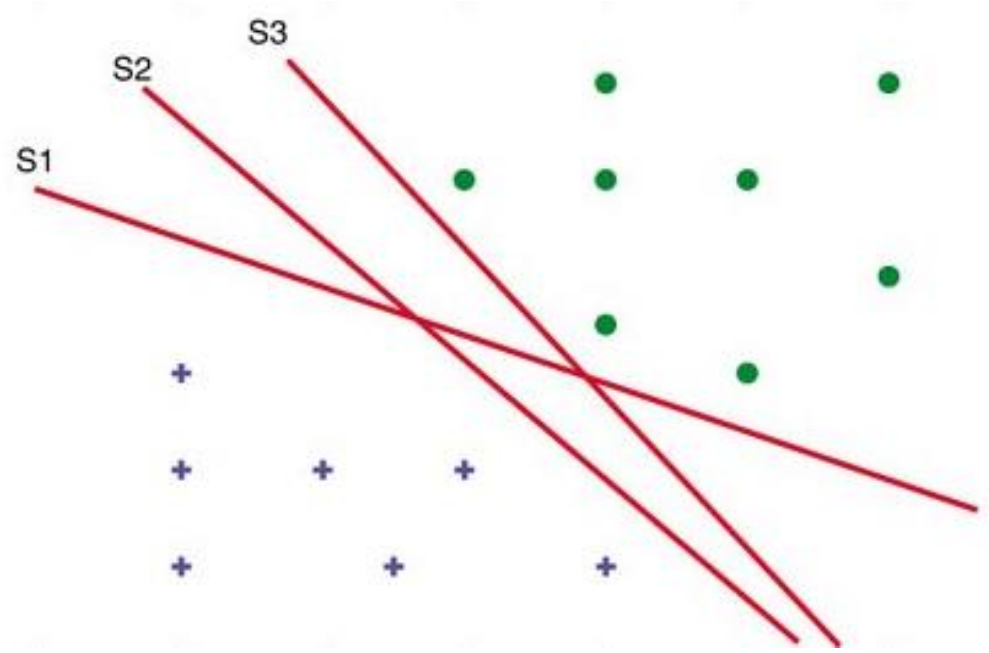
支持向量机

Support Vector Machines



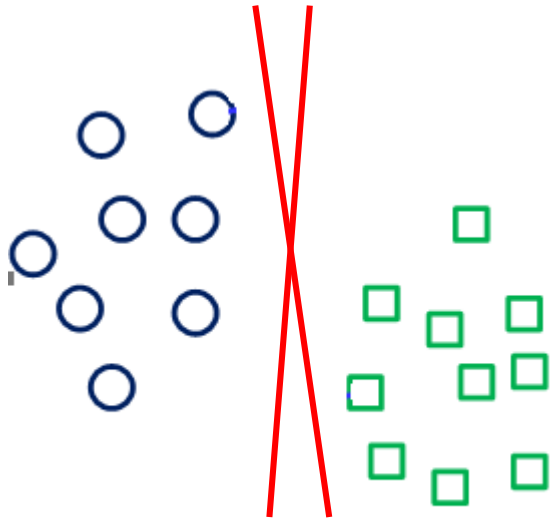
赵海涛

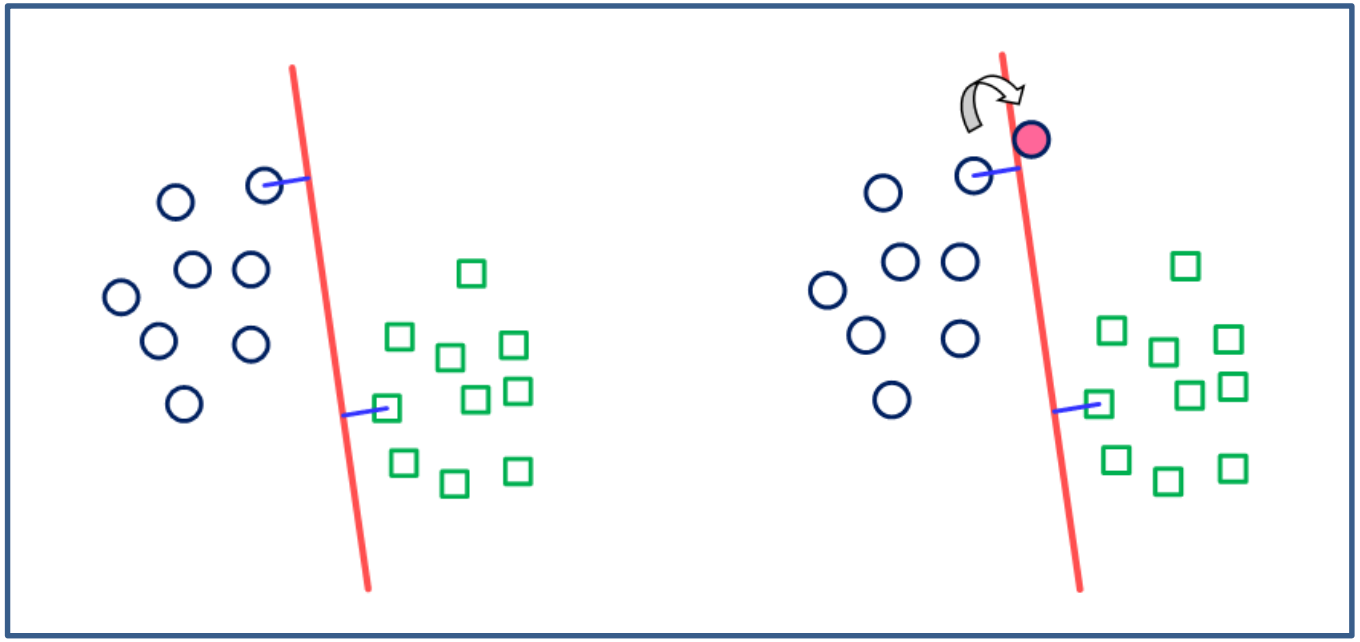
haitaozhao@ecust.edu.cn

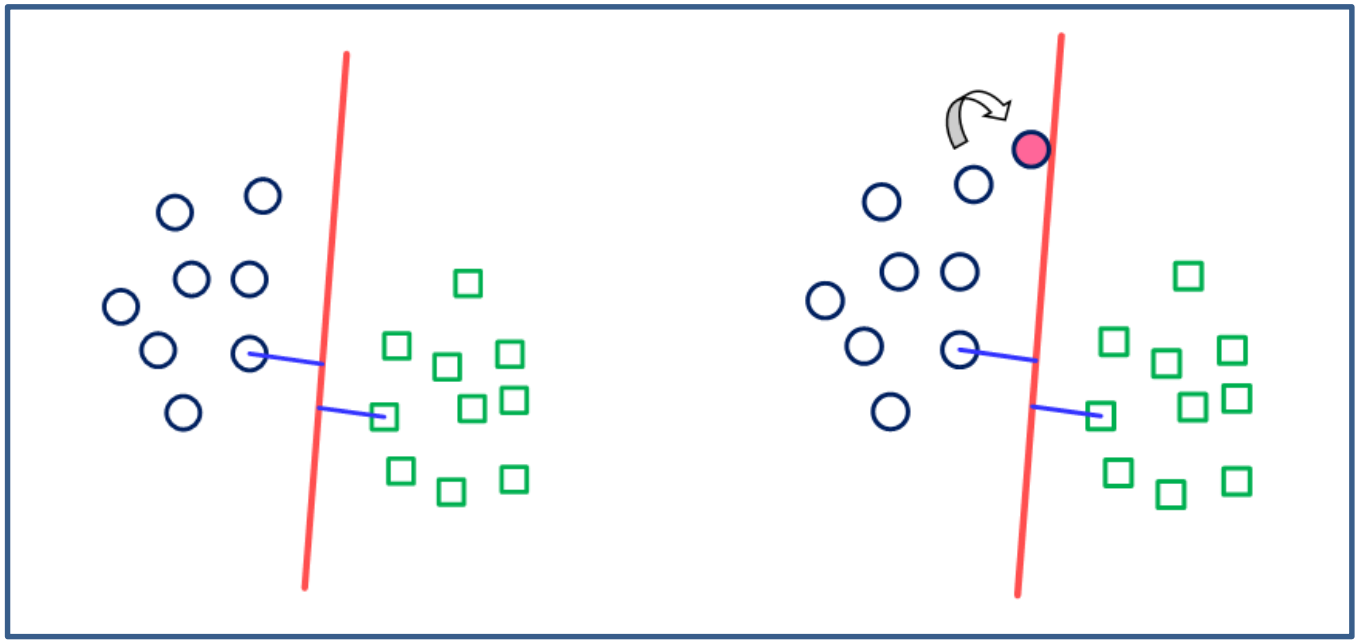
大纲

- 支持向量机（线性可分情况）
- 线性不可分
- 非线性扩展

开始一张图







学习和经验风险最小化

- 任何学习器的目的都是通过最小化某种误差函数来从有限的一组观测值中估计$g(x)$，例如经验风险：

$$R_{\text{emp}}(w, w_0) = \frac{1}{N}\sum_{k=1}^{N}[y_k - g(\boldsymbol{x}_k, w, w_0)]^2$$

类别标签:

$$y_k = \begin{cases} +1 & if\ \boldsymbol{x}_k \in \omega_1 \\ -1 & if\ \boldsymbol{x}_k \in \omega_2 \end{cases}$$

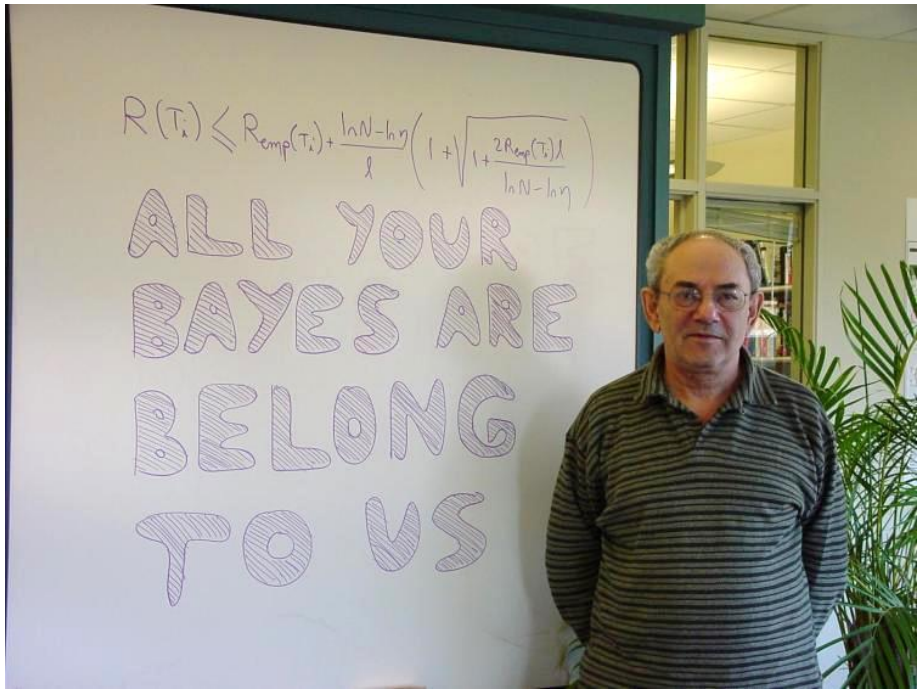
学习和经验风险最小化

- 对训练数据进行常规的经验性风险最小化，并不意味着可以很好地推广到测试数据。
  - 有许多不同的函数，它们都可能很好地划分训练数据集
  - 很难确定最能拟合数据分布真正基础结构的函数

统计学习： 容量 与 VC维

- 为了保证泛化误差的上界，必须控制学习函数的容量（**Capacity**）
- 在统计学习中，**Vapnik-Chervonenkis（VC）维度**是最常用的容量度量方法之一

Vladimir Naumovich Vapnik

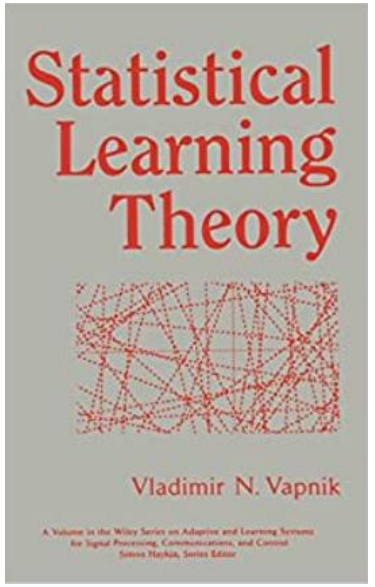


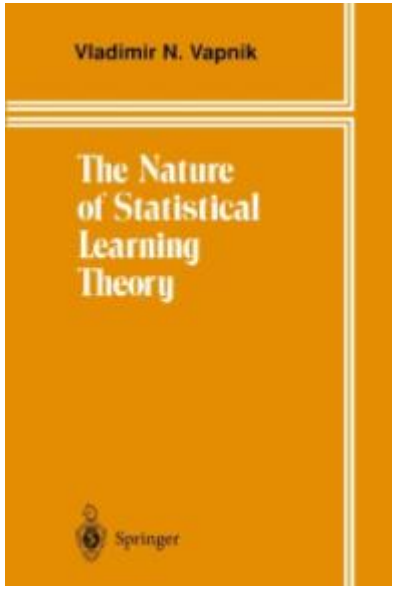
弗拉基米尔·纳乌莫维奇·瓦普尼克（英语：Vladimir Naumovich Vapnik，1936年12月6日－），俄罗斯统计学家、数学家。他是VC理论（Vapnik Chervonenkis theory）的主要创建人之一。

瓦普尼克出生于苏联。1958年，他在撒马尔罕（现属乌兹别克斯坦）的乌兹别克国立大学完成了硕士学业。1964年，他于莫斯科的控制科学学院获得博士学位。毕业后，他一直在该校工作直到1990年，在此期间，他成为了该校*计算机科学与研究系*的系主任。

1995年，他被伦敦大学聘为*计算机与统计科学专业*的教授。

1991至2001年间，他工作于AT&T贝尔实验室（后来的香农实验室），并和他的同事们一起发明了支持向量机理论。他们为机器学习的许多方法奠定了理论基础。





统计学习： 容量 与 VC维

- 数据集： 训练样本 $\boldsymbol{x}_i \in \mathbb{R}^n, i = 1,2,\cdots,N$， 对应的标签 $y_i \in \{-1,1\}$
- 假设： 训练样本 $(\boldsymbol{x}_i, y_i)$ 是独立同分布，从某个概率分布P中抽样得到
- 分类器： $y = h(\boldsymbol{x}, \theta)$ ， 其中θ 是需要学习的参数
- 分类器所属的集合： $\mathcal{H} = \{h(\cdot, \theta)\}$

$$\text{经验风险：}\ \mathrm{R}_{\mathrm{train}} = \frac{1}{2N}\sum\nolimits_i (y_i - \mathrm{sgn}(h(\boldsymbol{x}_i, \theta)))^2$$

$$\text{平均风险：}\ \mathrm{R}_{\mathrm{true}} = \frac{1}{2}\int (y - \mathrm{sgn}(h(\boldsymbol{x}, \theta)))^2 dP(\boldsymbol{x}, y)$$

结构风险最小化

- 一个函数具有：

（1）最小化经验风险

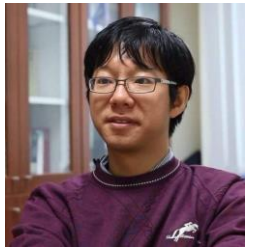
（2）VC维低

则无论输入空间的维度如何，都能很好地泛化 （**结构风险最小化**）

$$\mathrm{R}_{\mathrm{true}} \leq \mathrm{R}_{\mathrm{train}} + \sqrt{\frac{1}{8N}\log\frac{4(2N)^{d^{VC}}+4}{\delta}}$$

$$\text{with probability } (1-\delta)$$

.



[1] **Y. S. Abu-Mostafa**, M. Magdon-Ismail, and **H.-T. Lin**, Learning from Data – A Short Course, AMLbook.com, 2012

支持向量机概述

- 支持向量机（support vector machines）执行结构化风险最小化以实现良好的泛化
- 优化标准是两类的边界之间的宽度（分离裕度）
- 训练等同于解决带有**线性约束的二次规划问题**
- 主要用于二分类器，但可以扩展到多个类 （one vs. one, one vs. all）

分离裕度和最优超平面

- Vapnik已经证明，最大化类之间的分离裕度等同于最小化VC维
- 最优超平面是能够得到类之间最大分离裕度的超平面

Theorem (Vapnik, 1982):

- Given N data points in $\mathbb{R}^D$: $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ with $||\mathbf{x}_n|| \leq R$
- Define $\mathcal{H}_\gamma$: set of classifiers in $\mathbb{R}^D$ having margin γ on $\mathbf{X}$

The VC dimension of $\mathcal{H}_\gamma$ is bounded by:

$$VC(\mathcal{H}_\gamma) \leq \min\left\{D, \left\lceil \frac{4R^2}{\gamma^2} \right\rceil\right\}$$

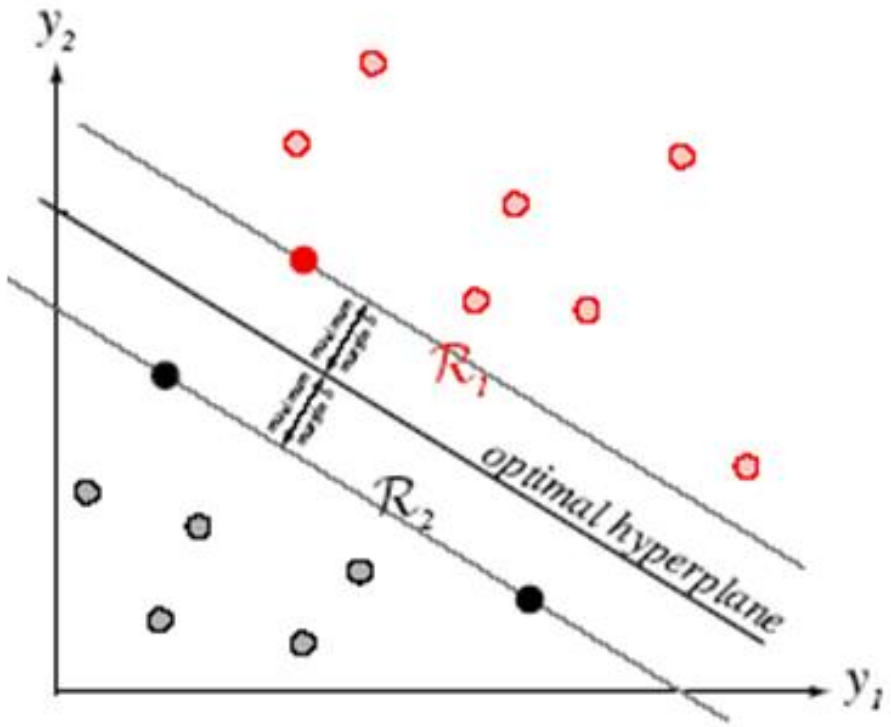
Generalization bound for the SVM:

$$\text{ExpectedLoss}(h) \leq \text{TrainingLoss}(h) + \sqrt{\frac{VC(\mathcal{H}_\gamma)(\log \frac{2N}{VC(\mathcal{H}_\gamma)} + 1) + \log \frac{4}{\delta}}{2N}}$$

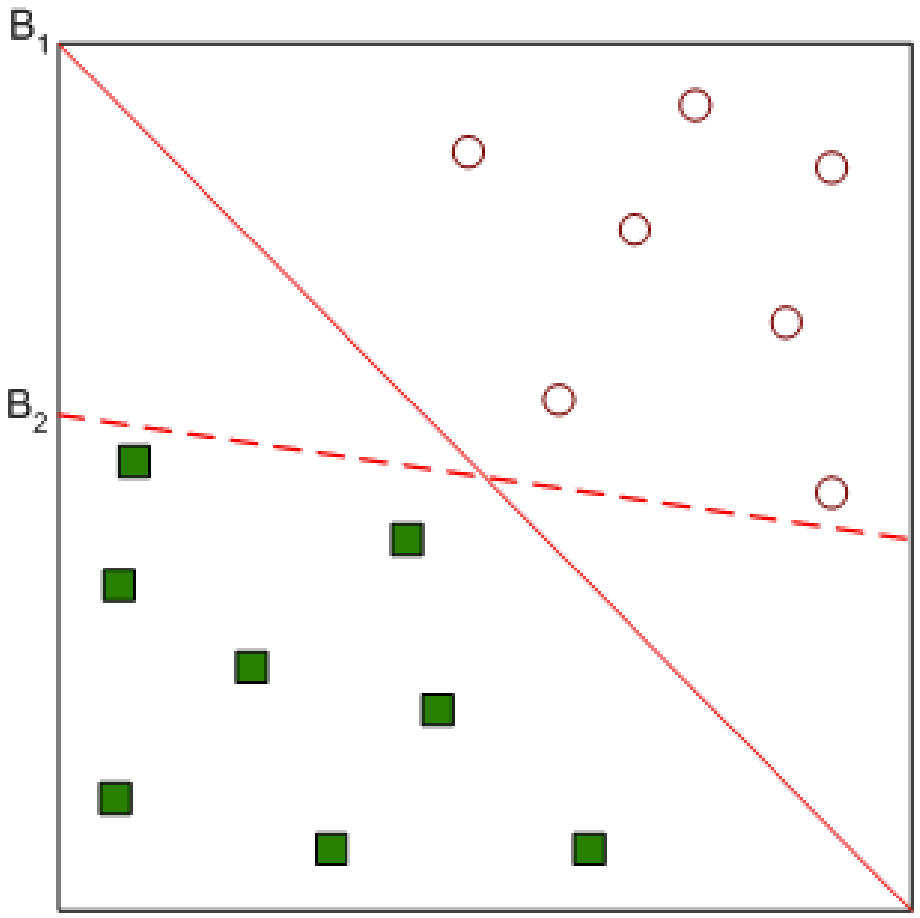
Large $\gamma \Rightarrow$ small VC dim. $\Rightarrow$ small complexity of $\mathcal{H}_\gamma \Rightarrow$ good generalization

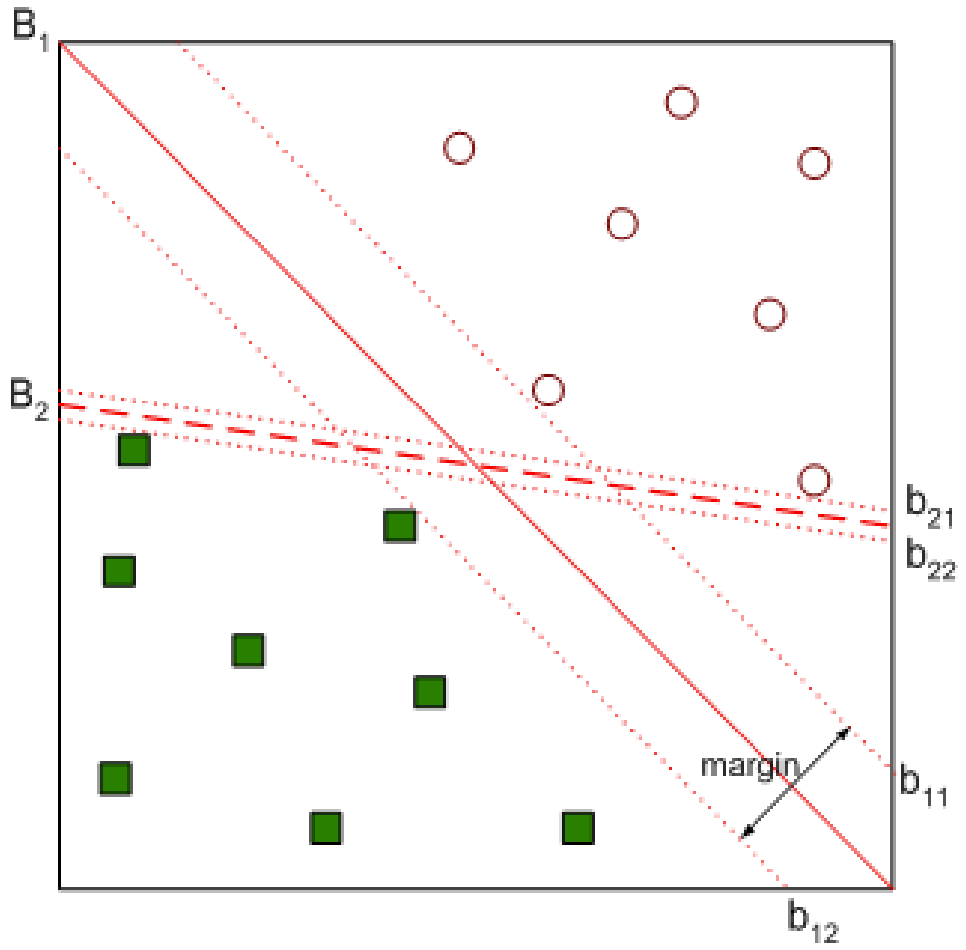
支持向量

- 决策边界周围的空区域，由距离最近的训练模式（即支持向量）确定
- 这些是最难分类的模式

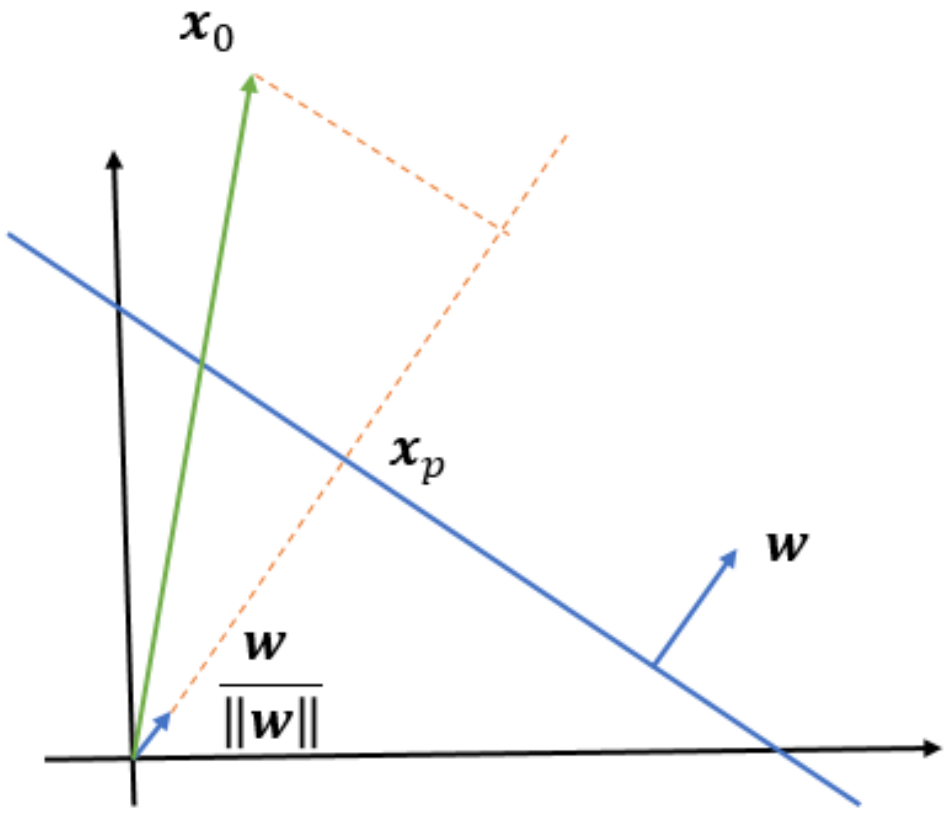


支持向量

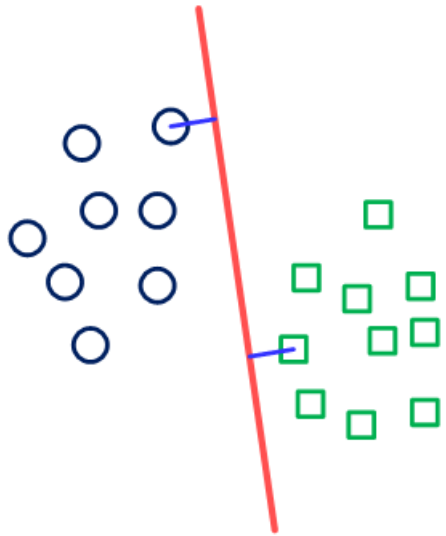




点到直线的距离计算 （revisited）



线性可分的目标函数



用$\boldsymbol{x}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$表示数据集，$y=\{y_1,y_2,\cdots,y_n\}$表示对应元素所属类的判别值。则有$y_i=\begin{cases}1 & \boldsymbol{x}_i\in\omega_1\\ -1 & \boldsymbol{x}_i\in\omega_2\end{cases}$

找到分类超平面$g(\boldsymbol{x})=\boldsymbol{0}$，并且有

$$g(\boldsymbol{x}_i)>0 \quad , \quad \boldsymbol{x}_i\in\omega_1$$

$$g(\boldsymbol{x}_i)<0 \quad , \quad \boldsymbol{x}_i\in\omega_2$$

或者可以写成：

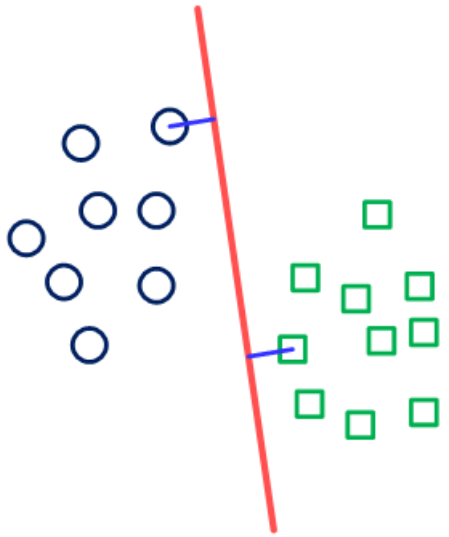
$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i+\boldsymbol{b})>0$$

那么所要求的参数可以通过解下面的优化问题获得：

$$\begin{cases}(\boldsymbol{w}^*,\boldsymbol{b}^*)=\underset{\boldsymbol{w},\boldsymbol{b}}{\operatorname{argmax}}\left(\min\limits_i\dfrac{|g(\boldsymbol{x}_i)|}{\|\boldsymbol{w}\|}\right)\\ s.t. \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i+\boldsymbol{b})\geq\varepsilon>0\end{cases}$$

其中，$\varepsilon=\min\limits_i y_i(\boldsymbol{w}^T\boldsymbol{x}_i+\boldsymbol{b})$。

线性可分的目标函数



假设有：$\tilde{\boldsymbol{w}} = \frac{\boldsymbol{w}}{\varepsilon}$　　$\tilde{\boldsymbol{b}} = \frac{\boldsymbol{b}}{\varepsilon}$

那么就有：

$$r = \frac{|g(\boldsymbol{x}_i)|}{\|\boldsymbol{w}\|} = \frac{|g(\boldsymbol{x}_i)|/\varepsilon}{\|\boldsymbol{w}\|/\varepsilon} = \frac{|(\boldsymbol{w}^T\boldsymbol{x}_i + \boldsymbol{b})|/\varepsilon}{\left\|\frac{\boldsymbol{w}}{\varepsilon}\right\|} = \frac{\left|\tilde{\boldsymbol{w}}^T\boldsymbol{x}_i + \tilde{\boldsymbol{b}}\right|}{\|\tilde{\boldsymbol{w}}\|} = \frac{|\tilde{g}(\boldsymbol{x}_i)|}{\|\tilde{\boldsymbol{w}}\|}$$

即：

$$\frac{|g(\boldsymbol{x}_i)|}{\|\boldsymbol{w}\|} = \frac{|\tilde{g}(\boldsymbol{x}_i)|}{\|\tilde{\boldsymbol{w}}\|}$$

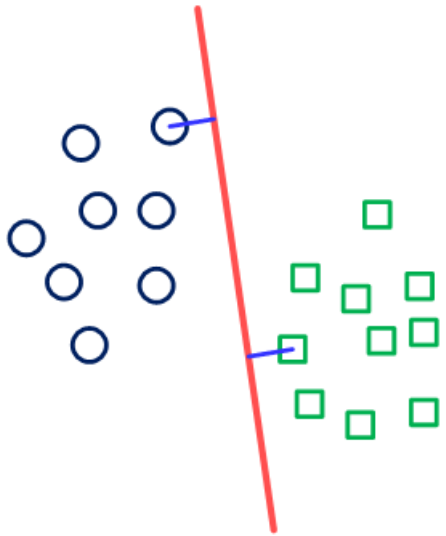
$$\begin{cases} (\boldsymbol{w}^*, \boldsymbol{b}^*) = \underset{\boldsymbol{w},\boldsymbol{b}}{\operatorname{argmax}} \left(\min_i \frac{|g(\boldsymbol{x}_i)|}{\|\boldsymbol{w}\|} \right) \\ s.t. \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + \boldsymbol{b}) \geq \varepsilon > 0 \end{cases}$$

其中，$\varepsilon = \min_i y_i(\boldsymbol{w}^T\boldsymbol{x}_i + \boldsymbol{b})$。

那么优化问题可以转化为：

$$\begin{cases} (\tilde{\boldsymbol{w}}^*, \tilde{\boldsymbol{b}}^*) = \underset{\boldsymbol{w},\boldsymbol{b}}{\operatorname{argmax}} \left(\min_i \frac{|\tilde{g}(\boldsymbol{x}_i)|}{\|\tilde{\boldsymbol{w}}\|} \right) \\ s.t. \quad y_i(\tilde{\boldsymbol{w}}^T\boldsymbol{x}_i + \tilde{\boldsymbol{b}}) \geq 1 \end{cases}$$

线性可分的目标函数



$$\begin{cases} (\tilde{\boldsymbol{w}}^*, \tilde{\boldsymbol{b}}^*) = \underset{\boldsymbol{w},\boldsymbol{b}}{\operatorname{argmax}} \left(\dfrac{1}{\|\tilde{\boldsymbol{w}}\|} \right) \\ s.t. \quad y_i(\tilde{\boldsymbol{w}}^T \boldsymbol{x}_i + \tilde{\boldsymbol{b}}) \geq 1 \ (i = 1,2,\cdots,n) \end{cases}$$

由于要求上述式子的极大，相当于求解

$$\begin{cases} (\tilde{\boldsymbol{w}}^*, \tilde{\boldsymbol{b}}^*) = \underset{\boldsymbol{w},\boldsymbol{b}}{\operatorname{argmin}}(\|\tilde{\boldsymbol{w}}\|) \\ s.t. \quad y_i(\tilde{\boldsymbol{w}}^T \boldsymbol{x}_i + \tilde{\boldsymbol{b}}) \geq 1 \ (i = 1,2,\cdots,n) \end{cases}$$

那么优化问题可以转化为：

$$\begin{cases} (\tilde{\boldsymbol{w}}^*, \tilde{\boldsymbol{b}}^*) = \underset{\boldsymbol{w},\boldsymbol{b}}{\operatorname{argmax}} \left(\min_i \dfrac{|\tilde{g}(\boldsymbol{x}_i)|}{\|\tilde{\boldsymbol{w}}\|} \right) \\ s.t. \quad y_i(\tilde{\boldsymbol{w}}^T \boldsymbol{x}_i + \tilde{\boldsymbol{b}}) \geq 1 \end{cases}$$

$$\mathcal{P}: \quad \begin{cases} (\boldsymbol{w}^*, \boldsymbol{b}^*) = \underset{\boldsymbol{w},\boldsymbol{b}}{\operatorname{argmin}} \left(\dfrac{1}{2} \|\boldsymbol{w}\|^2 \right) \\ s.t. \quad y_i(\boldsymbol{w}^T \boldsymbol{x}_i + \boldsymbol{b}) \geq 1 \end{cases}$$

线性分离超平面

- 训练数据: $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_N, y_N)$
- 寻找一个分离的超平面满足下面的约束

$$\forall i \in (1, \cdots, N) \text{ find } \boldsymbol{w}, b \text{ such that}$$

$$\boldsymbol{w}^T \boldsymbol{x}_i + b \geq +1 \quad \text{if } y_i = +1$$

$$\boldsymbol{w}^T \boldsymbol{x}_i + b \leq -1 \quad if \ y_i = -1$$

- 或者

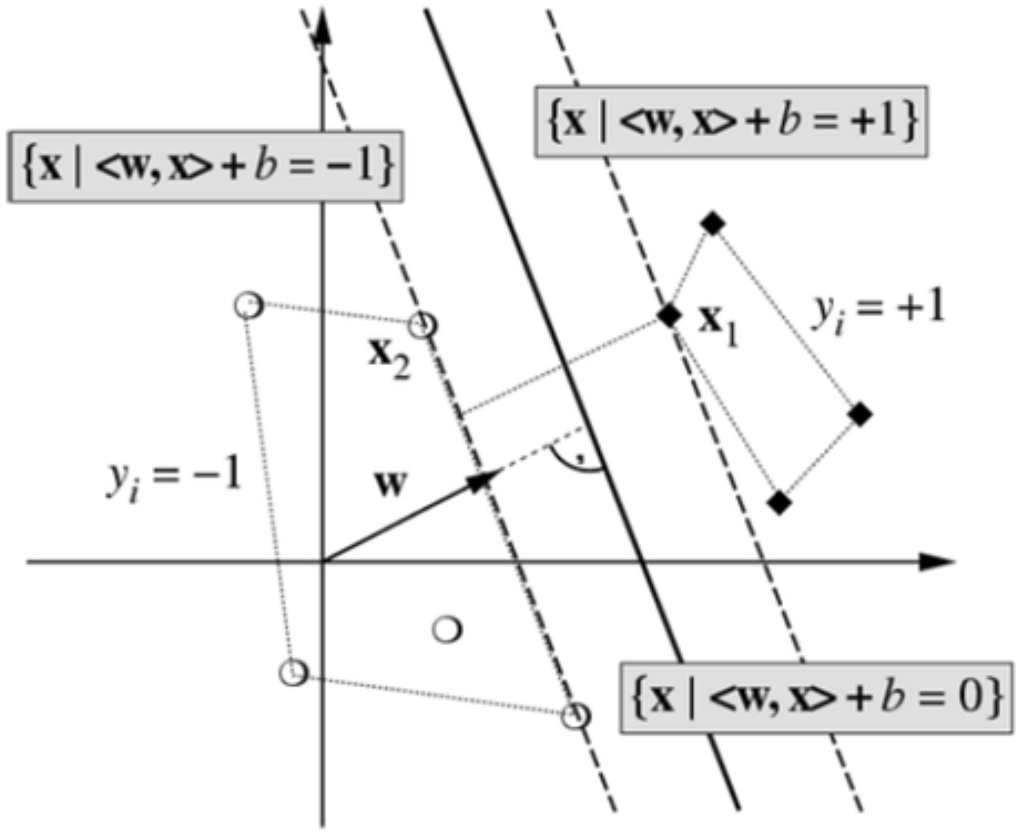
$$y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) - 1 \geq 0, \quad \forall i$$

分类器的裕度

- 取任意超平面（P0）分离数据
- 将平行于P0的超平面（P1）放置在类标签为1的最接近P0的样本点上
- 将第二个平行于P0的超平面（P2）放置在类标签为-1的最接近P0的样本点上
- 裕度是P1和P2之间的垂直距离
- 两超平面之间的距离

$$\text{margin} = \frac{2}{\| \mathbf{w} \|}$$

- 寻找超平面P0，最大化分离裕度



SVM: 带约束的优化问题

- 训练样本：$(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_N, y_N)$
- 最小化 $\|\boldsymbol{w}\|^2$

  subject to: $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - 1 \geq 0 \quad \forall i = (1, \cdots, N)$
- 引入拉格朗日乘子 $\alpha_i \geq 0$ ，有

$$L(\boldsymbol{w}, b, \alpha) = \frac{1}{2}||\boldsymbol{w}||^2 - \sum_{i=1}^{N} \alpha_i \left[y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - 1\right]$$

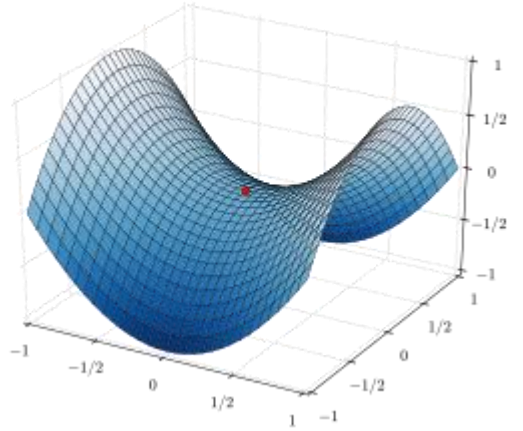
- L关于原问题变量 $\boldsymbol{w}$ 和 b 最小化 ，关于对偶问题变量α_i最大化

我是演草纸

$$\mathcal{P}: \quad \begin{cases} (\boldsymbol{w}^*, \boldsymbol{b}^*) = \underset{\boldsymbol{w}\,,\boldsymbol{b}}{\operatorname{argmin}} \left(\frac{1}{2} \|\boldsymbol{w}\|^2 \right) \\ s.t. \quad 1 - y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i + \boldsymbol{b} \right) \leq 0 \end{cases}$$

the Lagrangian function: $L(\boldsymbol{w}, b, \alpha) = \frac{1}{2} ||\boldsymbol{w}||^2 + \sum_{i=1}^{N} \alpha_i \left[1 - y_i (\boldsymbol{w}^T \boldsymbol{x}_i + b) \right]$

对偶问题的推导



- 在鞍点（saddle point，minimax point）

$$\frac{\partial}{\partial b}L(\boldsymbol{w},b,\boldsymbol{\alpha})=0,\quad \frac{\partial}{\partial \boldsymbol{w}}L(\boldsymbol{w},b,\boldsymbol{\alpha})=0$$

- 由此可得

$$\sum_{i=1}^{N}\alpha_i\,y_i=0,\quad \boldsymbol{w}=\sum_{i=1}^{N}\alpha_i\,y_i\boldsymbol{x}_i$$

- 代入 $L(\boldsymbol{w},b,\boldsymbol{\alpha})$ 得到对偶问题

KKT 条件 (Karush-Kuhn-Tucker (KKT) conditions)

$$\mathcal{P}: \quad \begin{cases} (\boldsymbol{w}^*, \boldsymbol{b}^*) = \underset{\boldsymbol{w}, \boldsymbol{b}}{\operatorname{argmin}} \left(\frac{1}{2} \|\boldsymbol{w}\|^2 \right) \\ s.t. \quad 1 - y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i + \boldsymbol{b} \right) \leq 0 \end{cases}$$

- KKT条件是$\boldsymbol{w}^*, \boldsymbol{b}^*$是原问题$\mathcal{P}$的一个局部极小点的一阶必要条件：存在拉格朗日乘子 α_i，满足下列条件

$$\begin{cases} \boldsymbol{w}^* - \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i = 0 & \text{Stationarity} \\ \sum_{i=1}^{N} \alpha_i y_i = 0 & \\ 1 - y_i \left(\boldsymbol{w}^{*T} \boldsymbol{x}_i + \boldsymbol{b}^* \right) \leq 0 & \text{Primal feasibility} \\ \alpha_i \geq 0 & \text{Dual feasibility} \\ \alpha_i \left[1 - y_i \left(\boldsymbol{w}^{*T} \boldsymbol{x}_i + \boldsymbol{b}^* \right) \right] = 0 & \text{Complementary slackness} \end{cases}$$

对偶问题

the Lagrangian function: $L(\boldsymbol{w}, b, \alpha) = \frac{1}{2}||\boldsymbol{w}||^2 + \sum_{i=1}^{N} \alpha_i \left[1 - y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b)\right]$

$$\mathfrak{D}: \begin{cases} \boldsymbol{\alpha}^* = \underset{\boldsymbol{\alpha}}{\operatorname{argmax}} \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_i \right) \\ s.t. \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \\ \alpha_i \geq 0 \ (i = 1,2,\cdots,n) \end{cases}$$

- 对于支持向量机，有

$$d^* = \max_{\alpha} \min_{\boldsymbol{w},b} L(\boldsymbol{w}, b, \alpha) = \min_{\boldsymbol{w},b} \max_{\alpha} L(\boldsymbol{w}, b, \alpha) = p^*$$

支持向量

$$\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i \, y_i \boldsymbol{x}_i$$

$$y_i[\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b] > 1 \Rightarrow \alpha_i = 0 \rightarrow \boldsymbol{x}_i \quad \text{无关点}$$

OR

$$y_i[\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b] = 1 \ \ (\text{On Margin}) \rightarrow \boldsymbol{x}_i \quad \text{支持向量}$$

分类器

- 考虑分类器

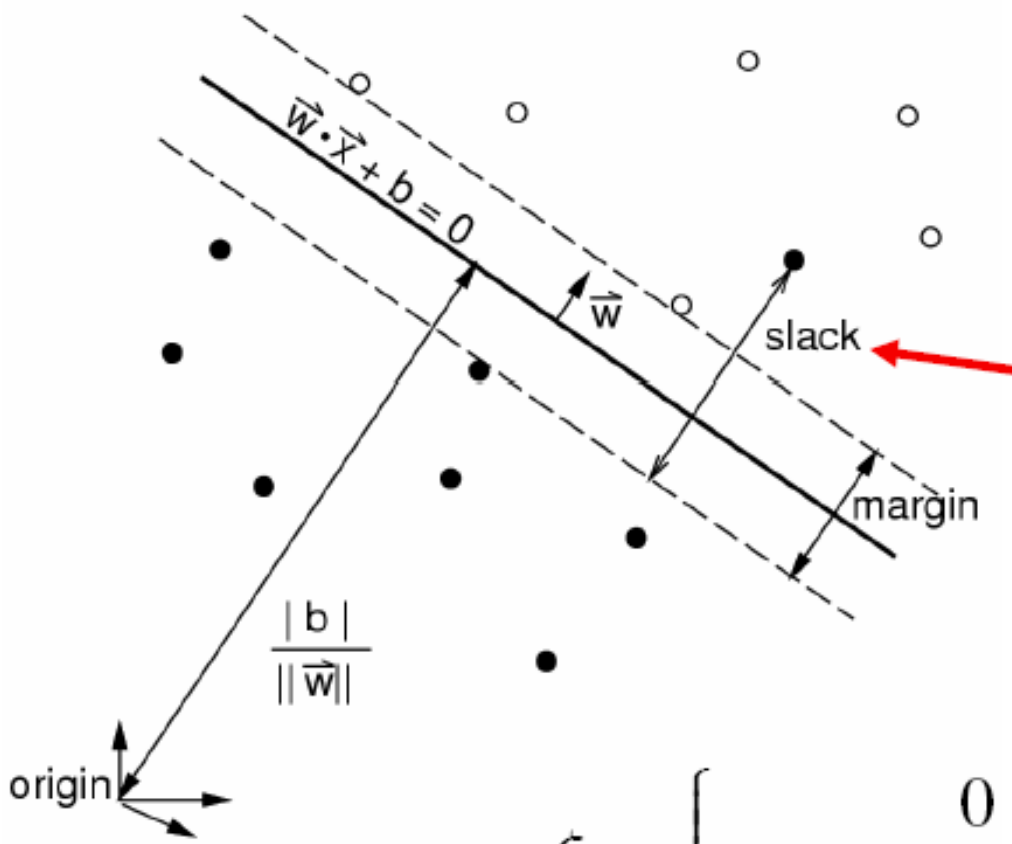
$$f(\boldsymbol{x}) = \operatorname{sgn}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b)$$

- 将 $\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i \, y_i \mathbf{x}_i$代入，则

$$f(\boldsymbol{x}) = \operatorname{sgn}\left(\sum_{i=1}^{N} \alpha_i \, y_i \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + b \right)$$

- 如何求解参数 b？

当数据不可分离时会发生什么?



Add a Slack Variable

ξ_i

$$\xi_i = \begin{cases} 0 & \text{if } \mathbf{x}_i \text{ correctly classified} \\ \text{distance to margin} & \text{otherwise} \end{cases}$$

软边界支持向量机:约束优化问题

- 训练样本：$(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_N, y_N)$
- 最小化 $\frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{N}\xi_i$

subject to: $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad \forall i = (1, \cdots, N)$

对偶问题(Non-separable data)

- Maximize

$$W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \, \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$$

- Subject to

$$0 \leq \alpha_i \leq C, \quad i = 1, \cdots, N$$

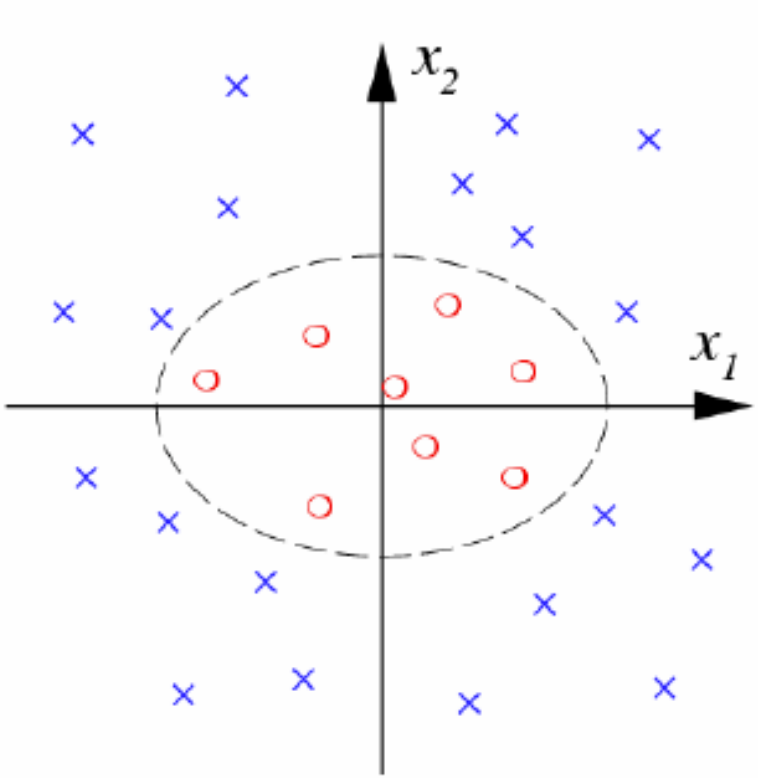
$$\sum_{i=1}^{N} \alpha_i \, y_i = 0$$

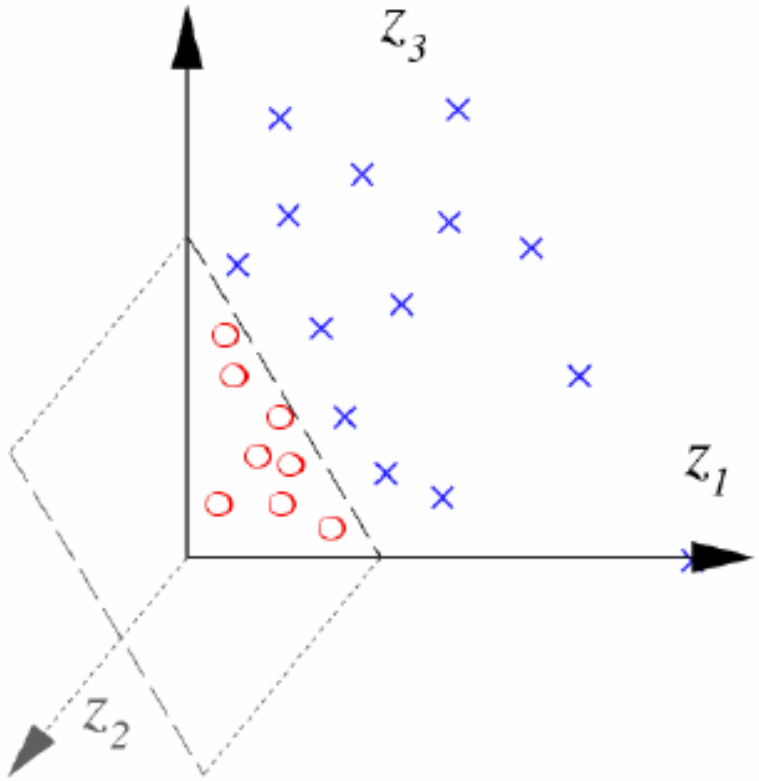
- Decision Boundary

$$f(\boldsymbol{x}) = sgn\left(\sum_{i=1}^{N} \alpha_i \, y_i \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle + b \right)$$

非线性映射后线性可分

$$\Phi : \mathbb{R}^2 \to \mathbb{R}^3$$
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2)$$



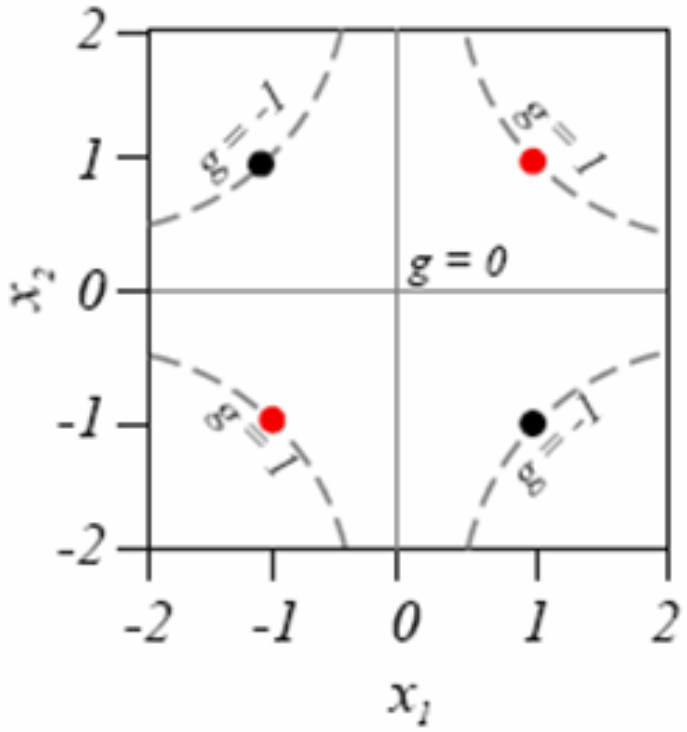


例

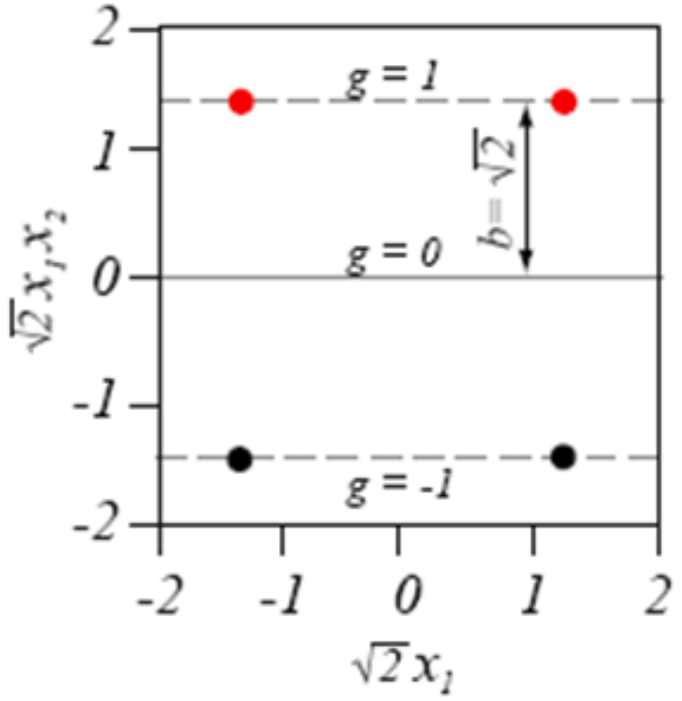
Consider the XOR problem which is non-linearly separable:

$(1,1)$ and $(-1,-1)$ belong to ω_1

$(1,1)$ and $(-1,1)$ belong to ω_2



例



Consider the following mapping (many other mapping could be used too):

$$\boldsymbol{z} = \Phi(\boldsymbol{x}) = \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1 \\ \sqrt{2}x_1x_2 \\ \sqrt{2}x_2 \\ x_2^2 \\ 1 \end{bmatrix}$$

例

- The above transformation maps $\boldsymbol{x}_k$ to a 6-dimensional space:

$$\boldsymbol{z}_1 = \Phi(\boldsymbol{x}_1) = \begin{bmatrix} 1 \\ \sqrt{2} \\ \sqrt{2} \\ \sqrt{2} \\ 1 \\ 1 \end{bmatrix} \quad \boldsymbol{z}_3 = \Phi(\boldsymbol{x}_3) = \begin{bmatrix} 1 \\ -\sqrt{2} \\ \sqrt{2} \\ -\sqrt{2} \\ 1 \\ 1 \end{bmatrix}$$

$$\boldsymbol{z}_2 = \Phi(\boldsymbol{x}_2) = \begin{bmatrix} 1 \\ \sqrt{2} \\ -\sqrt{2} \\ -\sqrt{2} \\ 1 \\ 1 \end{bmatrix} \quad \boldsymbol{z}_4 = \Phi(\boldsymbol{x}_4) = \begin{bmatrix} 1 \\ -\sqrt{2} \\ -\sqrt{2} \\ \sqrt{2} \\ 1 \\ 1 \end{bmatrix}$$

例

- We seek to maximize:

$$\sum_{k=1}^{4} \lambda_k - \frac{1}{2} \sum_{kj}^{4} \lambda_k \lambda_j y_k y_j \Phi(\boldsymbol{x}_j)^T \Phi(\boldsymbol{x}_k)$$

- subject to:

$$\sum_{k=1}^{4} y_k \lambda_k = 0, \ \lambda_k \geq 0, \quad k = 1,2,\cdots,4$$

- The solution turns out to be :

$$\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \frac{1}{8}$$

- Since all $\lambda_k \neq 0$,all $\boldsymbol{x}_k$ are support vectors!

例

- We can now compute $\boldsymbol{w}$:

$$\boldsymbol{w} = \sum_{k=1}^{4} y_k \lambda_k \Phi(\boldsymbol{x}_k) = \frac{1}{8}\begin{bmatrix}1\\ \sqrt{2}\\ \sqrt{2}\\ \sqrt{2}\\ 1\\ 1\end{bmatrix} - \frac{1}{8}\begin{bmatrix}1\\ \sqrt{2}\\ -\sqrt{2}\\ -\sqrt{2}\\ 1\\ 1\end{bmatrix} + \frac{1}{8}\begin{bmatrix}1\\ -\sqrt{2}\\ \sqrt{2}\\ -\sqrt{2}\\ 1\\ 1\end{bmatrix} - \frac{1}{8}\begin{bmatrix}1\\ -\sqrt{2}\\ -\sqrt{2}\\ \sqrt{2}\\ 1\\ 1\end{bmatrix} = \frac{1}{2}\begin{bmatrix}0\\ 0\\ \sqrt{2}\\ 0\\ 0\\ 0\end{bmatrix}$$

- The solution for b can be determined using any support vector, e.g, $\boldsymbol{x}_1$

$$\boldsymbol{w}^T \Phi(\boldsymbol{x}_1) + b = y_1 \text{ or } b = y_1 - \boldsymbol{w}^T \boldsymbol{x}_1 = 0$$

例

- The margin is computed as follows:

$$\frac{2}{||\boldsymbol{w}||} = 2\sqrt{2}$$

- The decision function is the following:

$$g(\boldsymbol{x}) = \boldsymbol{w}^T \Phi(\boldsymbol{x}) + b = x_1 x_2$$

- where we decide ω_1 if $g(\boldsymbol{x}) > 0$ and ω_2 if $g(\boldsymbol{x}) < 0$

非线性 SVMs

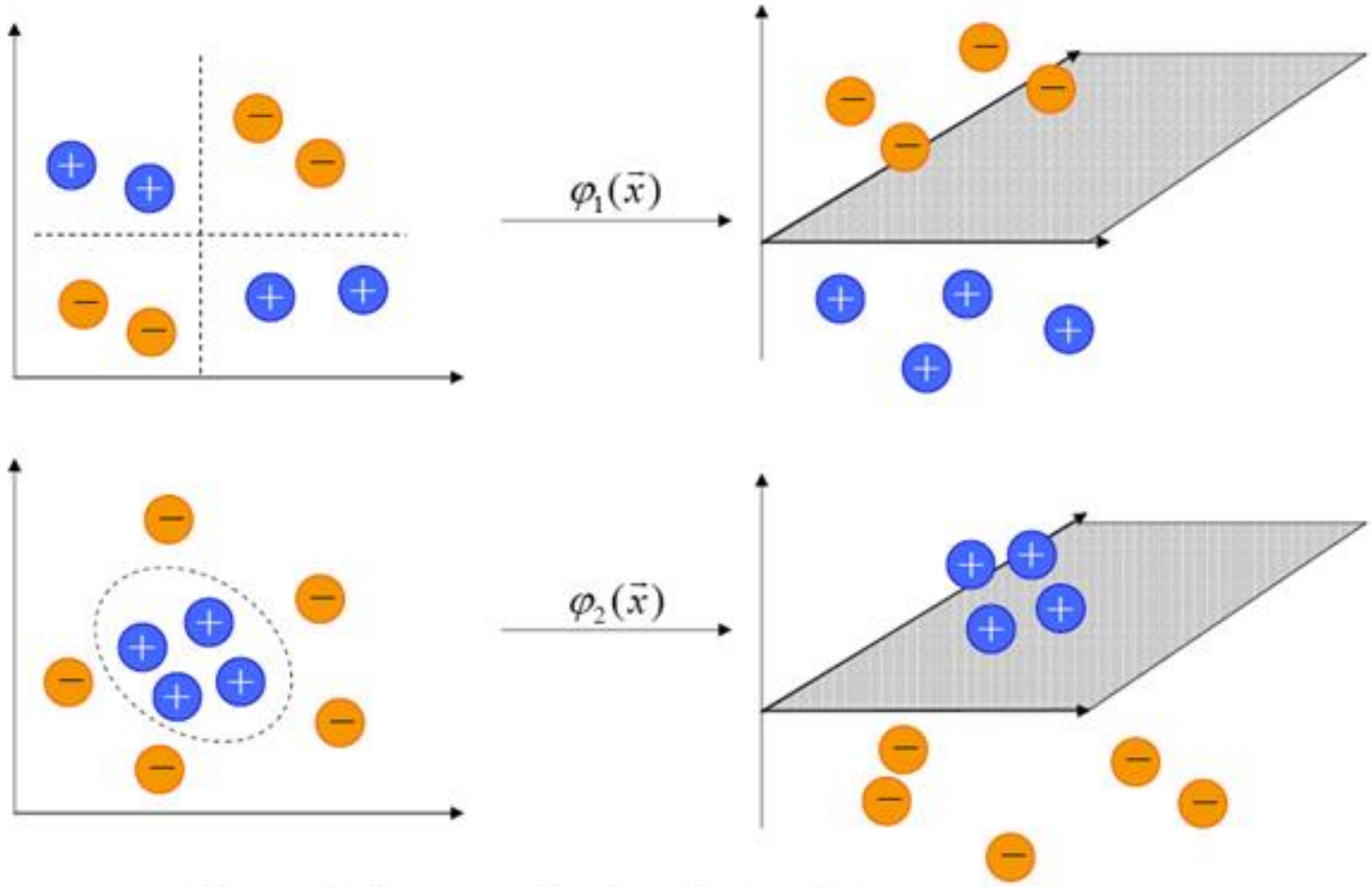
- 关键思想:注意决策边界和对偶优化公式都只依赖于输入空间中的点积!

$$\langle \boldsymbol{x}_i, \boldsymbol{x} \rangle$$

$$f(\boldsymbol{x}) = sgn(\sum_{i=1}^{N} \alpha_i \, y_i \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle + b)$$

$$W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \, \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$$

非线性 SVMs



Linearly Separable in Higher Dimension

核技巧

- Replace $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$ with

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle$$

- Maximize:

$$W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \, \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

- Boundary:

$$f(\boldsymbol{x}) = sgn(\sum_{i=1}^{N} \alpha_i \, y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b)$$

核矩阵的条件

$$\begin{aligned} K(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle \\ &= \langle \Phi(\boldsymbol{x}_j), \Phi(\boldsymbol{x}_i) \rangle \\ &= K(\boldsymbol{x}_j, \boldsymbol{x}_i) \end{aligned}$$

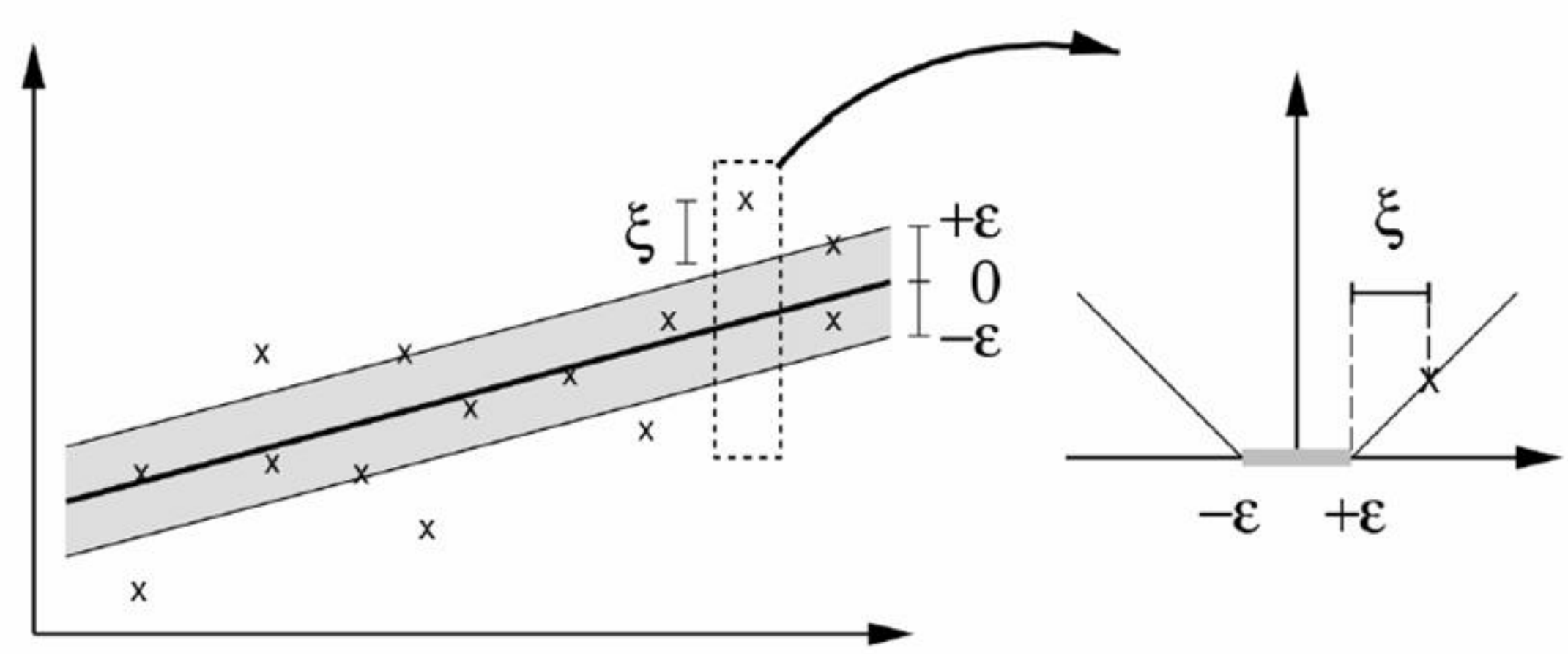
$$K = \begin{bmatrix} K(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & K(\boldsymbol{x}_1, \boldsymbol{x}_N) \\ \vdots & \ddots & \vdots \\ K(\boldsymbol{x}_N, \boldsymbol{x}_1) & \cdots & K(\boldsymbol{x}_N, \boldsymbol{x}_N) \end{bmatrix}$$

- 核矩阵 K 为 $N \times N$ 半正定矩阵

经常使用的核函数

- $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + c\right)^d$
- $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{1}{2\sigma^2}||\boldsymbol{x}_i - \boldsymbol{x}_j||^2\right)$

ε-SV 回归估计



优化问题的公式

- Estimate a linear regression

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$$

with precision ε by minimizing

- minimize

$$\tau(\mathbf{w}, \xi, \xi^*) = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*)$$

subject to

$$(\langle \mathbf{w}, \mathbf{x} \rangle + b) - y_i \leq \varepsilon + \xi_i$$
$$y_i - (\langle \mathbf{w}, \mathbf{x} \rangle + b) \leq \varepsilon + \xi_i^*$$

for all $i = 1, \cdots, m$.

关于核的对偶问题

- For $C > 0,\ \varepsilon \geq 0$ chosen a priori
- maximize

$$W(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*) = -\varepsilon \sum_{i=1}^{m} (\alpha_i^* + \alpha_i) + \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) y_i$$

$$-\frac{1}{2} \sum_{i,j=1}^{m} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) k(\mathbf{x}_i, \mathbf{x}_j)$$

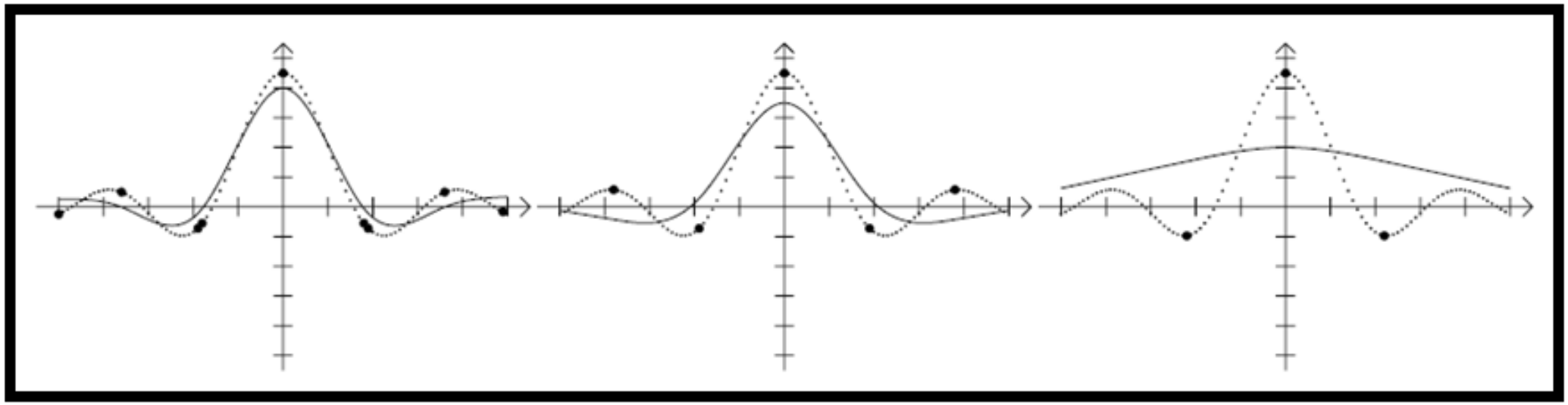
subject to

$$0 \leq \alpha_i, \alpha_i^* \leq C,\ i = 1, \cdots, m,\ \text{and} \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) = 0$$

- The regression estimate takes the form

$$f(\mathbf{x}) = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) k(\mathbf{x}_i, \mathbf{x}) + b$$

例



- Left to right: regression (solid line), data points (small dots) and SVs (big dots) for an approximation with $\varepsilon = 0.1, 0.2\ and\ \ 0.5$. Note the decrease in the number of SVs.

可选择的映射不是唯一的

- 例如： $\boldsymbol{x} \in R^2, \Phi(\boldsymbol{x}) = \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{bmatrix} \in R^3$ ， $K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^T\boldsymbol{y})^2$

$$(\boldsymbol{x}^T\boldsymbol{y})^2 = (x_1y_1 + x_2y_2)^2$$

$$\Phi(x) \cdot \Phi(y) = x_1^2y_1^2 + 2x_1y_1x_2y_2 + x_2^2y_2^2 = (x_1y_1 + x_2y_2)^2$$

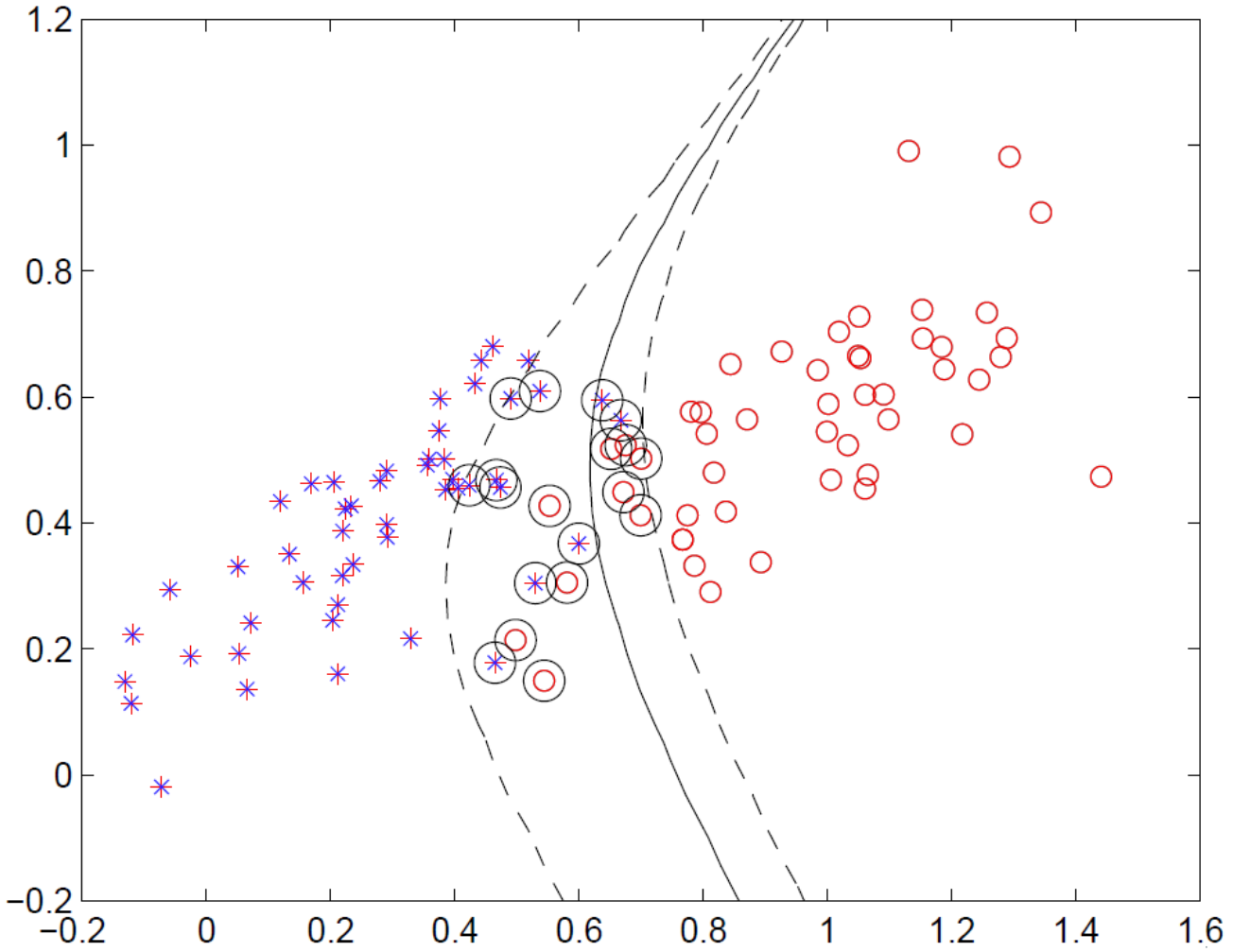
- 非线性映射 $\Phi(\cdot)$和高维特征空间都不是唯一的

$$\Phi(x) = \frac{1}{\sqrt{2}}\begin{bmatrix} (x_1^2 - x_2^2) \\ 2x_1x_2 \\ (x_1^2 + x_2^2) \end{bmatrix} \in R^3 \text{ or } \Phi(x) = \begin{bmatrix} x_1^2 \\ x_1x_2 \\ x_1x_2 \\ x_2^2 \end{bmatrix} \in R^4$$

支持向量机的优缺点

- 全局优化方法，没有局部最优(即基于精确优化，而不是近似方法)。
- 支持向量机的性能取决于核函数及其参数的选择
  - 对于给定的问题，核的最佳选择仍然是一个难题
- 它的复杂性取决于支持向量的数量，而不是转换空间的维数
- 利用小的训练集可以避免高维空间的过拟合和很好的泛化

MATLAB Code



谢谢各位同学！