

Support Vector Machine (SVM)

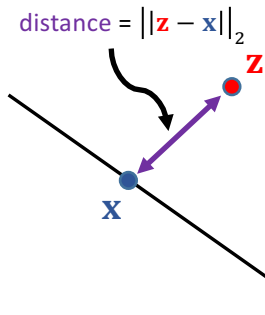
Shusen Wang

Project a Point onto a Hyperplane

Project a Point onto a Hyperplane

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

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



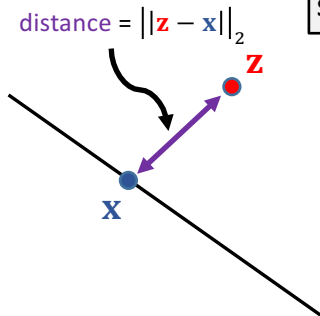
$$\bullet \min_{\mathbf{x}} \|\mathbf{z} - \mathbf{x}\|_2^2; \quad \text{s.t. } \mathbf{w}^T \mathbf{x} + b = 0$$

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

Project a Point onto a Hyperplane

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

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



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

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

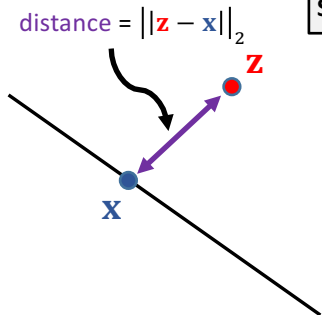
$$\begin{cases} \frac{\partial \|\mathbf{z} - \mathbf{x}\|_2^2}{\partial \mathbf{x}} + \lambda \frac{\partial (\mathbf{w}^T \mathbf{x} + b)}{\partial \mathbf{x}} = 0; \\ \mathbf{w}^T \mathbf{x} + b = 0. \end{cases}$$

- 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} to the hyperplane?

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



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

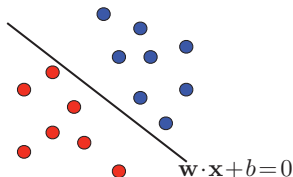
- Solution: $\mathbf{x} = \mathbf{z} - \frac{\mathbf{w}^T \mathbf{z} + b}{\|\mathbf{w}\|_2^2} \mathbf{w}$
- The ℓ_2 distance between \mathbf{z} and the hyperplane is

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

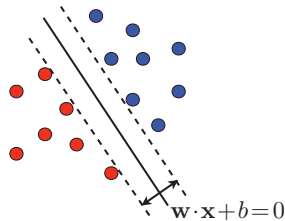
Support Vector Machine

Support Vector Machine (SVM)

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



An arbitrary hyperplane.

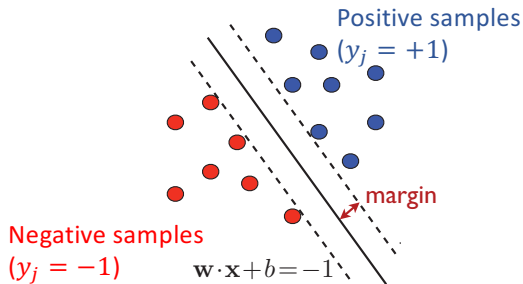


The hyperplane that maximizes the margin.

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

Support Vector Machine (SVM)

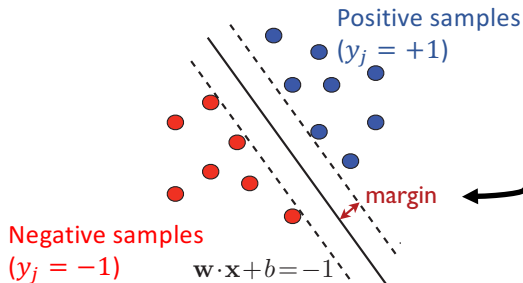
Separate data by a hyperplane (assume the data are separable)



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 \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): $\max_{\mathbf{w}, b} \min_j \frac{y_j(\mathbf{w}^T \mathbf{x}_j + b)}{\|\mathbf{w}\|_2}$

Support Vector Machine (SVM)

Support Vector Machine (SVM): $\max_{\mathbf{w}, b} \min_j \frac{y_j(\mathbf{w}^T \mathbf{x}_j + b)}{\|\mathbf{w}\|_2}$

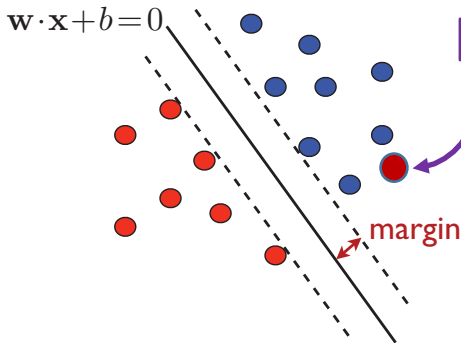
Support Vector Machine (SVM)

$$\text{Support Vector Machine (SVM): } \max_{\mathbf{w}, b} \min_j \frac{y_j(\mathbf{w}^T \mathbf{x}_j + b)}{\|\mathbf{w}\|_2}$$

$$\begin{aligned} \max_{\mathbf{w}, b} \min_j \frac{y_j(\mathbf{w}^T \mathbf{x}_j + b)}{\|\mathbf{w}\|_2} &= \max_{\mathbf{w}, b} \frac{\min_j y_j(\mathbf{w}^T \mathbf{x}_j + b)}{\|\mathbf{w}\|_2} \\ &= \max_{\mathbf{w}, b} \frac{1}{\|\mathbf{w}\|_2}, \quad \text{s.t.} \quad \min_j y_j(\mathbf{w}^T \mathbf{x}_j + b) = 1 \\ &= \min_{\mathbf{w}, b} \|\mathbf{w}\|_2, \quad \text{s.t.} \quad \min_j y_j(\mathbf{w}^T \mathbf{x}_j + b) = 1 \\ &= \min_{\mathbf{w}, b} \|\mathbf{w}\|_2, \quad \text{s.t.} \quad y_j(\mathbf{w}^T \mathbf{x}_j + b) \geq 1 \text{ for all } j \end{aligned}$$

Support Vector Machine (SVM)

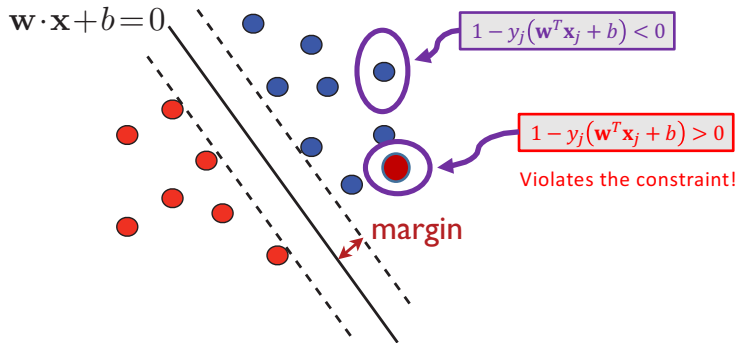
$$\min_{\mathbf{w}, b} \|\mathbf{w}\|_2, \quad \text{s.t.} \quad 1 - y_j(\mathbf{w}^T \mathbf{x}_j + b) \leq 0 \text{ for all } j \in \{1, \dots, n\}.$$



What if the data is inseparable?

Support Vector Machine (SVM)

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



Support Vector Machine (SVM)

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



Relax

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

- $[\xi_j]_+ = \max\{\xi_j, 0\}$
- $\xi_j \leq 0$ means the constraint $1 - y_j(\mathbf{w}^T \mathbf{x}_j + b) \leq 0$ 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)

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



Relax

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



Equivalent

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

Support Vector Machine (SVM)

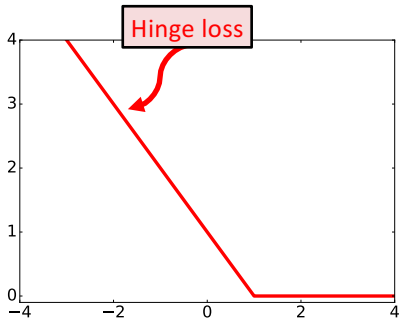
$$\min_{\mathbf{w}, b} ||\mathbf{w}||_2 + \lambda \sum_j [1 - y_j(\mathbf{w}^T \mathbf{x}_j + b)]_+.$$

Support Vector Machine (SVM)

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

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

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



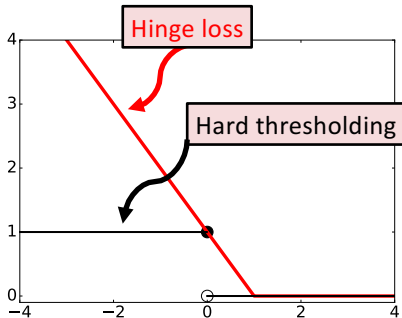
Comparisons

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

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

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

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



Comparisons

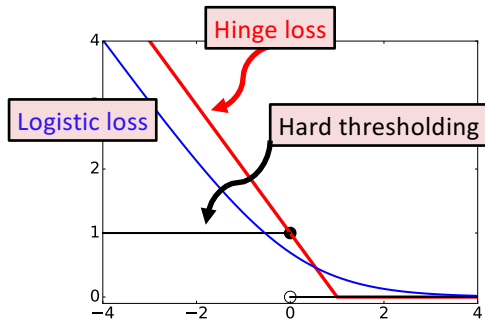
$$\min_{\mathbf{w}, b} \|\mathbf{w}\|_2 + \lambda \sum_j [1 - y_j(\mathbf{w}^T \mathbf{x}_j + b)]_+.$$

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

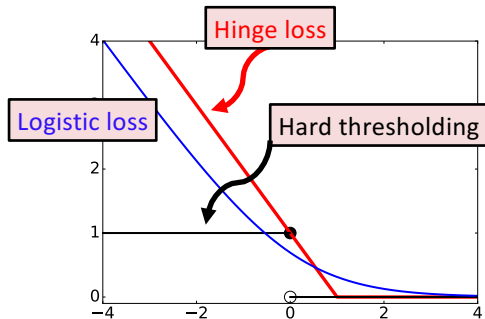
$$\text{Hinge loss: } g(z) = [1 - z]_+.$$

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

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



Comparisons



- Convexity
 - Hinge loss and logistic loss are convex.
 - Global optimal 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.