. } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i \qquad \forall i = 1, \ldots, n$$
$$\xi_i \geq 0 \qquad \forall i = 1, \ldots, n$$

- True or False: As C increases, the solution becomes more sensitive to outliers like the one shown

  (A) True

  (B) False



Poll 3

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i$$

$$\text{s.t. } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i \qquad \forall i = 1,\ldots,n$$

$$\xi_i \geq 0 \qquad \forall i = 1,\ldots,n$$

- True or False: If $C = 0$, the OSM hyperplane recovers the max-margin hyperplane in the case of linearly separable data.

  (A) True

  (B) False ✓

$w = 0$ is optimal

Closing Thoughts

- The optimal soft margin hyperplane classifier is a special case of a much more general classifier that we will study soon: the support vector machine