统计机器学习

（小班研讨）

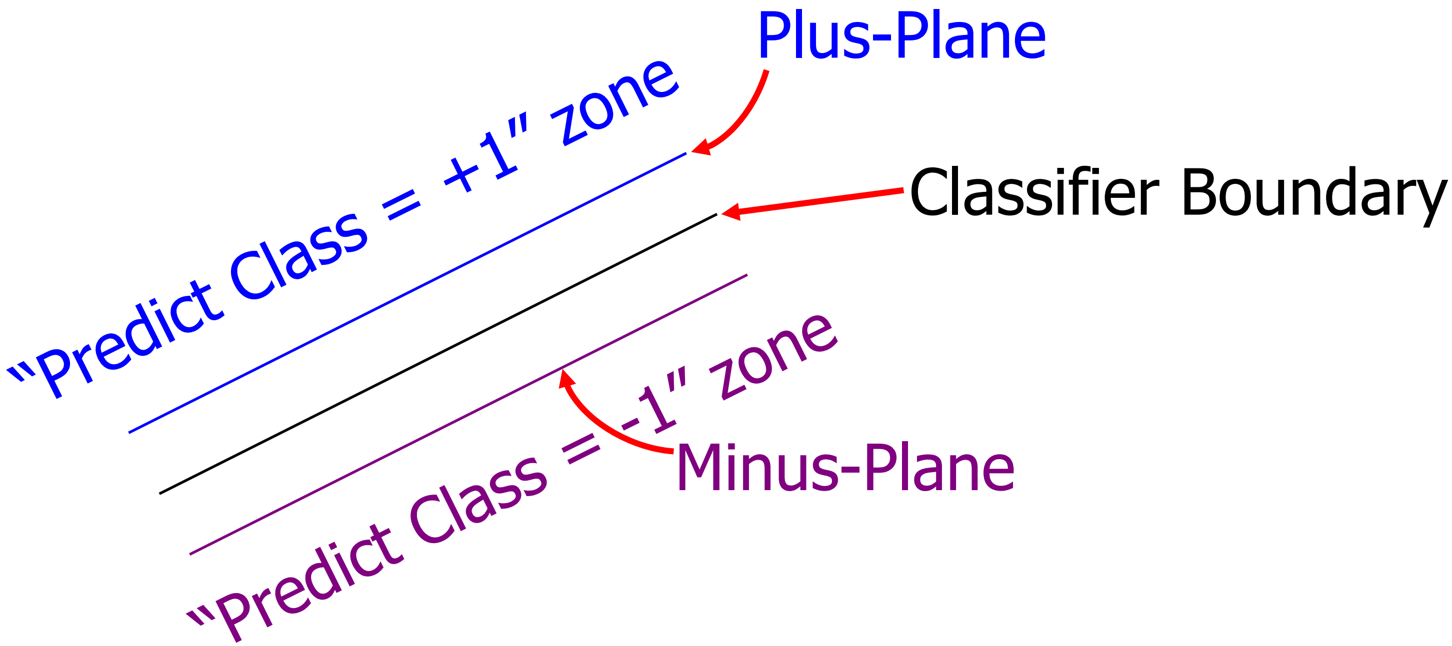
主讲教师： 刘峤

第4章 支持向量机与核方法

Support Vector Machines and Kernel Methods

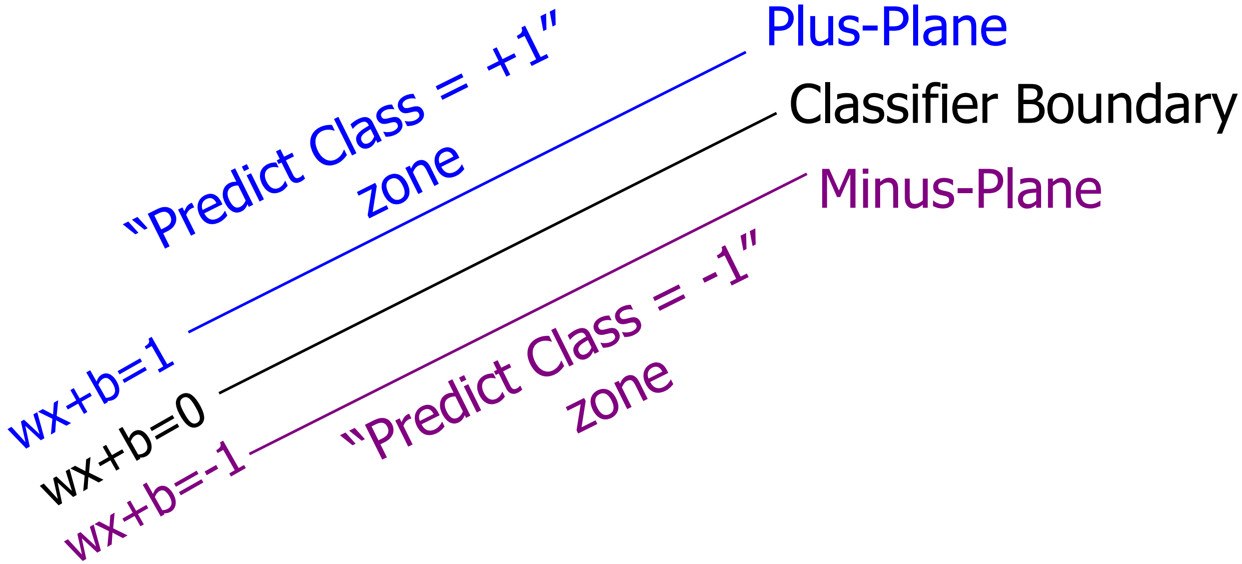
线性可分支持向量机

Specifying a line and margin



- How do we represent this mathematically?
- …in *m* input dimensions?

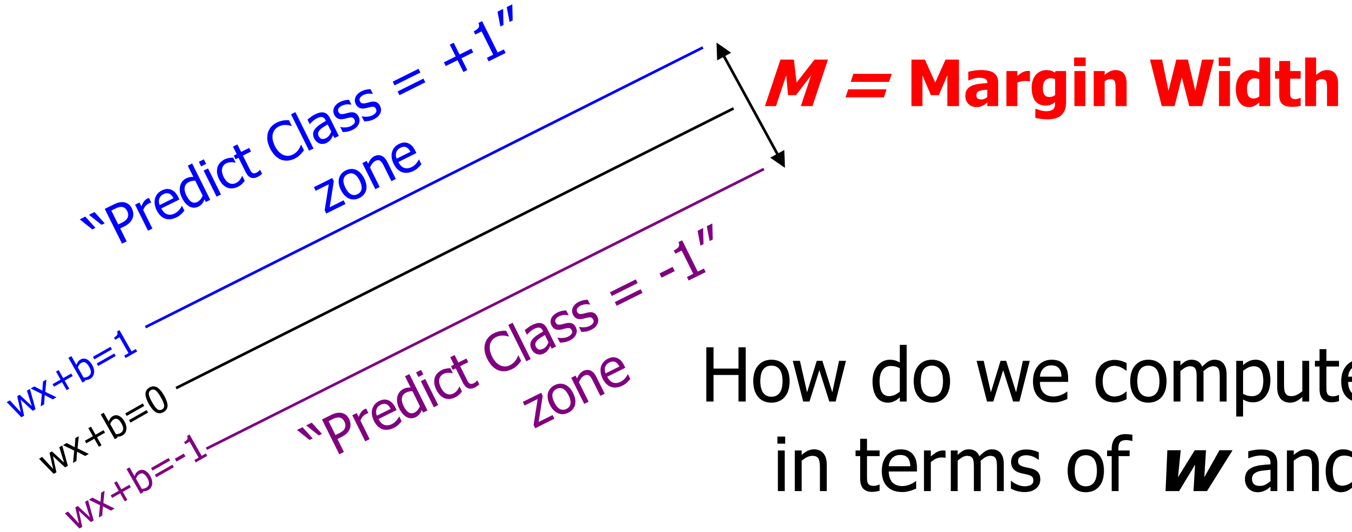
Specifying a line and margin



- Plus-plane = $\{ \mathbf{x} : \mathbf{w} \,.\, \mathbf{x} + b = +1 \}$
- Minus-plane = $\{ \mathbf{x} : \mathbf{w} \,.\, \mathbf{x} + b = -1 \}$

Classify as..	+1	if	$\mathbf{w} \,.\, \mathbf{x} + b >= 1$
	-1	if	$\mathbf{w} \,.\, \mathbf{x} + b <= -1$
	Universe explodes	if	$-1 < \mathbf{w} \,.\, \mathbf{x} + b < 1$

Computing the margin width



How do we compute M in terms of $\boldsymbol{w}$ and b?

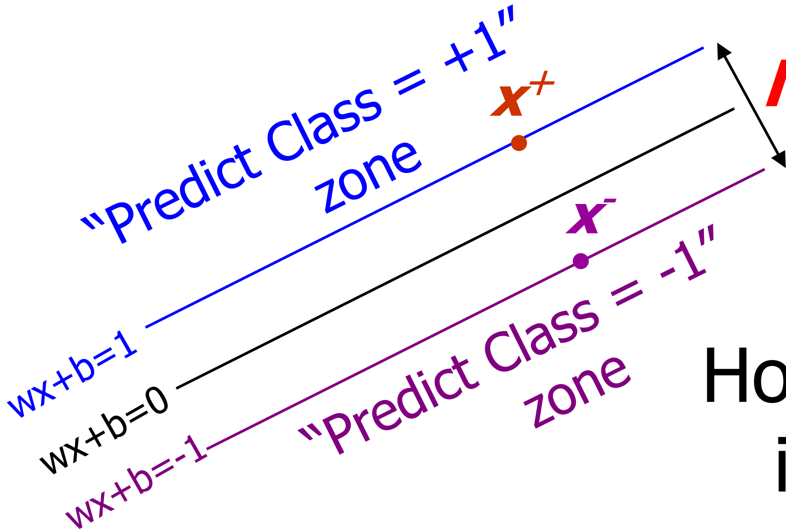
- Plus-plane = $\{ \boldsymbol{x} : \boldsymbol{w} \cdot \boldsymbol{x} + b = +1 \}$
- Minus-plane = $\{ \boldsymbol{x} : \boldsymbol{w} \cdot \boldsymbol{x} + b = -1 \}$

Claim: The vector **w** is perpendicular to the Plus Plane.

Why?

- Let **u** and **v** be two vectors on the Plus Plane. What is $\boldsymbol{w} \cdot (\boldsymbol{u} - \boldsymbol{v})$?
- And so of course the vector **w** is also perpendicular to the Minus Plane

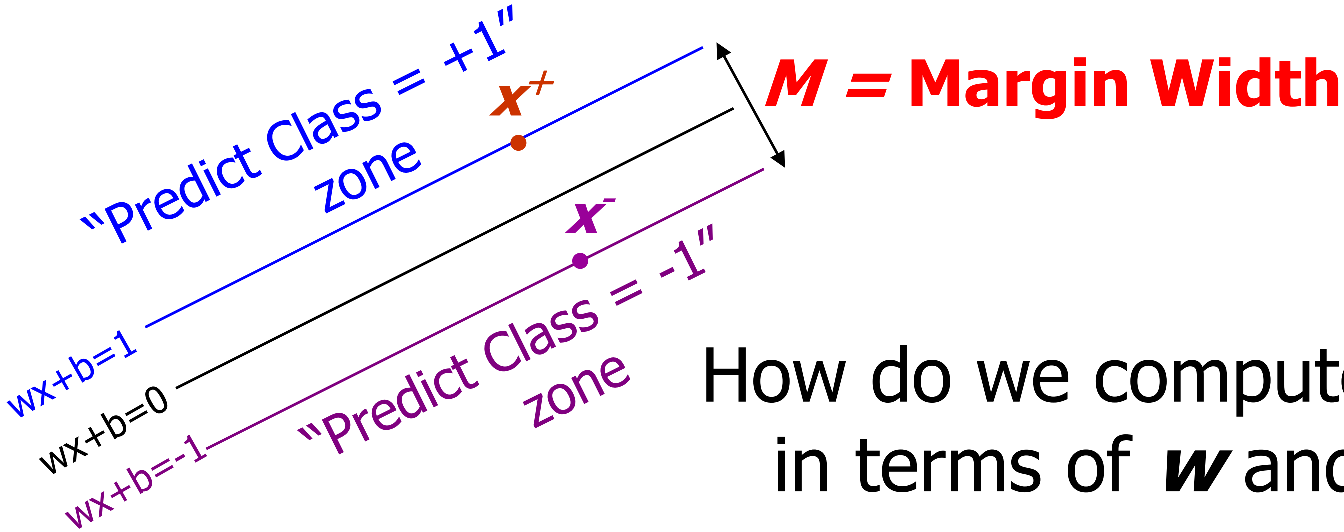
Computing the margin width



How do we compute M in terms of $\boldsymbol{w}$ and b?

- Plus-plane = $\{ \boldsymbol{x} : \boldsymbol{w} . \boldsymbol{x} + b = +1 \}$
- Minus-plane = $\{ \boldsymbol{x} : \boldsymbol{w} . \boldsymbol{x} + b = -1 \}$
- The vector **w** is perpendicular to the Plus Plane
- Let $\boldsymbol{x}^-$ be any point on the minus plane
  - Any location in R^m: not necessarily a datapoint
- Let $\boldsymbol{x}^+$ be the closest plus-plane-point to $\boldsymbol{x}^-$.
- **Claim:** $\boldsymbol{x}^+ = \boldsymbol{x}^- + \lambda \boldsymbol{w}$ for some value of λ. **Why?**

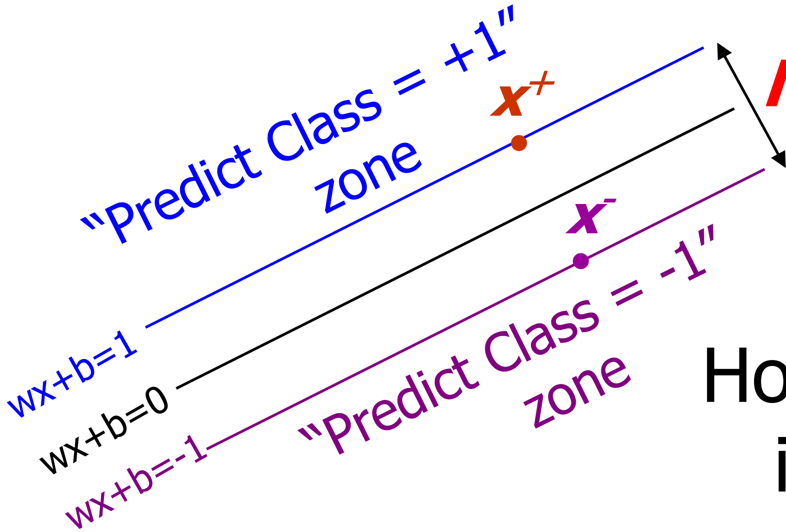
Computing the margin width



How do we compute M in terms of $\boldsymbol{w}$ and b?

- Plus-plane = $\{ \boldsymbol{x} : \boldsymbol{w} \cdot \boldsymbol{x} + b = +1 \}$
- Minus-plane = $\{ \boldsymbol{x} : \boldsymbol{w} \cdot \boldsymbol{x} + b = -1 \}$
- **Claim:** $\boldsymbol{x}^+ = \boldsymbol{x}^- + \lambda \boldsymbol{w}$ for some value of λ. **Why?**
  - The line from $\boldsymbol{x}^-$ to $\boldsymbol{x}^+$ is perpendicular to the planes.
  - So to get from $\boldsymbol{x}^-$ to $\boldsymbol{x}^+$ travel some distance in direction $\boldsymbol{w}$.

Computing the margin width



***M* = Margin Width**

How do we compute *M* in terms of ***w*** and *b*?

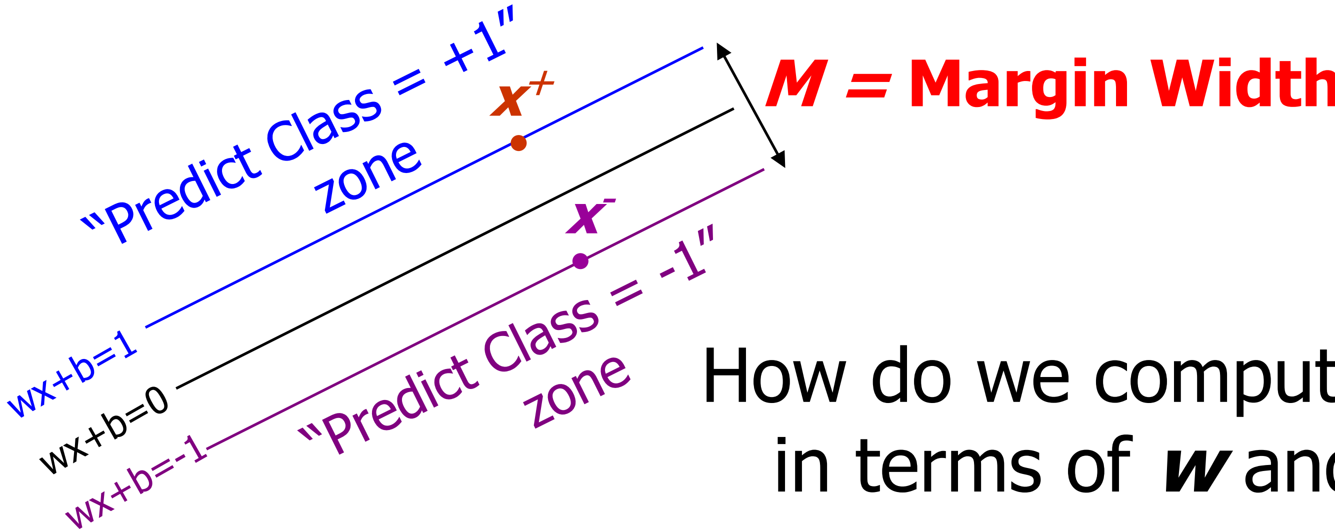
- What we know:

$$\mathbf{w} \cdot \mathbf{x}^+ + b = +1 \qquad \mathbf{w} \cdot \mathbf{x}^- + b = -1$$

$$\mathbf{x}^+ = \mathbf{x}^- + \lambda \mathbf{w} \qquad \|\mathbf{x}^+ - \mathbf{x}^-\|_2 = M$$

- It's now easy to get M in terms of w and b

Computing the margin width



How do we compute M in terms of $\boldsymbol{w}$ and b?

- What we know:

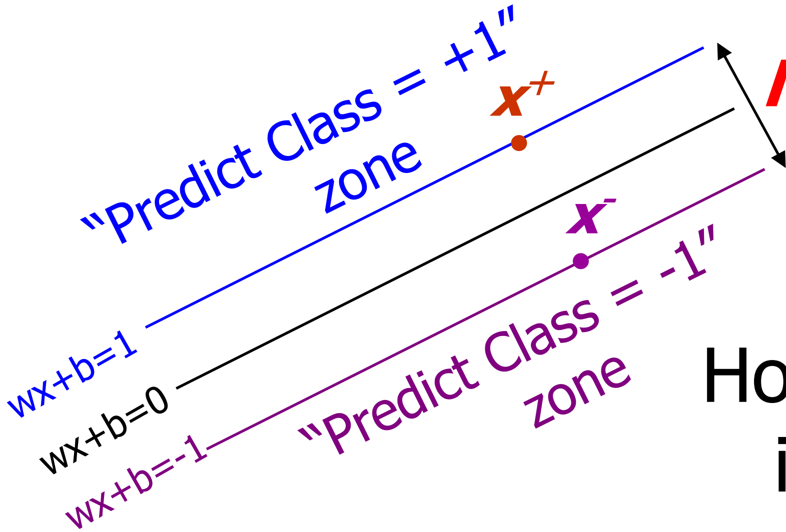
$$\mathbf{w} \cdot \mathbf{x}^+ + b = +1 \qquad \mathbf{w} \cdot (\mathbf{x}^- + \lambda \mathbf{w}) + b = 1$$

$$\mathbf{w} \cdot \mathbf{x}^- + b = -1 \qquad \mathbf{w} \cdot \mathbf{x}^- + b + \lambda \mathbf{w} \cdot \mathbf{w} = 1$$

$$\mathbf{x}^+ = \mathbf{x}^- + \lambda \mathbf{w} \qquad -1 + \lambda \mathbf{w} \cdot \mathbf{w} = 1$$

$$\|\mathbf{x}^+ - \mathbf{x}^-\|_2 = M \qquad \lambda = \frac{2}{\|\mathbf{w}\|_2^2}$$

Computing the margin width



How do we compute *M* in terms of ***w*** and *b* ?

- What we know:

$$\mathbf{w} \cdot \mathbf{x}^+ + b = +1$$

$$\mathbf{w} \cdot \mathbf{x}^- + b = -1$$

$$\mathbf{x}^+ = \mathbf{x}^- + \lambda \mathbf{w}$$

$$\|\mathbf{x}^+ - \mathbf{x}^-\|_2 = M$$

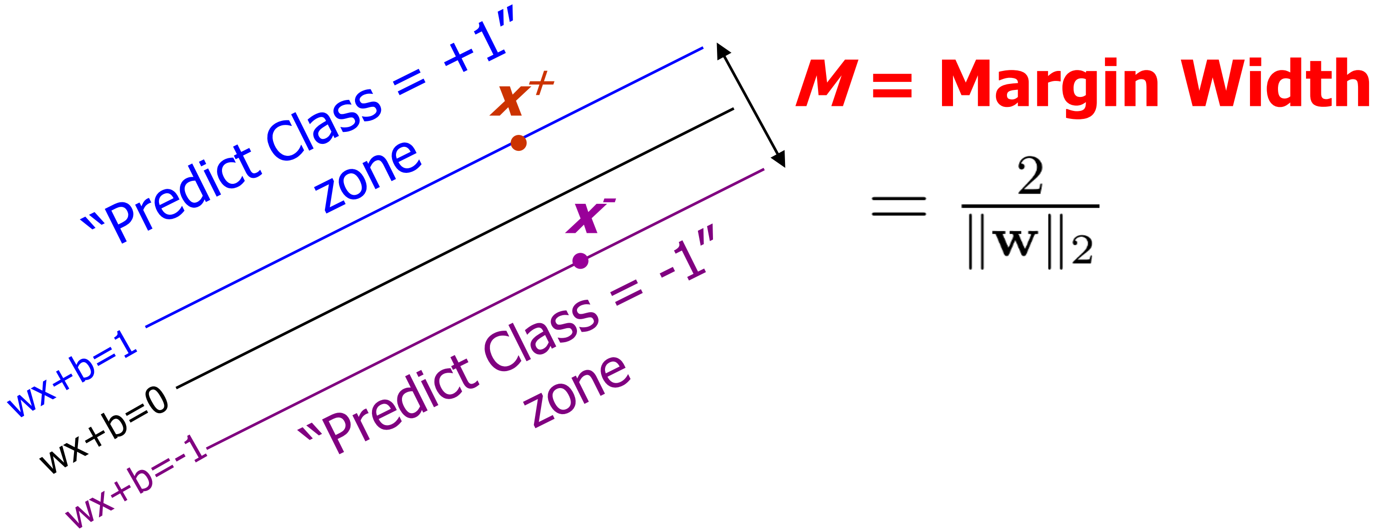
$$\lambda = \frac{2}{\|\mathbf{w}\|_2^2}$$

$$M = \|\mathbf{x}^+ - \mathbf{x}^-\|_2 = \|\lambda \mathbf{w}\|_2 =$$

$$= \lambda \|\mathbf{w}\|_2 = \lambda \sqrt{\|\mathbf{w}\|_2^2}$$

$$= \frac{2\sqrt{\|\mathbf{w}\|_2^2}}{\|\mathbf{w}\|_2^2} = \frac{2}{\|\mathbf{w}\|_2}$$

Computing the margin width

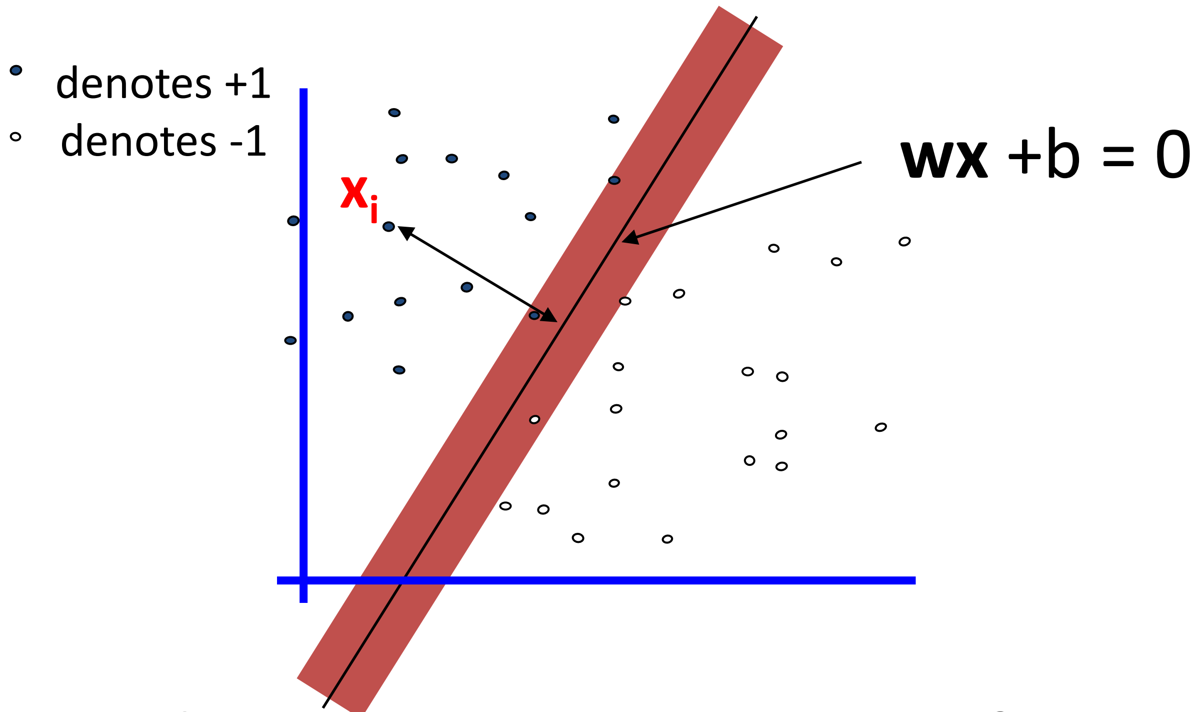


$$= \frac{2}{\|\mathbf{w}\|_2}$$

- Given a guess of **w** and b we can
  - Compute whether all data points in the correct half-planes
  - Compute the width of the margin **How?**
- We just need to write a program to search the space of w's and b's to find the **widest** margin that matches all the datapoints.
  - Gradient descent? Simulated Annealing?
  - Matrix Inversion? EM? Newton's Method?

Estimate the Margin

Estimate the Margin



- What is the distance expression for a point $(\mathbf{x}_i, \mathbf{y}_i)$ to a line **wx**+b= 0?

$$d(\mathbf{x}_i) = \frac{y_i(\mathbf{x}\cdot\mathbf{w}+b)}{\|\mathbf{w}\|_2} = \frac{y_i(\mathbf{x}\cdot\mathbf{w}+b)}{\sqrt{\mathbf{w}^T\mathbf{w}}}$$

Estimate the Margin

- What is the expression for margin?

$$\text{margin} \equiv \min_{\mathbf{x} \in D} d(\mathbf{x}) = \min_{\mathbf{x} \in D} \frac{y_i(\mathbf{x} \cdot \mathbf{w} + b)}{\sqrt{\mathbf{w}^T \mathbf{w}}}$$

$$\operatorname*{argmax}_{\mathbf{w},b} \text{margin}(\mathbf{w}, b, D) = \operatorname*{argmax}_{\mathbf{w},b} \min_{\mathbf{x}_i \in D} d(\mathbf{x}_i)$$

$$= \operatorname*{argmax}_{\mathbf{w},b} \min_{\mathbf{x}_i \in D} \frac{y_i(\mathbf{x} \cdot \mathbf{w} + b)}{\sqrt{\mathbf{w}^T \mathbf{w}}}$$

$$\text{subject to } \forall \mathbf{x}_i \in D : y_i \left(\mathbf{x}_i \cdot \mathbf{w} + b \right) \geq 0$$

- Min-max problem → game problem

Estimate the Margin

- Min-max problem

$$\underset{\mathbf{w},b}{\operatorname{argmax}} \ \min_{\mathbf{x}_i \in D} \frac{y_i(\mathbf{x}\cdot\mathbf{w}+b)}{\sqrt{\mathbf{w}^T\mathbf{w}}}$$

$$\text{subject to } \ \forall \mathbf{x}_i \in D : y_i\left(\mathbf{x}_i \cdot \mathbf{w} + b\right) \geq 0$$

- 线性可分支持向量机的最优化问题

$$\underset{\mathbf{w},b}{\operatorname{argmin}} \ \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

Quadratic criterion

$$\text{subject to } \ \forall \mathbf{x}_i \in D : y_i\left(\mathbf{x}_i \cdot \mathbf{w} + b\right) \geq 1$$

多少个不等式?

linear constraints

- How to solve it? **Quadratic Programming?**

Quadratic Programming

- optimal $\mathbf{u} = QP(\mathbf{Q}, \mathbf{p}, \mathbf{A}, \mathbf{c})$

$$\min_{\mathbf{u}} \ \frac{1}{2}\mathbf{u}^T\mathbf{Q}\mathbf{u} + \mathbf{p}^T\mathbf{u}$$

$$\text{subject to} \quad \mathbf{a}_i^T\mathbf{u} \geq c_i, \quad \text{for } i = 1, 2, \ldots, N$$

- optimal $(b, \mathbf{w}) = QP(\mathbf{Q}, \mathbf{p}, \mathbf{A}, \mathbf{c})$

$$\min_{\mathbf{u}} \ \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

$$\text{subject to} \quad y_i\left(\mathbf{x}_i \cdot \mathbf{w} + b\right) \geq 1 \ \text{ for } i = 1, 2, \ldots, N$$

- objective function: $\mathbf{u} = \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix}$ $\mathbf{Q} = \begin{bmatrix} 0 & \mathbf{0}_k^T \\ \mathbf{0}_k & \mathbf{I}_k \end{bmatrix}$ $\mathbf{p} = \mathbf{0}_{k+1}$

- constraints: $\mathbf{a}_i^T = y_i[1 \ \ \mathbf{x}_i^T]$ $\quad c_i = 1 \quad$ $N: \ \#\ \text{data set}$

Quize

- objective function: $\mathbf{u} = \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix}$ $\quad \mathbf{Q} = \begin{bmatrix} 0 & \mathbf{0}_k^T \\ \mathbf{0}_k & \mathbf{I}_k \end{bmatrix}$ $\quad \mathbf{p} = \mathbf{0}_{k+1}$
- constraints: $\mathbf{a}_i^T = y_i[1 \;\; \mathbf{x}_i^T]$ $\quad c_i = 1$ $\quad N: \#\ \text{data set}$

- Consider two **negative** examples with $\mathbf{x}_1 = (0, 0)$ and $\mathbf{x}_2 = (2, 2)$
- And two **positive** examples with $\mathbf{x}_3 = (2, 0)$ and $\mathbf{x}_4 = (3, 0)$
- Define **u, Q, p, c** as those listed above. **Reference Answer: 4**
- What are $\mathbf{a}_i^T$ that need to be fed into the QP solver?

(1) $a_1^T = [-1, 0, 0],\ a_2^T = [-1, 2, 2],\ a_3^T = [-1, 2, 0],\ a_4^T = [-1, 3, 0]$

(2) $a_1^T = [1, 0, 0],\ a_2^T = [1, -2, -2],\ a_3^T = [-1, 2, 0],\ a_4^T = [-1, 3, 0]$

(3) $a_1^T = [1, 0, 0],\quad a_2^T = [1, 2, 2],\quad a_3^T = [1, 2, 0],\quad a_4^T = [1, 3, 0]$

(4) $a_1^T = [-1, 0, 0],\ a_2^T = [-1, -2, -2],\ a_3^T = [1, 2, 0],\ a_4^T = [1, 3, 0]$

Estimate the Margin

- QP with k + 1 variables and N constraints

$$(b, \mathbf{w}) = QP(\mathbf{Q}, \mathbf{p}, \mathbf{A}, \mathbf{c}) \qquad \mathbf{Q} = \begin{bmatrix} 0 & \mathbf{0}_k^T \\ \mathbf{0}_k & \mathbf{I}_k \end{bmatrix} \qquad \mathbf{a}_i^T = y_i [1 \;\; \mathbf{x}_i^T] \qquad c_i = 1$$

- Challenging if k large ? could reach infinite size in theory!
- Goal: SVM without dependence on k
  - Original SVM : convex QP of
    - **k+1** variables, N constraints
  - Equivalent SVM : convex QP of
    - **N** variables, N+1 constraints
- How? Lagrange Multipliers
  - Build Equivalent SVM based on dual problem of Original SVM

Recap : Linear SVM

- Min-max problem

$$\underset{\mathbf{w},b}{\operatorname{argmax}} \ \min_{\mathbf{x}_i \in D} \frac{y_i(\mathbf{x}\cdot\mathbf{w}+b)}{\sqrt{\mathbf{w}^T\mathbf{w}}}$$

$$\text{subject to} \ \ \forall \mathbf{x}_i \in D : y_i\left(\mathbf{x}_i \cdot \mathbf{w} + b\right) \geq 0$$

- 线性可分支持向量机的最优化问题

$$\underset{\mathbf{w},b}{\operatorname{argmin}} \ \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

Quadratic criterion

$$\text{subject to} \ \ \forall \mathbf{x}_i \in D : y_i\left(\mathbf{x}_i \cdot \mathbf{w} + b\right) \geq 1$$

linear constraints

- How to solve it? **Quadratic Programming?**

Lagrange Dual Problem

- 定义拉格朗日函数：引入拉格朗日乘子 $\alpha_i \geq 0,\ i = 1, \ldots, N$

$$\mathcal{L}(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|_2^2 - \sum_{i=1}^{N} \alpha_i \{1 - y_i(\mathbf{w} \cdot \mathbf{x}_i + b)\}$$

- 固定 $(\mathbf{w}, b)$，对于任意给定的 α' (其中 $\alpha_i' \geq 0$)

$$\min_{\mathbf{w},b} \left(\max_{all\ \alpha_i \geq 0} \mathcal{L}(\mathbf{w}, b, \alpha) \right) \geq \min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \alpha')$$

- 对于上式右侧的最优 α' (其中 $\alpha_i' \geq 0$) ，有：

$$\min_{\mathbf{w},b} \left(\max_{all\ \alpha_i \geq 0} \mathcal{L}(\mathbf{w}, b, \alpha) \right) \geq \max_{all\ \alpha_i' \geq 0} \left(\min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \alpha') \right)$$

- 拉格朗日对偶问题：
  - 外层是关于 α 的最大化问题，内层则是原始问题的下界

Estimate the Margin

- 由拉格朗日对偶性可知：原始问题的对偶问题是max-min问题

$$\min_{\mathbf{w},b}\left(\max_{all\ \alpha_i\geq 0}\ \mathcal{L}(\mathbf{w},b,\alpha)\right)\geq \max_{all\ \alpha_i'\geq 0}\left(\min_{\mathbf{w},b}\ \mathcal{L}(\mathbf{w},b,\alpha')\right)$$

- 强对偶性：若QP问题满足以下条件（constraint qualification）
  - convex primal
  - feasible primal (true if $\Phi(\mathbf{x})$ separable)
  - linear constraints
- 满足强对偶性：因此求解对偶问题即可
  - 思考：如何求解上式右侧的二次规划问题?

Solving Lagrange Dual

$$\max_{all\ \alpha_i \geq 0} \left(\min_{\mathbf{w},b}\ \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{N} \alpha_i (1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)) \right)$$

- 注意到内部的最小化问题是无约束的
- 求解该对偶问题，应首先求得 $\mathcal{L}$ 对 $\mathbf{w}, b$ 的极小，再求对 $\mathcal{L}$ 的极大

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, b, \alpha) = \mathbf{w} - \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i = 0 \quad \Rightarrow \mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i$$

$$\nabla_b \mathcal{L}(\mathbf{w}, b, \alpha) = -\sum_{i=1}^{N} \alpha_i y_i = 0 \quad \Rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0$$

Solving Lagrange Dual

- 代入拉格朗日函数得到:

$$\mathcal{L}(\mathbf{w}, b, \alpha) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{i=1}^{N}\alpha_i y_i\left(\left(\sum_{j=1}^{N}\alpha_j y_j \mathbf{x}_j\right)\cdot \mathbf{x}_i + b\right) + \sum_{i=1}^{N}\alpha_i$$

$$= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(\mathbf{x}_i \cdot \mathbf{x}_j) + \sum_{i=1}^{N}\alpha_i \qquad b\sum_{i=1}^{N}\alpha_i y_i = 0$$

- 即:

$$\min_{\mathbf{w},b}\mathcal{L}(\mathbf{w}, b, \alpha) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(\mathbf{x}_i \cdot \mathbf{x}_j) + \sum_{i=1}^{N}\alpha_i$$

- 接下来求解上式对 α 的极大，即对偶问题

Solving Lagrange Dual

- 求解对偶问题

$$\max_{\alpha}\left(-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(\mathbf{x}_i\cdot\mathbf{x}_j)+\sum_{i=1}^{N}\alpha_i\right)$$

$$\text{s.t.}\quad \sum_{i=1}^{N}\alpha_i y_i=0;\quad \alpha_i\geq 0,\ i=1,2,\cdots,N$$

- 将其转化为求极小，得到如下等价（对偶）问题：

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(\mathbf{x}_i\cdot\mathbf{x}_j)-\sum_{i=1}^{N}\alpha_i$$

$$\text{s.t.}\quad \sum_{i=1}^{N}\alpha_i y_i=0;\quad \alpha_i\geq 0,\ i=1,2,\cdots,N$$

Convex QP of N variables & N + 1 constraints!

Dual SVM with QP Solver

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{i=1}^{N} \alpha_i$$

$$\text{s.t.} \quad \sum_{i=1}^{N} \alpha_i y_i = 0; \quad \alpha_i \geq 0, \; i = 1, 2, \cdots, N$$

- optimal $\alpha = QP(\mathbf{Q}, \mathbf{p}, \mathbf{A}, \mathbf{c})$

$$\min_{\alpha} \quad \tfrac{1}{2} \alpha^T \mathbf{Q} \mathbf{u} + \mathbf{p}^T \alpha$$

$$\text{subject to} \quad \mathbf{a}_i^T \alpha \geq c_i, \quad \text{for } i = 1, 2, \ldots, N$$

- where : $\mathbf{Q}_{ij} = y_i y_j \mathbf{x}_i^T \mathbf{x}_j$

- Problem: if N = 30,000, dense Q takes > 3G RAM
  - need special solver for not storing whole Q by utilizing special constraints properly to scale up to large N

Karush-Kuhn-Tucker Optimality Conditions

- 如果原始问题和对偶问题有相同的最优解 $(\mathbf{w}, b, \alpha)$，需满足:
  - primal feasible: $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$
  - dual feasible: $\alpha_i \geq 0$
  - dual-inner optimal: $\sum_{i=1}^{N} \alpha_i y_i = 0 \quad \mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i$
  - primal-inner optimal: complimentary slackness
    - at optimal all Lagrange terms disappear

$$\alpha_i \left(1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)\right) = 0$$

$$\min_{\mathbf{w},b} \left(\max_{all\ \alpha_i \geq 0} \mathcal{L}(\mathbf{w}, b, \alpha) \right) \geq \max_{all\ \alpha'_i \geq 0} \left(\min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \alpha') \right)$$

Quize

- For a single variable **w**, consider minimizing $\frac{1}{2}\mathbf{w}^2$ subject to two linear constraints $\mathbf{w} \geq 1$ and $\mathbf{w} \leq 3$. We know that the Lagrange function

$$\mathcal{L}(\mathbf{w}, \alpha) = \frac{1}{2}\mathbf{w}^2 + \alpha_1(1 - \mathbf{w}) + \alpha_2(\mathbf{w} - 3)$$

- Which of the following equations that contain α are among the KKT conditions of the optimization problem?

(1) $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$

(2) $\mathbf{w} = \alpha_1 - \alpha_2$

(3) $\alpha_1(1 - \mathbf{w}) = 0$ and $\alpha_2(\mathbf{w} - 3)$

(4) all of the above

Reference Answer: 4

KKT Optimality Conditions

- 根据定理2和3可知，KKT条件成立
  - 设 $\alpha^* = (\alpha_1^*, \alpha_2^*, \cdots, \alpha_l^*)$ 是对偶优化问题的解，由KKT条件可知

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}^*, b^*, \alpha^*) = \mathbf{w}^* - \sum_{i=1}^{N} \alpha_i^* y_i \mathbf{x}_i = 0$$

$$\nabla_b \mathcal{L}(\mathbf{w}^*, b^*, \alpha^*) = -\sum_{i=1}^{N} \alpha_i^* y_i = 0$$

$$\alpha_i^* (y_i (\mathbf{w}^* \cdot \mathbf{x}_i + b^*) - 1) = 0, \quad i = 1, 2, \ldots, N$$

$$y_i (\mathbf{w}^* \cdot \mathbf{x}_i + b^*) - 1 \geq 0, \quad i = 1, 2, \ldots, N$$

$$\alpha_i^* \geq 0, \quad i = 1, 2, \ldots, N$$

- 由此可得：

$$\text{for any given } \alpha_j > 0$$

$$\mathbf{w}^* = \sum_{i=1}^{N} \alpha_i^* y_i \mathbf{x}_i \qquad b^* = y_j - \sum_{i=1}^{N} \alpha_i^* y_i (\mathbf{x}_i \cdot \mathbf{x}_j)$$

Quize

- Consider two transformed examples $(\mathbf{x}_1, +1)$ and $(\mathbf{x}_2, -1)$ with $\mathbf{x}_1 = \mathbf{x}$ and $\mathbf{x}_2 = -\mathbf{x}$. After solving the dual problem of hard-margin SVM, assume that the optimal α_1 and α_2 are both strictly positive. What is the optimal *b*?

(1) -1 (2) 0 (3) 1

(4) not certain with the descriptions above

- **Hints :** with the descriptions, **x** located on the margin

$$y_i(\mathbf{w}^* \cdot \mathbf{x} + b^*) - 1 = 0 \Rightarrow b^* = y_i - \mathbf{w}^* \cdot \mathbf{x}$$

$$b^* = 1 - \mathbf{w}^* \cdot \mathbf{x} = -1 - \mathbf{w}^* \cdot (-\mathbf{x})$$

Reference Answer: 2

Estimate the Margin

- 分类超平面可以表示为：

$$\sum_{i=1}^{N} \alpha_i^* y_i (\mathbf{x} \cdot \mathbf{x}_i) + b^* = 0$$

- 分类决策函数可以表示为：

$$f(x) = \mathrm{sign}\left(\sum_{i=1}^{N} \alpha_i^* y_i (\mathbf{x} \cdot \mathbf{x}_i) + b^* \right)$$

- 分类决策函数只依赖于输入x和训练样本特征向量的内积
- 上式称为线性可分支持向量机的对偶形式

Support Vectors Revisited

- 将 $\alpha_i > 0$ 的样本$(\mathbf{x}_i, y_i)$称为 support vectors

$$SV \text{ (positive } \alpha_i) \subseteq SV \text{ candidates (on boundary)}$$

- only SV needed to compute w:

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i = \sum_{SV} \alpha_i y_i \mathbf{x}_i$$

- only SV needed to compute b:

$$b = y_i - \mathbf{w}^T \mathbf{x}_i \ \text{ with any SV : } (\mathbf{x}_i; y_i)$$

- 支持向量机的基本思想：求解最大边界的分类超平面
  - 方法：通过求解对偶问题识别出支持向量

SVM与PLA的对比

- SVM

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i (y_i \mathbf{x}_i)$$

- α_i from dual solution

- PLA

$$\mathbf{w} = \sum_{i=1}^{N} \beta_i (y_i \mathbf{x}_i)$$

- β_i by # mistake corrections

$$\mathbf{w} = \text{linear combination of } y_i \mathbf{x}_i$$

- also true for SGD-based LogReg/LinReg when $w_0 = 0$
- call **w** ***represented by data***
- SVM: represent **w** by support vectors only

Summary: Two Forms of Hard-Margin SVM

- Primal Hard-Margin SVM
  - k + 1 variables, N constraints
  - suitable when k small
  - physical meaning: locate specially-scaled (b, w)
- Dual Hard-Margin SVM
  - N variables, N + 1 simple constraints
  - suitable when N small 思考：真的与k无关吗?
  - physical meaning: locate SVs (x_i, y_i) & their \alpha_i
- both result in optimal (b,w) for fattest hyperplane

$$g_{svm}(\mathbf{x}) = sign(\mathbf{w}^T\mathbf{x} + b)$$

电子科技大学

UeSTC

1956