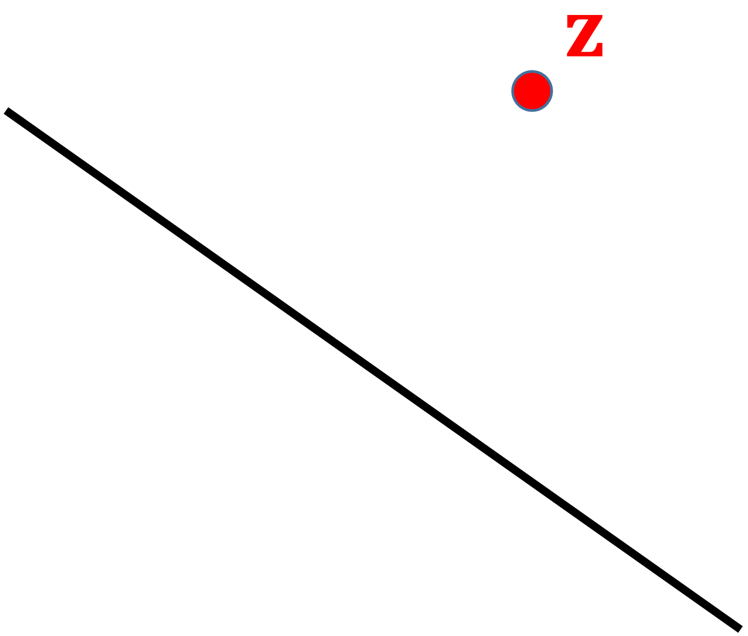
Support Vector Machine (SVM)

Shusen Wang

Project a Point onto a Hyperplane

Project a Point onto a Hyperplane

Question: how to project $\mathbf{z}$ onto the hyperplane?



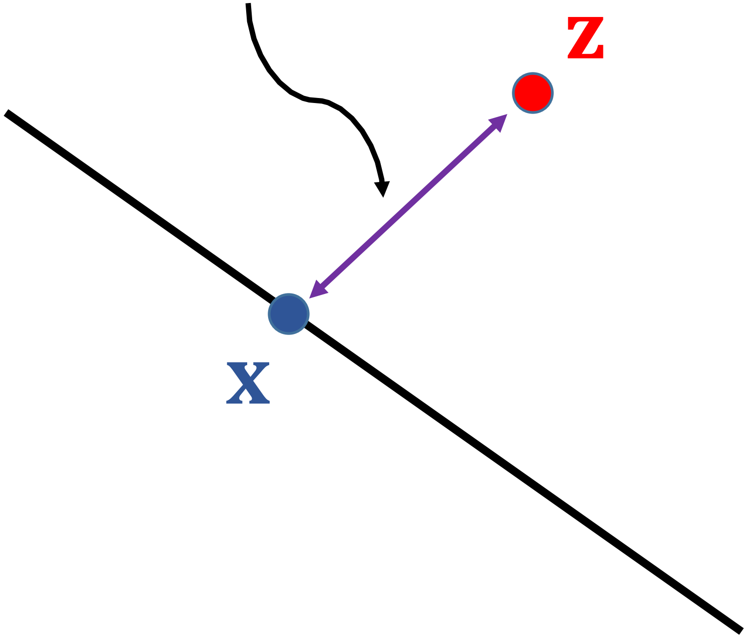
Hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$

Project a Point onto a Hyperplane

Question: how to project $\mathbf{z}$ onto the hyperplane?

Solution: find $\mathbf{x}$ on the hyperplane such that $\|\mathbf{z} - \mathbf{x}\|_2^2$ is minimized.

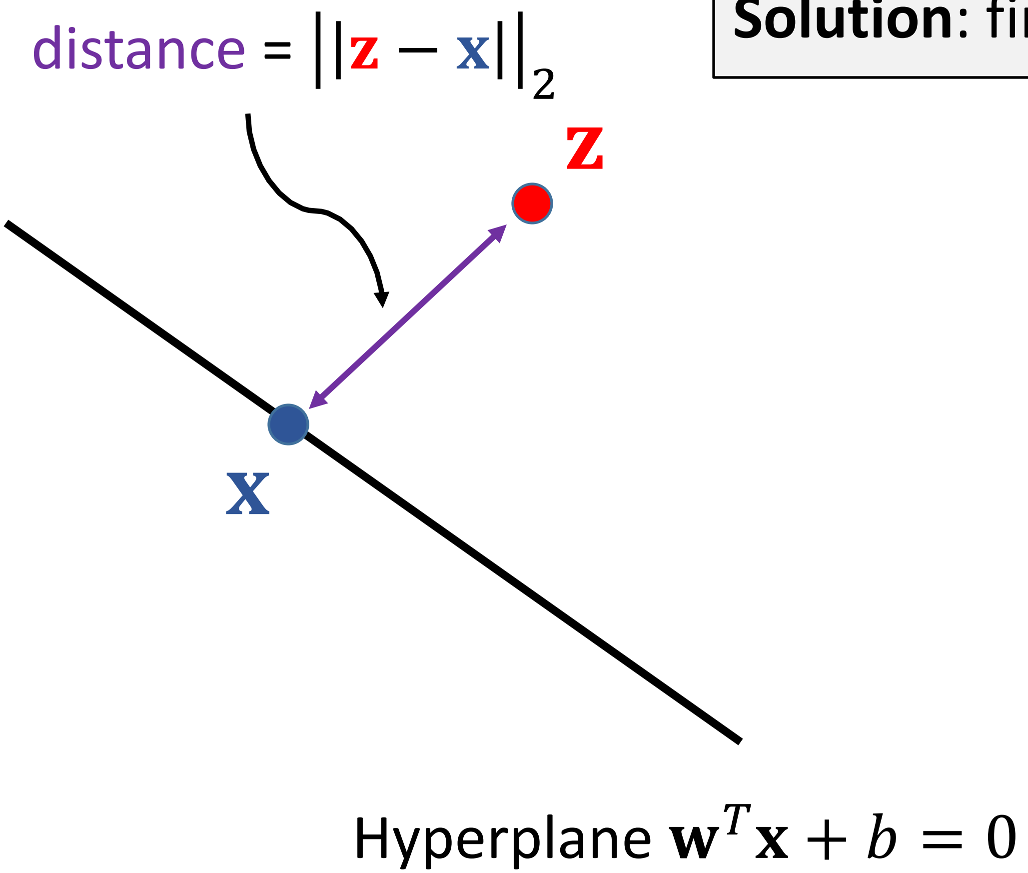
distance = $\|\mathbf{z} - \mathbf{x}\|_2$



Hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$

- $\min_{\mathbf{x}} \|\mathbf{z} - \mathbf{x}\|_2^2$;   s.t. $\mathbf{w}^T\mathbf{x} + b = 0$

Project a Point onto a Hyperplane



Question: how to project $\mathbf{z}$ onto the hyperplane?

Solution: find $\mathbf{x}$ on the hyperplane such that $||\mathbf{z} - \mathbf{x}||_2^2$ is minimized.

- $\min_{\mathbf{x}} ||\mathbf{z} - \mathbf{x}||_2^2$; s.t. $\mathbf{w}^T\mathbf{x} + b = 0$
- Solve the problem using the KKT conditions:

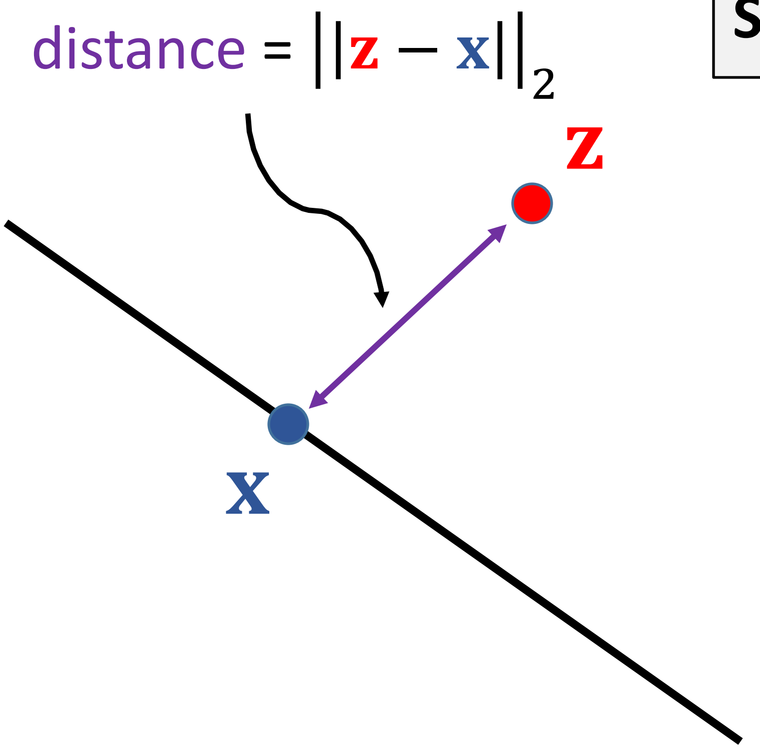
$$\begin{cases} \dfrac{\partial ||\mathbf{z} - \mathbf{x}||_2^2}{\partial \mathbf{x}} + \lambda \dfrac{\partial\, (\mathbf{w}^T\mathbf{x} + b)}{\partial \mathbf{x}} = 0; \\ \mathbf{w}^T\mathbf{x} + b = 0. \end{cases}$$

- Solution: $\mathbf{x} = \mathbf{z} - \frac{\mathbf{w}^T\mathbf{z} + b}{||\mathbf{w}||_2^2}\mathbf{w}$

Project a Point onto a Hyperplane

Question: how to project $\mathbf{z}$ onto the hyperplane?

Solution: find $\mathbf{x}$ on the hyperplane such that $||\mathbf{z}-\mathbf{x}||_2^2$ is minimized.



Hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$

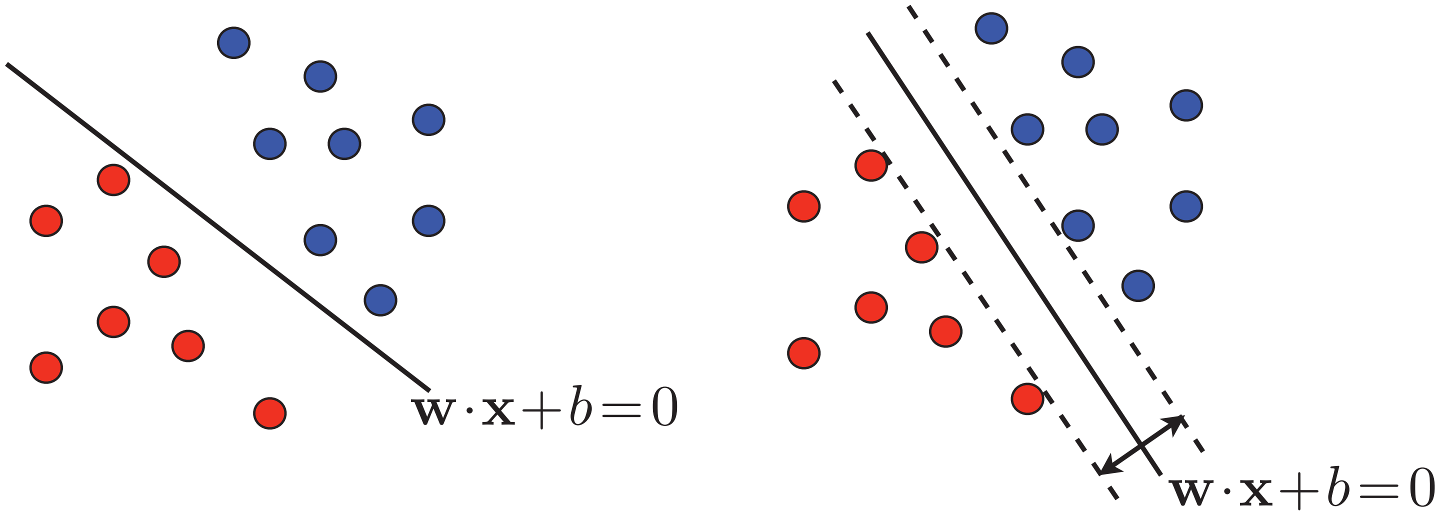
- Solution: $\mathbf{x} = \mathbf{z} - \frac{\mathbf{w}^T\mathbf{z}+b}{||\mathbf{w}||_2^2}\mathbf{w}$
- The ℓ_2 distance between $\mathbf{z}$ and the hyperplane is

$$\boxed{||\mathbf{z}-\mathbf{x}||_2 = \frac{|\mathbf{w}^T\mathbf{z}+b|}{||\mathbf{w}||_2}.}$$

Support Vector Machine (SVM)

Support Vector Machine (SVM)

Separate data by a hyperplane (assume the data are separable)



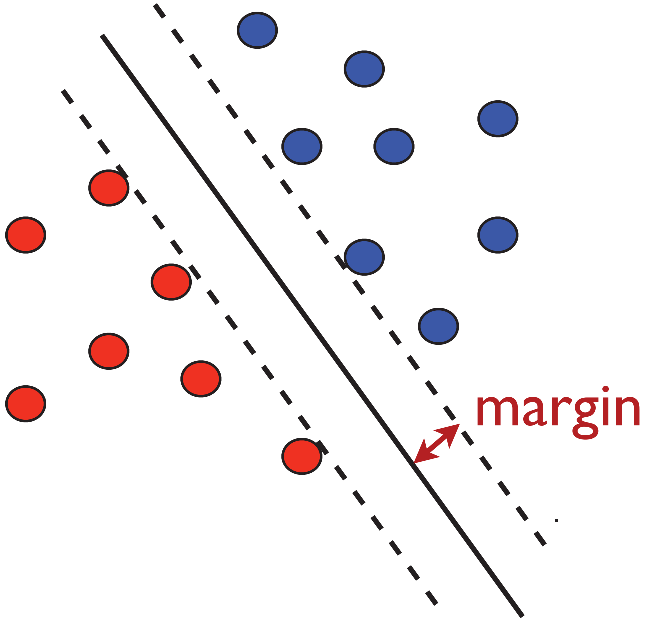
An arbitrary hyperplane.

The hyperplane that maximizes the margin.

The figure is from the book "*Foundations of Machine Learning*"

Support Vector Machine (SVM)

Separate data by a hyperplane (assume the data are separable)

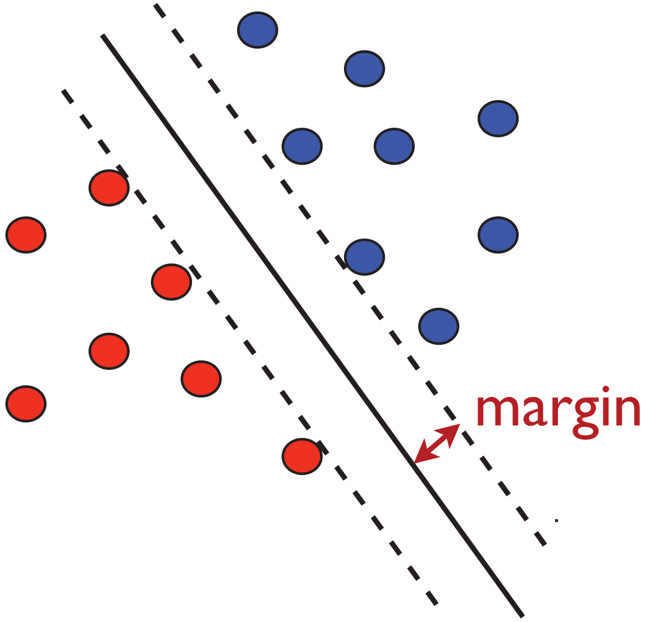


Hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$

The figure is from the book "*Foundations of Machine Learning*"

Support Vector Machine (SVM)

Separate data by a hyperplane (assume the data are separable)



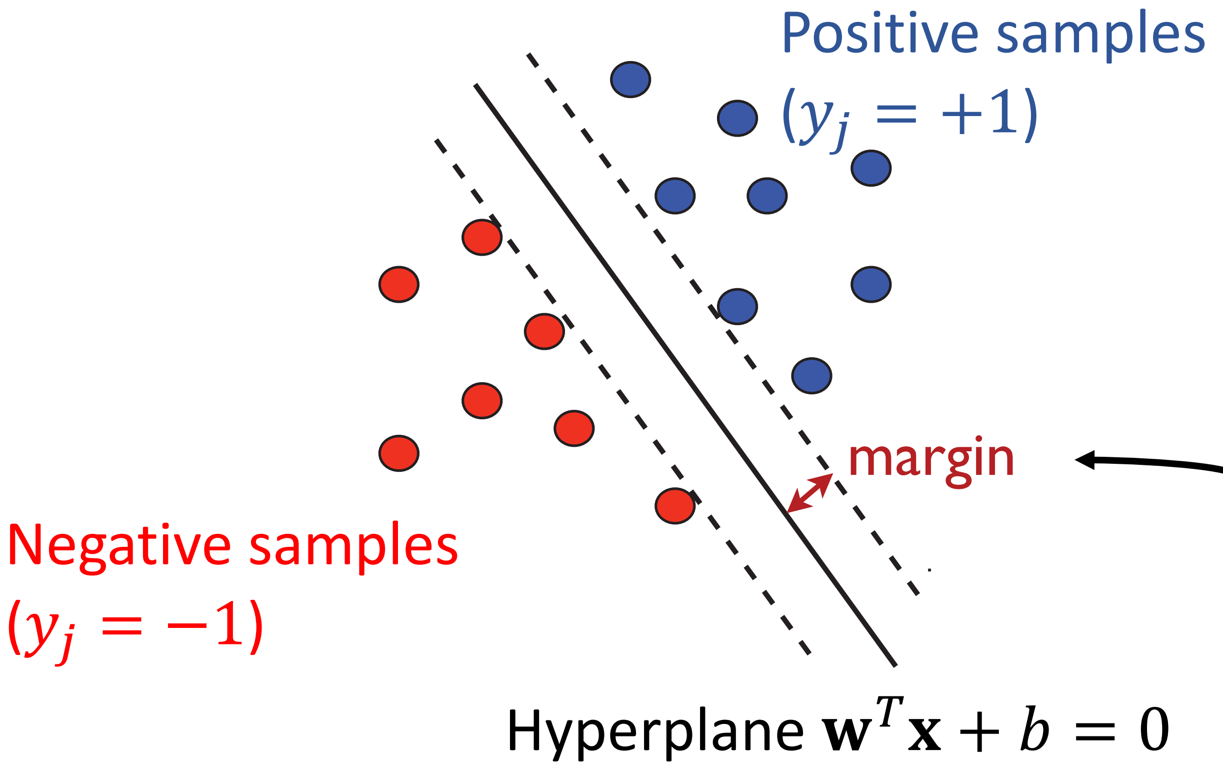
Hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$

- The distance between any feature vector, $\mathbf{x}$, and the hyperplane is $\frac{|\mathbf{w}^T\mathbf{x}+b|}{||\mathbf{w}||_2}$.
- The margin is $\min_j \frac{|\mathbf{w}^T\mathbf{x}_j+b|}{||\mathbf{w}||_2}$

The figure is from the book "*Foundations of Machine Learning*"

Support Vector Machine (SVM)

Separate data by a hyperplane (assume the data are separable)



- The distance between any feature vector, $\mathbf{x}$, and the hyperplane is $\frac{|\mathbf{w}^T\mathbf{x}+b|}{||\mathbf{w}||_2}$.
- The margin is
$$\min_j \frac{|\mathbf{w}^T\mathbf{x}_j+b|}{||\mathbf{w}||_2} = \min_j \frac{y_j(\mathbf{w}^T\mathbf{x}_j+b)}{||\mathbf{w}||_2}$$

The figure is from the book "*Foundations of Machine Learning*"

Support Vector Machine (SVM)

Margin = $\min_j \frac{y_j(\mathbf{w}^T\mathbf{x}_j+b)}{||\mathbf{w}||_2}$; we want to maximize the margin.

Support Vector Machine (SVM)

Margin = $\min_j \frac{y_j\left(\boxed{\mathbf{w}^T\mathbf{x}_j+b}\right)}{||\mathbf{w}||_2}$; we want to maximize the margin.

Define $\bar{\mathbf{x}}_j = [\mathbf{x}_j; 1] \in \mathbb{R}^{d+1}$

Define $\bar{\mathbf{w}} = [\mathbf{w}, b] \in \mathbb{R}^{d+1}$

➔ $\boxed{\mathbf{x}_j^T\mathbf{w} + b = \bar{\mathbf{x}}_j^T\bar{\mathbf{w}}}$

Support Vector Machine (SVM)

Margin $= \min_{j} \frac{y_j \mathbf{w}^T \mathbf{x}_j}{||\mathbf{w}||_2}$; we want to maximize the margin.

$\Downarrow$

Support Vector Machine (SVM): $\max_{\mathbf{w}} \min_{j} \frac{y_j \mathbf{w}^T \mathbf{x}_j}{||\mathbf{w}||_2}$

Support Vector Machine (SVM)

Support Vector Machine (SVM): $$\max_{\mathbf{w}} \min_{j} \frac{y_j \mathbf{w}^T \mathbf{x}_j}{||\mathbf{w}||_2}$$

Support Vector Machine (SVM)

Support Vector Machine (SVM): $\max_{\mathbf{w}} \min_{j} \frac{y_j \mathbf{w}^T \mathbf{x}_j}{||\mathbf{w}||_2}$

$$\operatorname*{argmax}_{\mathbf{w}} \min_{j} \frac{y_j \mathbf{w}^T \mathbf{x}_j}{||\mathbf{w}||_2} = \operatorname*{argmax}_{\mathbf{w}} \frac{\min_{j} y_j \mathbf{w}^T \mathbf{x}_j}{||\mathbf{w}||_2}$$

$$= \operatorname*{argmax}_{\mathbf{w}} \frac{1}{||\mathbf{w}||_2}, \quad \text{s.t.} \quad \min_{j} \; y_j \mathbf{w}^T \mathbf{x}_j = 1$$

$$= \operatorname*{argmin}_{\mathbf{w}} ||\mathbf{w}||_2^2, \quad \text{s.t.} \quad \min_{j} \; y_j \mathbf{w}^T \mathbf{x}_j = 1$$

$$= \operatorname*{argmin}_{\mathbf{w}} ||\mathbf{w}||_2^2, \quad \text{s.t.} \quad y_j \mathbf{w}^T \mathbf{x}_j \geq 1 \text{ for all } j$$

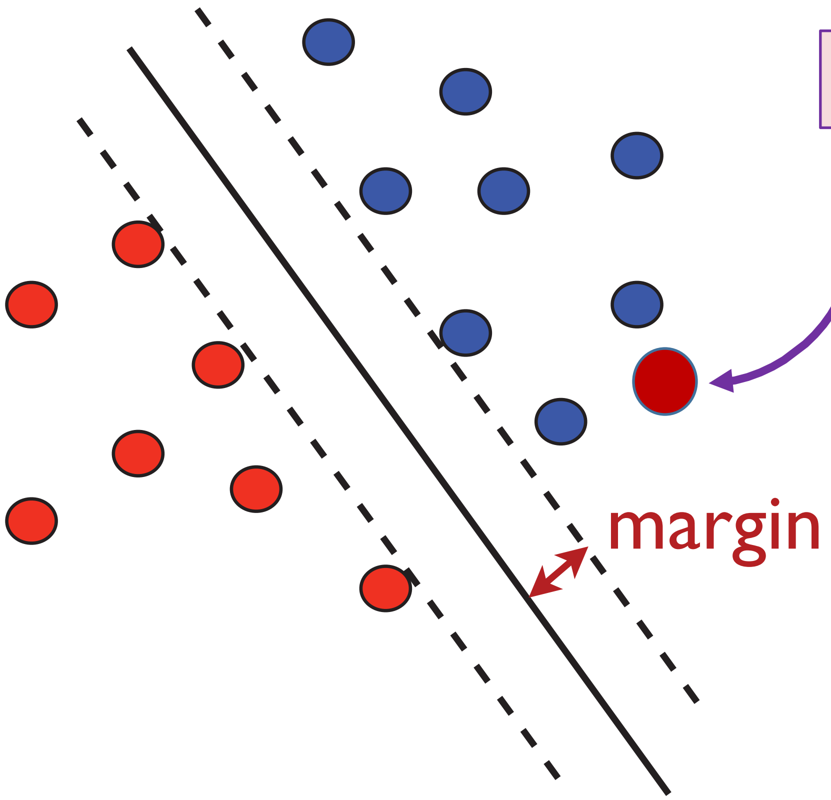
Support Vector Machine (SVM)

$$\min_{\mathbf{w}} ||\mathbf{w}||_2^2, \quad \text{s.t.} \quad 1 - y_j \mathbf{w}^T \mathbf{x}_j \leq 0 \ \text{ for all } j \in \{1, \cdots, n\}.$$

Equivalent form of SVM

Support Vector Machine (SVM)

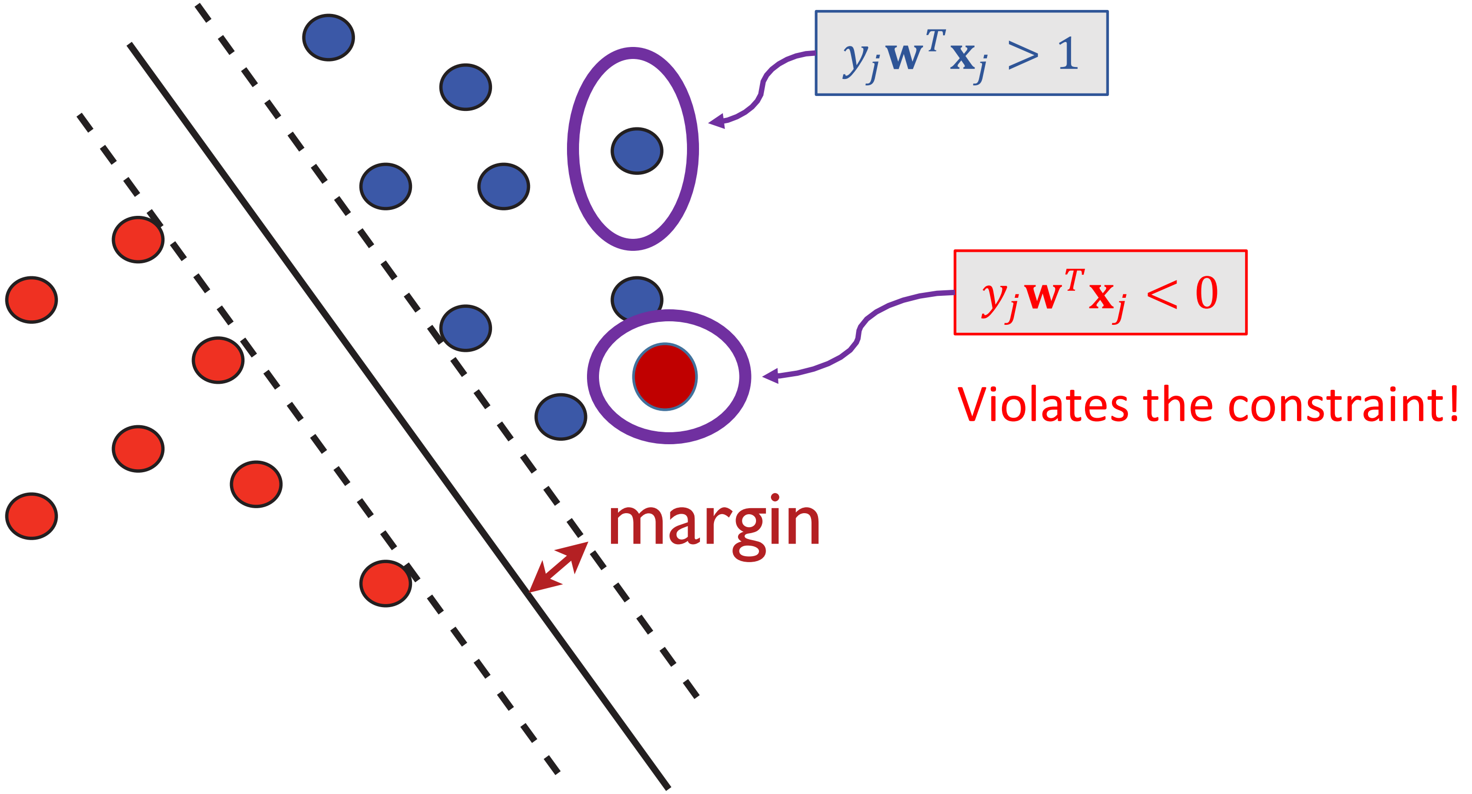
$$\min_{\mathbf{w}} \|\mathbf{w}\|_2^2, \quad \text{s.t.} \quad 1 - y_j \mathbf{w}^T \mathbf{x}_j \leq 0 \ \text{ for all } j \in \{1, \cdots, n\}.$$



What if the data is inseparable?

Support Vector Machine (SVM)

$$\min_{\mathbf{w}} \|\mathbf{w}\|_2^2, \quad \text{s.t.} \quad 1 - y_j \mathbf{w}^T \mathbf{x}_j \le 0 \ \text{ for all } j \in \{1, \cdots, n\}.$$



Support Vector Machine (SVM)

$$\min_{\mathbf{w}} \|\mathbf{w}\|_2^2, \quad \text{s.t.} \quad 1 - y_j \mathbf{w}^T \mathbf{x}_j \leq 0 \;\text{ for all } j \in \{1, \cdots, n\}.$$

$$\Big\Downarrow \text{ Relax}$$

$$\min_{\mathbf{w}, \xi_j} \|\mathbf{w}\|_2^2 + \lambda \sum_j [\xi_j]_+, \quad \text{s.t.} \quad 1 - y_j \mathbf{w}^T \mathbf{x}_j = \xi_j \;\text{ for all } j \in \{1, \cdots, n\}.$$

- $[\xi_j]_+ = \max\{\xi_j, 0\}$

Support Vector Machine (SVM)

$$\min_{\mathbf{w}} \|\mathbf{w}\|_2^2, \quad \text{s.t.} \quad 1 - y_j \mathbf{w}^T \mathbf{x}_j \leq 0 \text{ for all } j \in \{1, \cdots, n\}.$$

Relax

$$\min_{\mathbf{w}, \xi_j} \|\mathbf{w}\|_2^2 + \lambda \sum_j [\xi_j]_+, \quad \text{s.t.} \quad 1 - y_j \mathbf{w}^T \mathbf{x}_j = \xi_j \text{ for all } j \in \{1, \cdots, n\}.$$

- $[\xi_j]_+ = \max\{\xi_j, 0\}$
- $\xi_j \leq 0$ means the constraint $1 - y_j \mathbf{w}^T \mathbf{x}_j \leq 0$ is satisfied
  ➔ no penalty!
- $\xi_j > 0$ means the constraint is violated (because the data is inseparable)
  ➔ penalize the violation ξ_j.

Support Vector Machine (SVM)

$$\min_{\mathbf{w}} \|\mathbf{w}\|_2^2, \quad \text{s.t.} \quad 1 - y_j \mathbf{w}^T \mathbf{x}_j \leq 0 \ \text{ for all } j \in \{1, \cdots, n\}.$$

Relax

$$\min_{\mathbf{w}, \xi_j} \|\mathbf{w}\|_2^2 + \lambda \sum_j [\xi_j]_+, \quad \text{s.t.} \quad 1 - y_j \mathbf{w}^T \mathbf{x}_j = \xi_j \ \text{ for all } j \in \{1, \cdots, n\}.$$

Equivalent

$$\min_{\mathbf{w}, b} \|\mathbf{w}\|_2^2 + \lambda \sum_j [1 - y_j \mathbf{w}^T \mathbf{x}_j]_+.$$

Comparisons

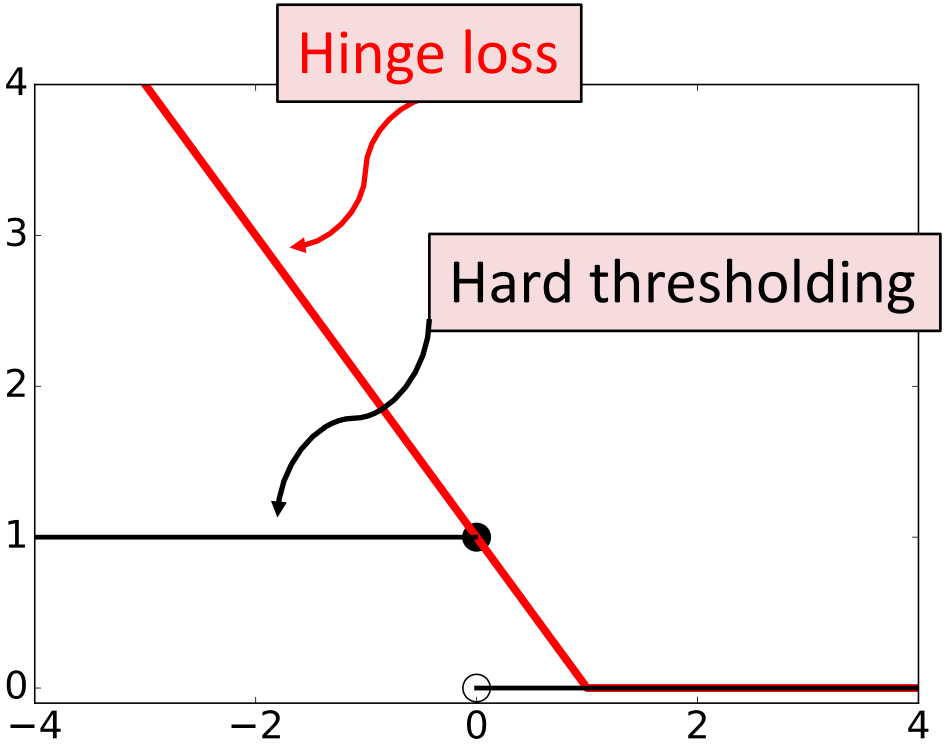
SVM: $\min_{\mathbf{w}} ||\mathbf{w}||_2^2 + \lambda \sum_j g(y_j \mathbf{w}^T \mathbf{x}_j)$.

Hinge loss: $g(z) = [1 - z]_+$.

Comparisons

SVM: $\min_{\mathbf{w}} ||\mathbf{w}||_2^2 + \lambda \sum_j g(y_j \mathbf{w}^T \mathbf{x}_j)$.

Hinge loss: $g(z) = [1 - z]_+$.

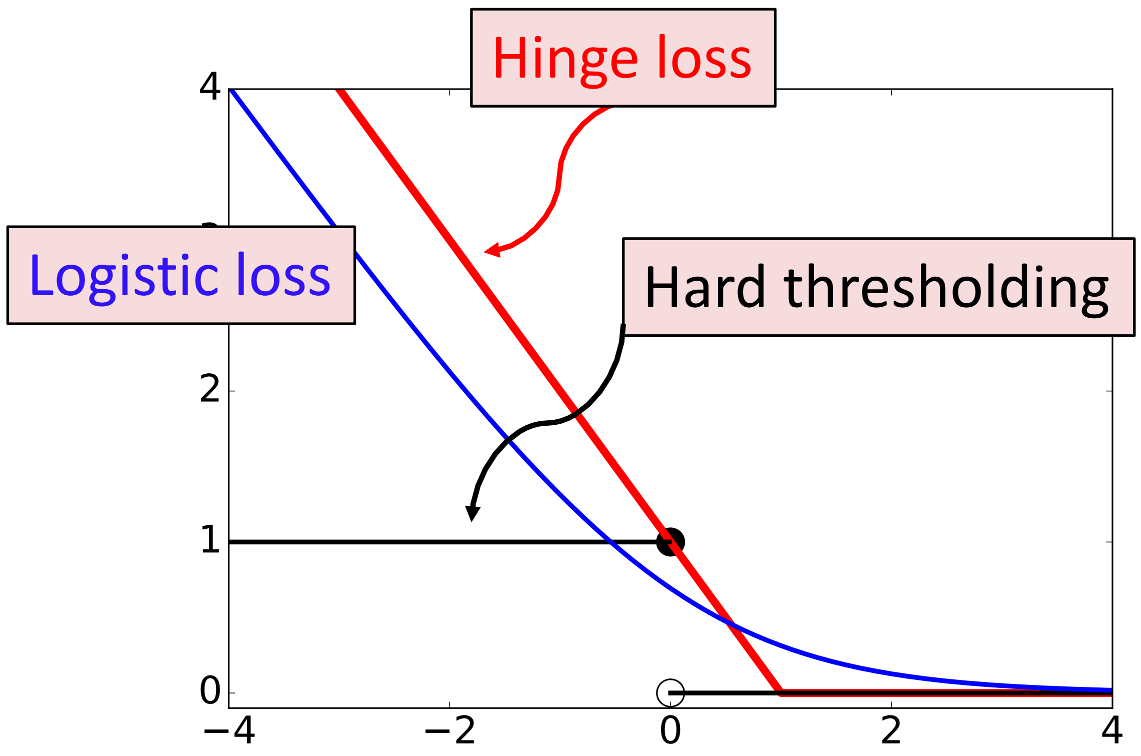


Hard thresholding: $h(\mathrm{z}) = \begin{cases} 1, & \text{if } z < 0; \\ 0, & \text{if } z \geq 0. \end{cases}$

Comparisons

SVM: $\min_{\mathbf{w}} ||\mathbf{w}||_2^2 + \lambda \sum_j g\left(y_j \mathbf{w}^T \mathbf{x}_j\right)$.

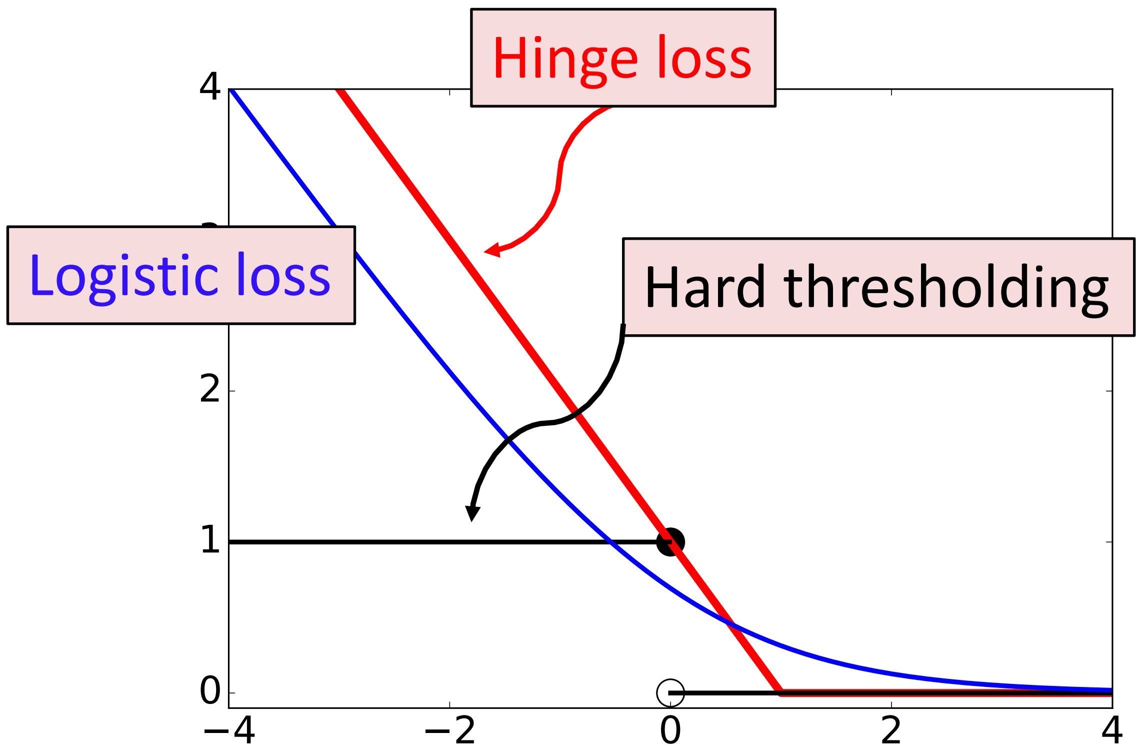
Hinge loss: $g(z) = [1 - z]_+$.



Hard thresholding: $h(\mathrm{z}) = \begin{cases} 1, & \text{if } z < 0; \\ 0, & \text{if } z \geq 0. \end{cases}$

Logistic loss: $l(\mathrm{z}) = \log(1 + e^{-z})$.

Comparisons



- Convexity
  - Hinge loss and logistic loss are convex.
  - Global optima can be efficiently found.
- Smoothness
  - Hinge loss is non-smooth.
  - Logistic loss is smooth.
- Logistic regression is easier to solve than SVM.
  - GD for logistic regression has linear convergence.
  - Algorithms for SVM have sub-linear convergence.