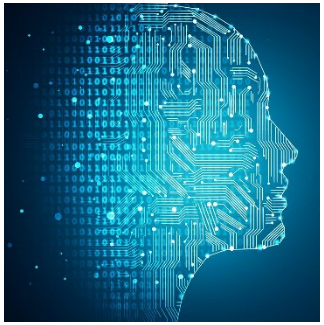
Machine Learning

Separating Hyperplanes

Kevin Moon (kevin.moon@usu.edu)

STAT/CS 5810/6655

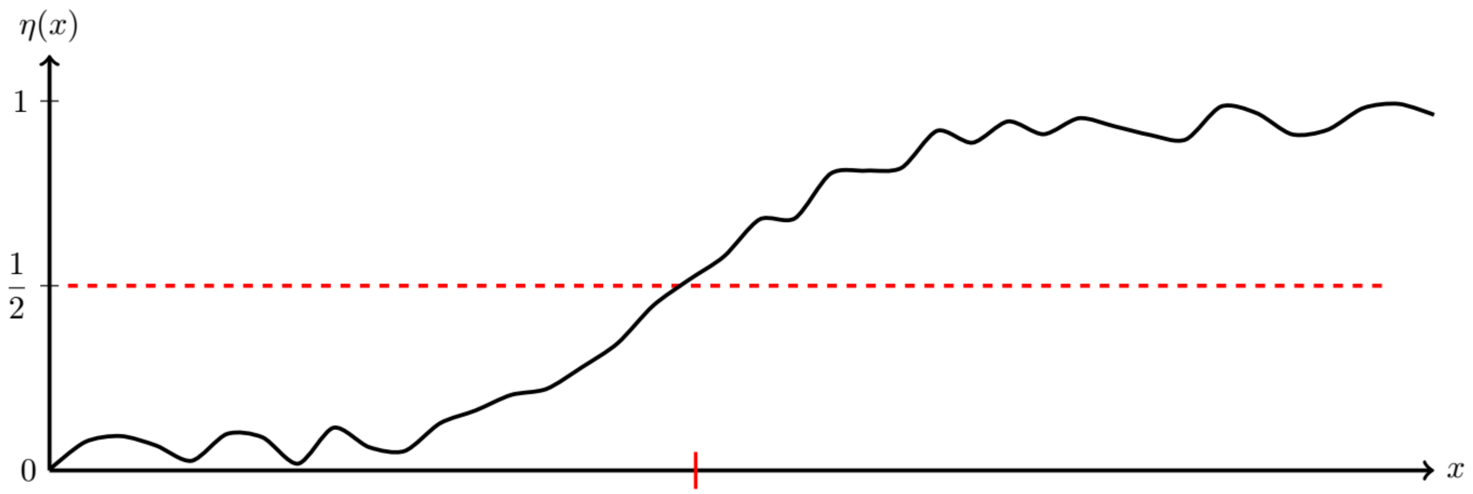


Outline

1. Hyperplanes
2. Max-margin hyperplanes
3. Optimal soft-margin hyperplanes
4. ERM and the optimal soft-margin hyperplane

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
- Maxim attribute to Vladimir Vapnik, a machine learning pioneer (paraphrased): "Don't solve a harder problem than you have to."



$\eta(x)$ 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

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

  - where $\text{sign}(t) = \begin{cases} 1 & t \geq 0 \\ -1 & t < 0 \end{cases}$
- How can we use the training data to directly optimize for $\boldsymbol{w}$ and b?

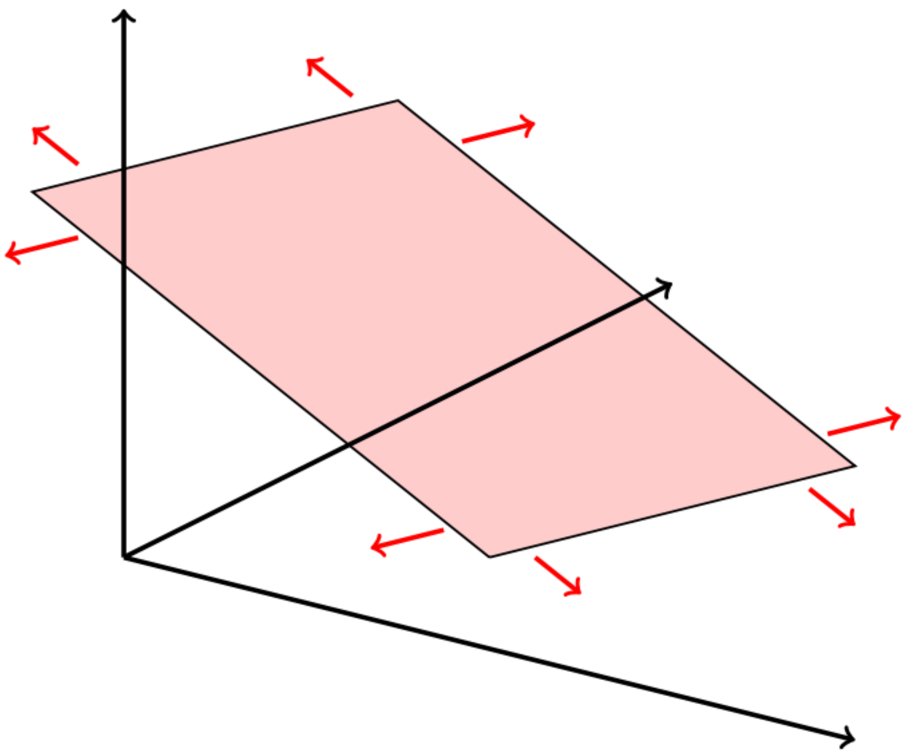
$$\min_{\mathbf{w},\mathrm{b}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{y_i \neq \text{sign}(\boldsymbol{w}^T \boldsymbol{x}_i + b)\}}$$

Hyperplanes

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

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

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

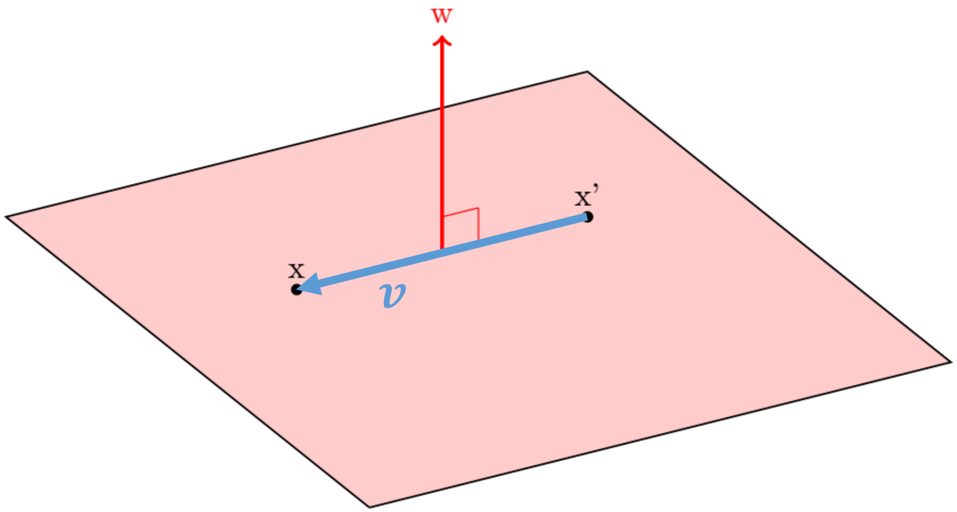


- In general, a hyperplane is an affine subspace of dimensions $d - 1$

Normal vectors

- The vector $\boldsymbol{w}$ is orthogonal to the hyperplane and is called a normal vector
- *Proof*: Suppose $\boldsymbol{v}$ is parallel to $\mathcal{H}$. Then we can write $v = \boldsymbol{x} - \boldsymbol{x}'$ where $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{H}$. Then

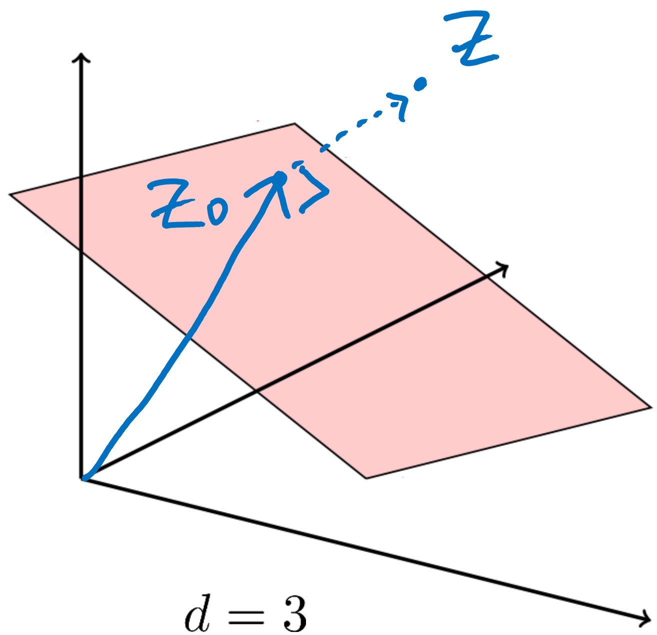
$$\begin{aligned}
\boldsymbol{w}^T\boldsymbol{v} &= \boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{x}') \\
&= \boldsymbol{w}^T\boldsymbol{x} - \boldsymbol{w}^T\boldsymbol{x}' \\
&= \boldsymbol{w}^T\boldsymbol{x} + b \\
&\quad -(\boldsymbol{w}^T\boldsymbol{x}' + b) \\
&= 0
\end{aligned}$$



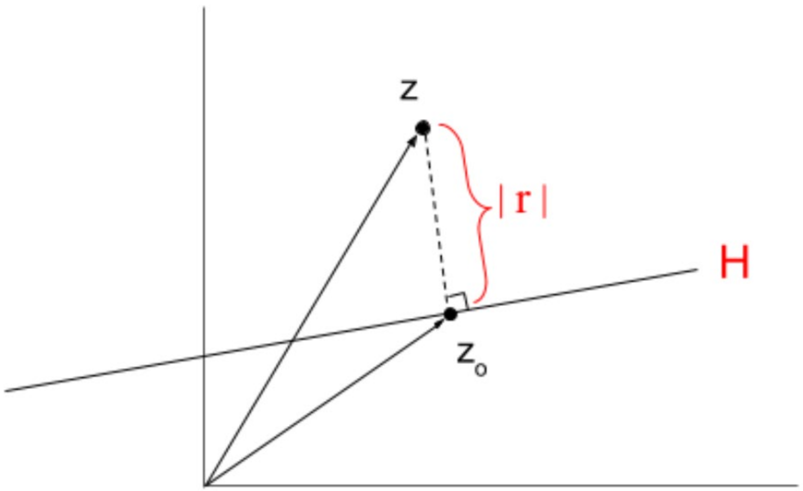
Distance to a Hyperplane

- Given a hyperplane $\mathcal{H} = \{\boldsymbol{x} | \boldsymbol{w}^T\boldsymbol{x} + b = 0\}$ and a point $\boldsymbol{z} \notin \mathcal{H}$, what is the distance of $\boldsymbol{z}$ to $\mathcal{H}$?
- We can write $\boldsymbol{z}$ as

$$\boldsymbol{z} = \boldsymbol{z}_0 + r\frac{\boldsymbol{w}}{||\boldsymbol{w}||}$$



$d = 3$



$d = 2$

Distance to a Hyperplane

- Then

$$\begin{aligned}\boldsymbol{w}^T\boldsymbol{z}+b &= \boldsymbol{w}^T\left(\boldsymbol{z}_0+r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)+b\\ &= \boldsymbol{w}^T\boldsymbol{z}_0+b+r\frac{\boldsymbol{w}^T\boldsymbol{w}}{\|\boldsymbol{w}\|}\\ &= r\|\boldsymbol{w}\|\end{aligned}$$

- Hence,

$$|r|=\frac{\left|\boldsymbol{w}^T\boldsymbol{z}+b\right|}{\|\boldsymbol{w}\|}$$

Separating Hyperplanes

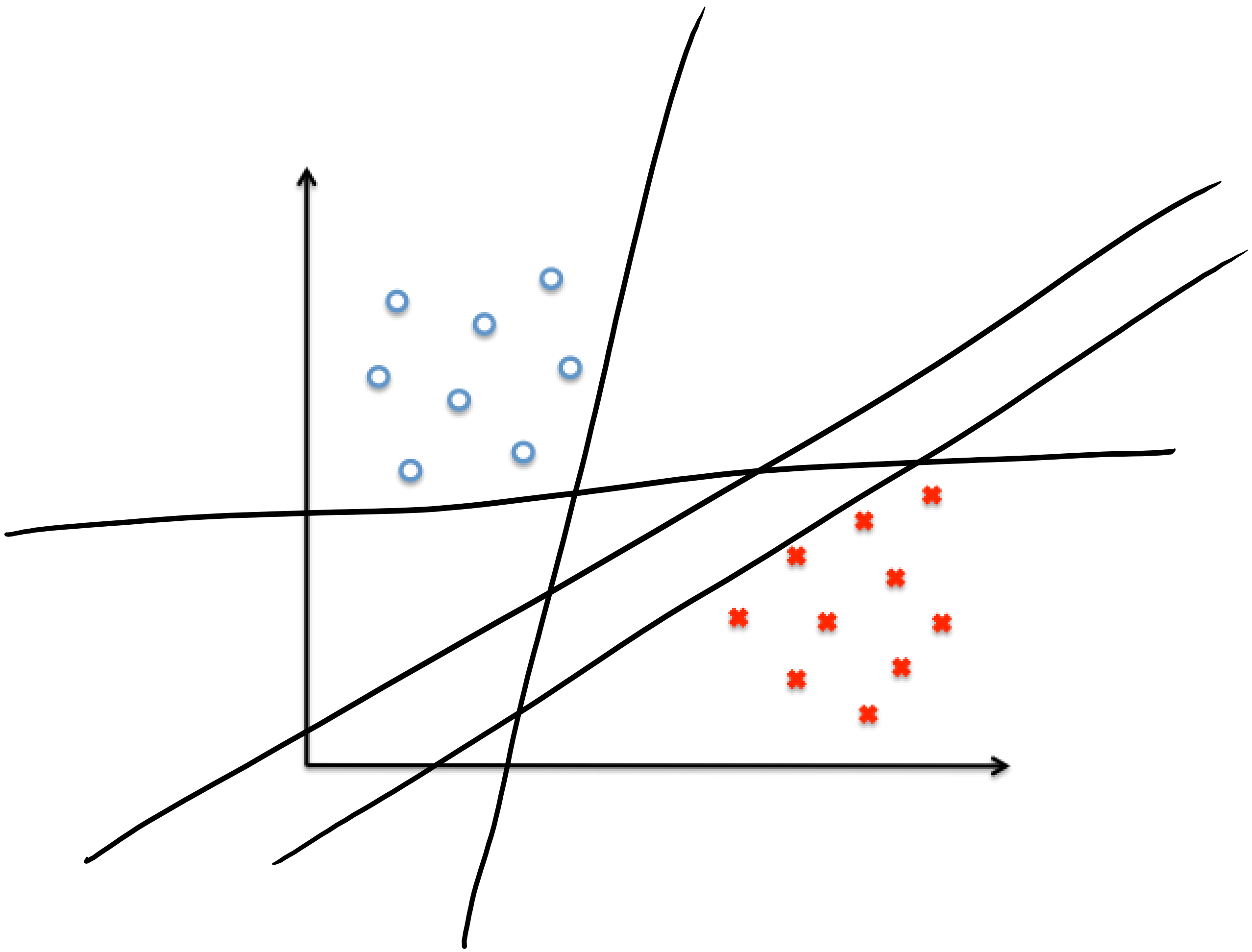
- Let $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ be training data for a binary classification problem
- Assume $y_i \in \{-1,1\}$
- We say the training data are *linearly separable* if there exists $\boldsymbol{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$ such that

$$y_i\left(\boldsymbol{w}^T \boldsymbol{x}_i + b\right) > 0 \ \ \forall i$$

- In this case we refer to $\mathcal{H} = \left\{\boldsymbol{x} \middle| \boldsymbol{w}^T \boldsymbol{x} + b = 0\right\}$ as a *separating hyperplane*

Separating Hyperplanes

- Are all separating hyperplanes equally good?



Max-Margin Hyperplane

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

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

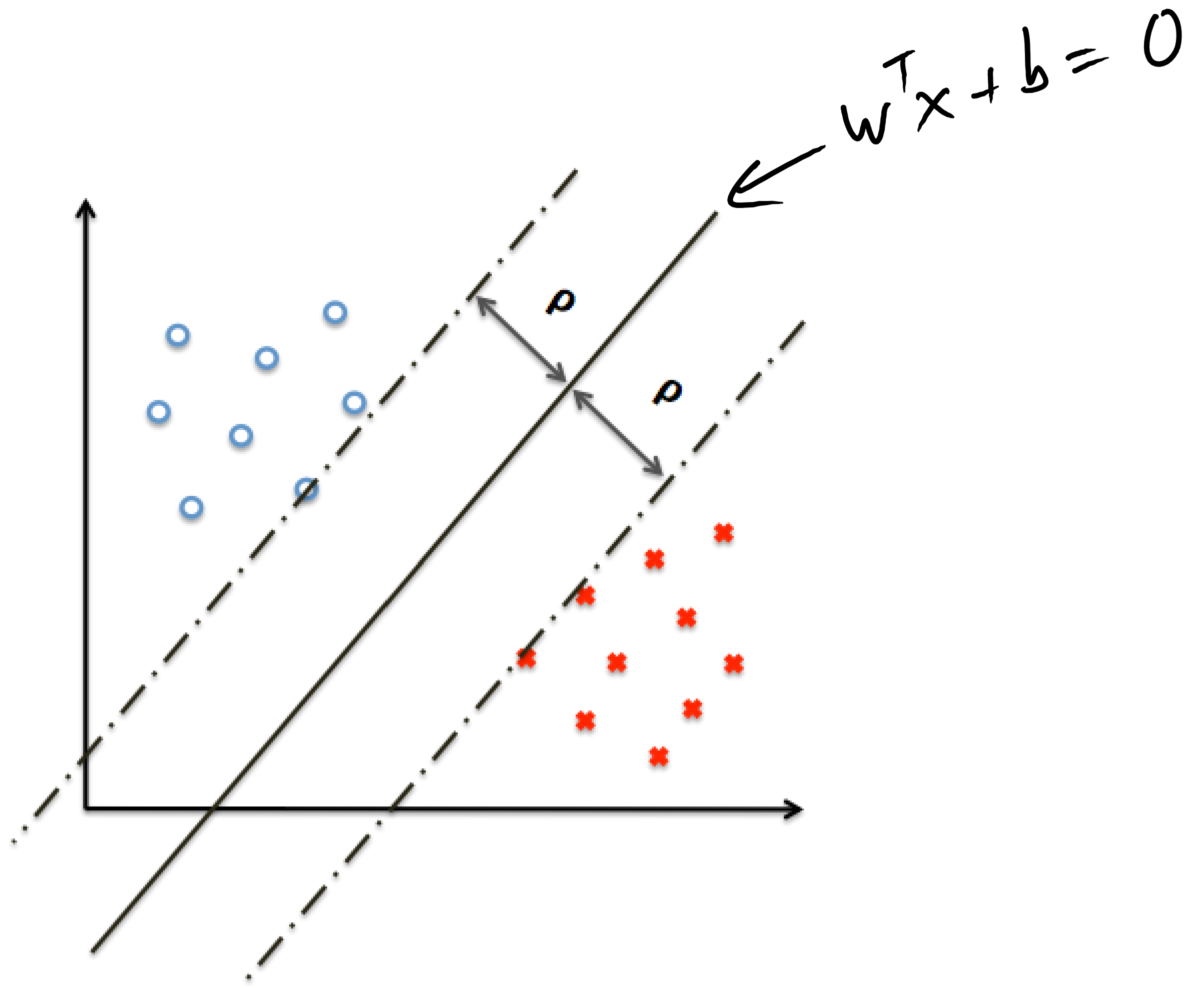
- The *maximum margin* or *optimal* separating hyperplane is the solution of

$$\max_{\boldsymbol{w},\mathrm{b}} \rho\left(\boldsymbol{w}, b\right) \qquad \text{s.t. } y_i\left(\boldsymbol{w}^T \boldsymbol{x}_i + b\right) > 0 \;\; \forall i$$

$$\Updownarrow$$

$$\max_{\boldsymbol{w},b} \left(\min_{i=1,\dots,n} \frac{\left|\boldsymbol{w}^T \boldsymbol{x}_i + b\right|}{||\boldsymbol{w}||} \right) \qquad \text{s.t. } y_i\left(\boldsymbol{w}^T \boldsymbol{x}_i + b\right) > 0 \;\; \forall i$$

Max-Margin Hyperplane



Canonical Form

- A separating hyperplane is said to be in *canonical form* if w and b are such that

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 \qquad \forall i$$
$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) = 1 \quad \text{for some } i$$

- Every separating hyperplane can be expressed in canonical form. Suppose $\mathcal{H} = \{\boldsymbol{x} : \boldsymbol{w}_1^T\boldsymbol{x} + b_1 = 0\}$ is a separating hyperplane (not necessarily in canonical form). Let

$$m := \min_{i=1,\ldots,n} |\boldsymbol{w}_1^T\boldsymbol{x}_i + b_1|$$

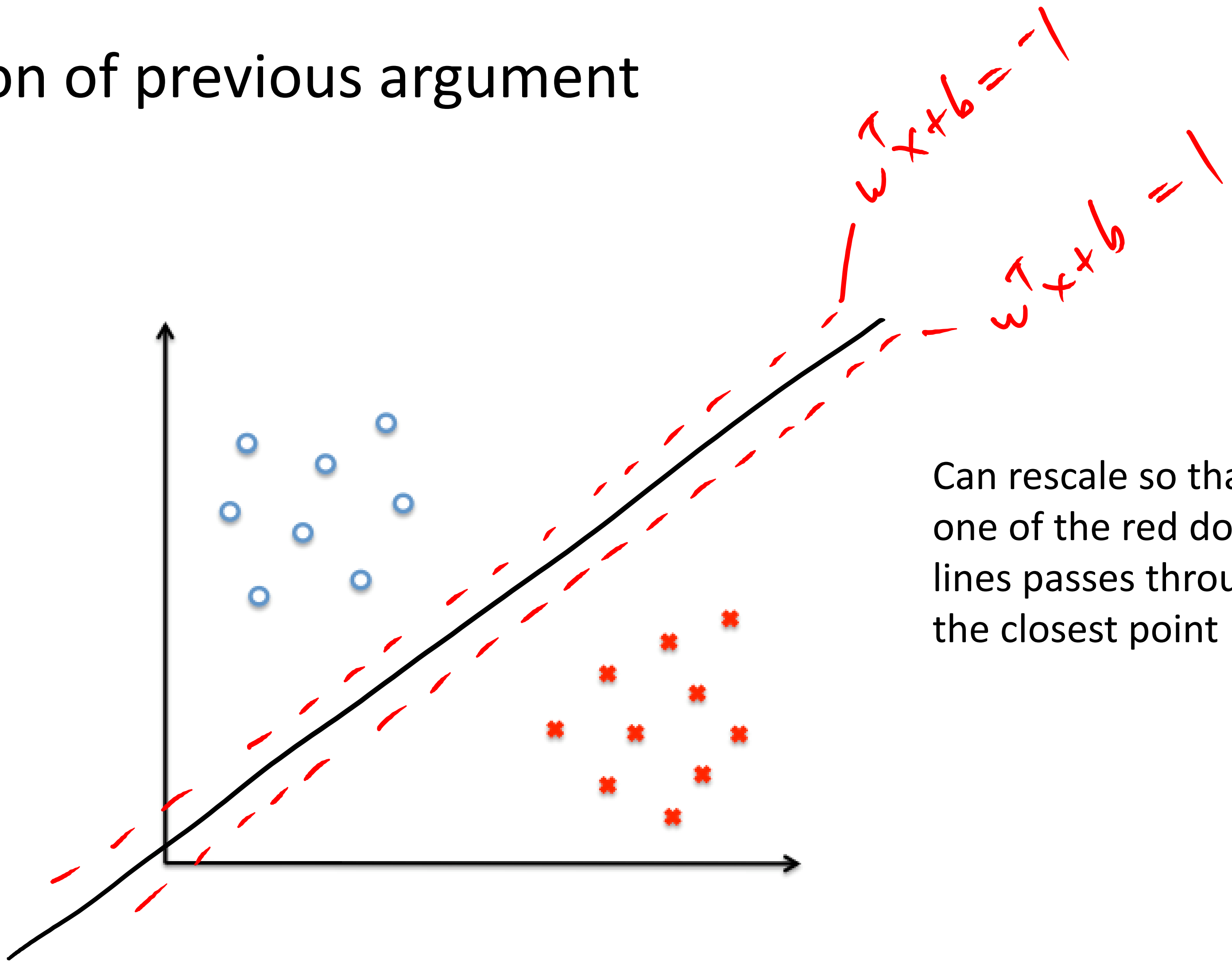
and define

$$\boldsymbol{w}_2 = \frac{\boldsymbol{w}_1}{m}, \quad b_2 = \frac{b_1}{m}.$$

then $\boldsymbol{w}_2$, b_2 express $\mathcal{H}$ in canonical form.

Canonical Form

- Illustration of previous argument



Can rescale so that one of the red dotted lines passes through the closest point

Max-Margin Hyperplane

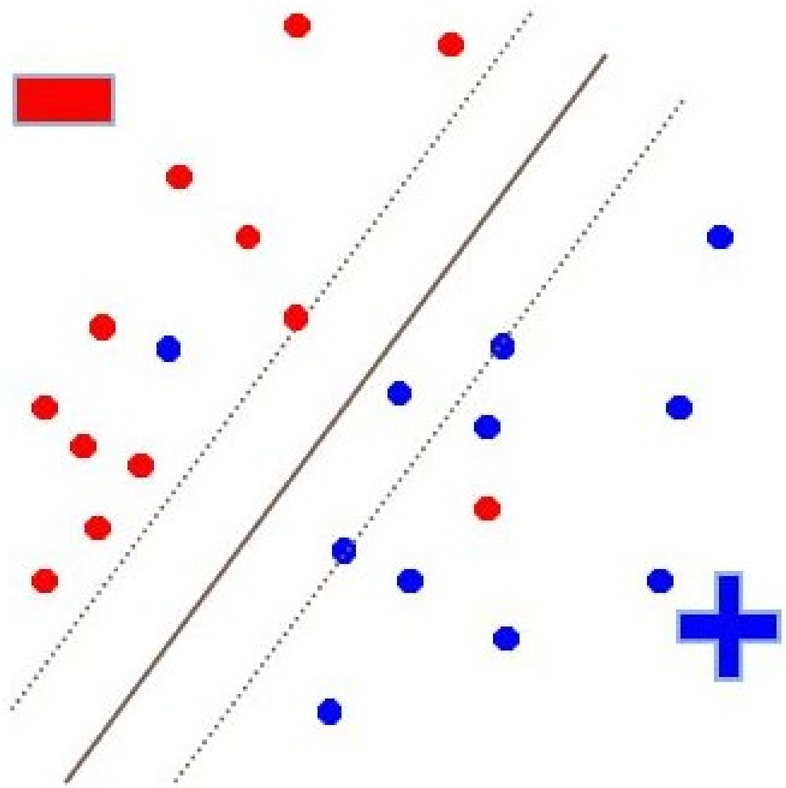
- This allows us to write the max-margin hyperplane as

$$
\begin{aligned}
\max_{\boldsymbol{w},b} \quad & \min_{i=1,\ldots,n} \frac{\left|\boldsymbol{w}^T\boldsymbol{x}_i + b\right|}{\|\boldsymbol{w}\|} \\
\text{s.t.} \quad & y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right) \geq 1 \qquad \forall i \\
& y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right) = 1 \quad \text{for some } i
\end{aligned}
$$

  - Previously, we had $y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right) > 0 \ \forall i$

Non-Separable Data

- What if the training data are not linearly separable?



Optimal Soft-Margin Hyperplane

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

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

$$\text{s.t. } y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right) \geq 1 - \xi_i \quad \forall i$$

$$\xi_i \geq 0 \qquad \forall i$$

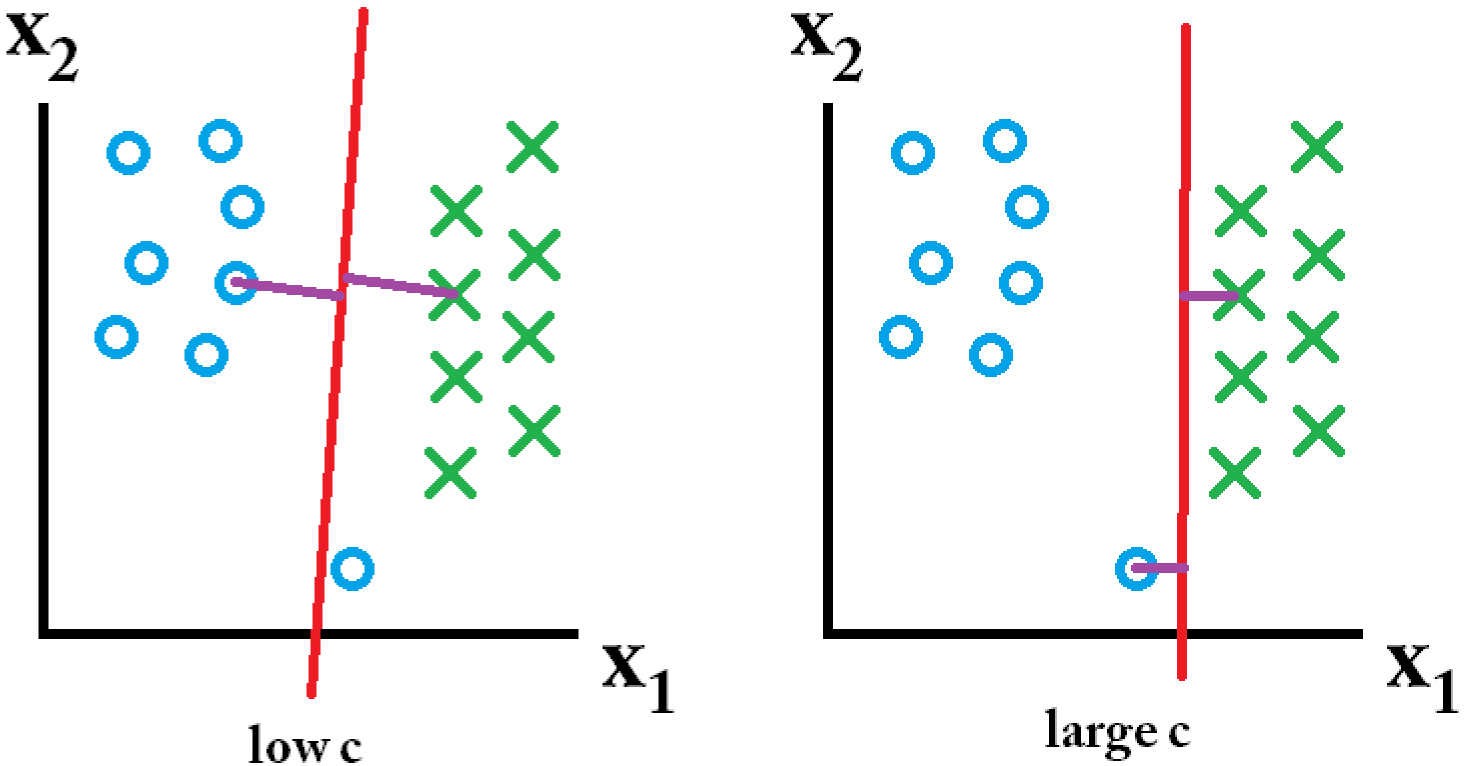
- C is a user-defined parameter
- OSM hyperplane is a special case of the *support vector machine*

Group Exercise

1. Argue that if $\boldsymbol{x}_i$ is misclassified by the OSM hyperplane, then $\xi_i \geq 1$.
2. Use the previous fact to show that the training error is bounded by
$$\frac{1}{n}\sum_{i=1}^{n}\xi_i.$$
3. What is the impact of the constant C in the optimal soft-margin hyperplane? Consider the case where outliers are present in the training data.

Effects of the (regularization) parameter C

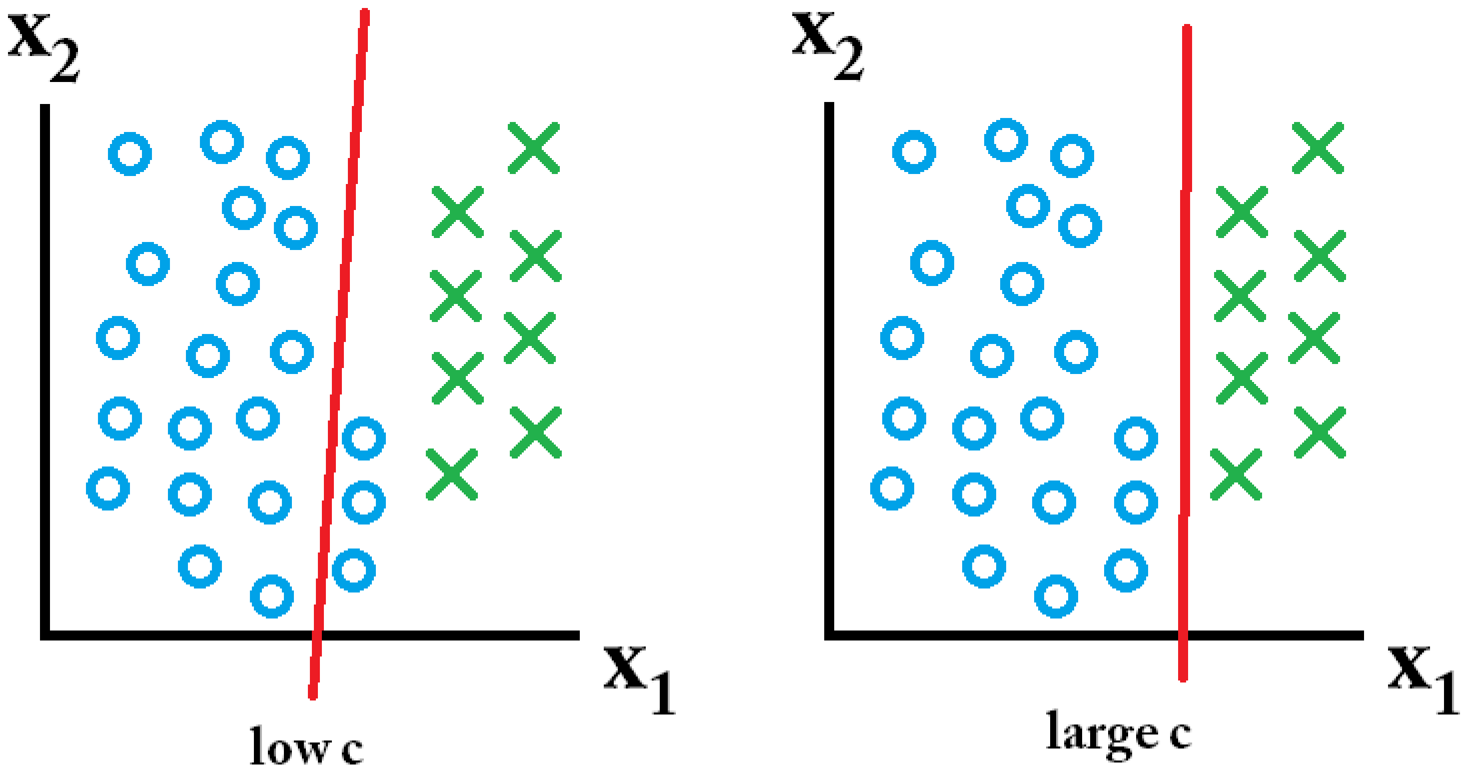
- Which is better in this case?



- It depends on future data

Effects of the (regularization) parameter C

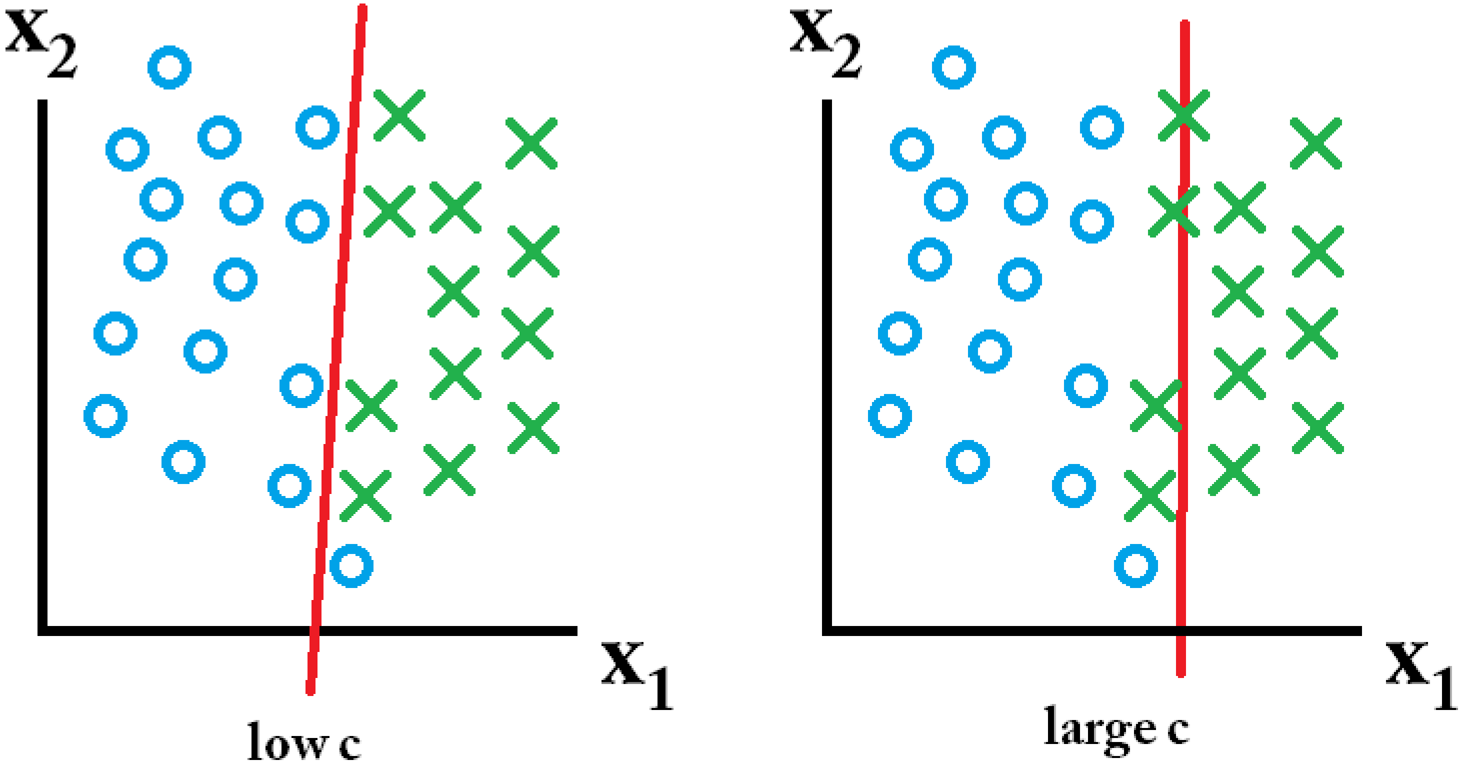
- One scenario



- Large C is best

Effects of the (regularization) parameter C

- Another scenario



- Small C is best

Effects of the (regularization) parameter C

- How can we choose C?
- No good theory for this
- Best practice right now is to use cross validation

ERM and OSM Hyperplane

- Recall the optimal soft margin hyperplane solves:

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

- If $\lambda = \frac{1}{C}$, then the solution $(\boldsymbol{w}^*, b^*)$ also solves

$$\min_{\boldsymbol{w},b}\left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n}\max\left(0,1 - y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right)\right)\right)$$

  - Proof on next slide
- Conclusion: the OSM hyperplane corresponds to regularized ERM with hinge loss

ERM and OSM Hyperplane

- The statement on the previous slide can be seen by scaling the objective function of (OSM) by $\frac{1}{C}$, which doesn't change the solution, and merging the constraints into a single constraint (for each i):

$$\left.\begin{aligned} y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) &\geq 1 - \xi_i \\ \xi_i &\geq 0 \end{aligned}\right\} \quad \Longleftrightarrow \quad \xi_i \geq \max(0, 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b))$$

So (OSM) reduces to

$$\begin{aligned} \min_{\boldsymbol{w}, b, \boldsymbol{\xi}} \quad & \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & \xi_i \geq \max\{0, 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)\} \end{aligned}$$

Clearly the solution must satisfy

$$\xi_i = \max\{0, 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)\} \quad \forall i$$

(otherwise we could decrease the objective), which reduces the problem to ERM with hinge loss.

Further reading

- ESL Sections 4.5.2 and 12.2
- ISL Sections 9.1 and 9.2