安徽理工大学
ANHUI UNIVERSITY OF SCIENCE & TECHNOLOGY

- 人工智能专业 学科基础教育必修模块

2025

Python与机器学习

Python and Machine Learning

Chapter 10: Support Vector Machine

- Lecturer ： 孟亦凡
- E-mail：myf@aust.edu.cn

安徽理工大學
ANHUI UNIVERSITY OF SCIENCE & TECHNOLOGY

Chapter 10: Support Vector Machine

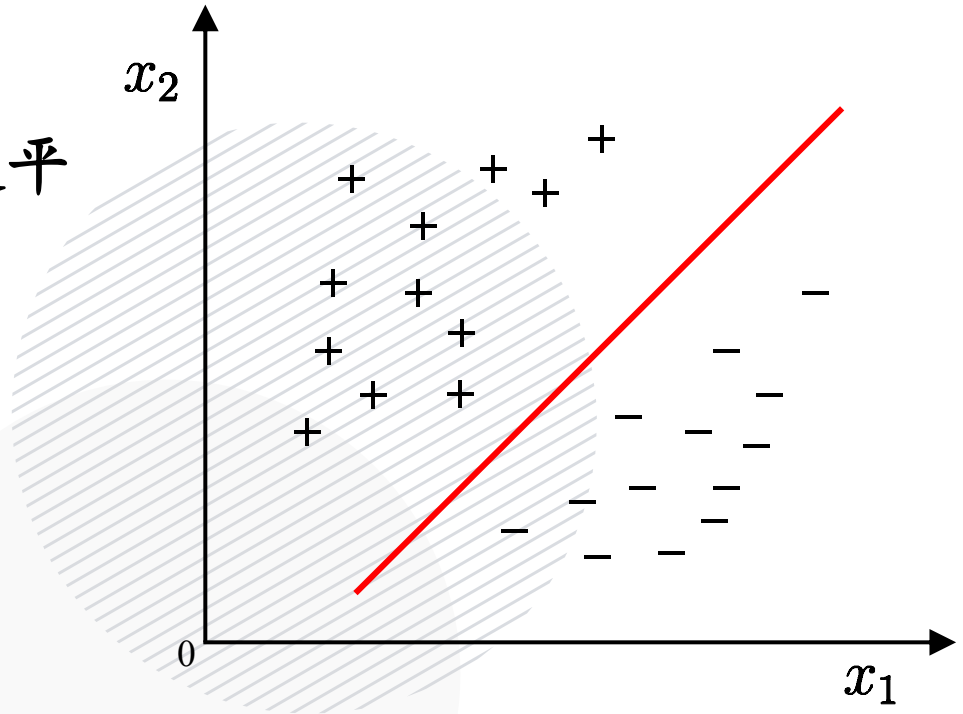
CONTENTS

10.1 Margins and Support Vectors
间隔与支持向量

Limitations of Linear Classifiers and Improvements
线性分类器的局限和改进

➢ Starting from Linear Classification: Which Hyperplane is Better?

- **Linear classifiers can find infinite separating hyperplanes. 线性分类器可以找到无数个分类超平面**
- **Key question: Which hyperplane has the best generalization ability? 哪个超平面最好?**

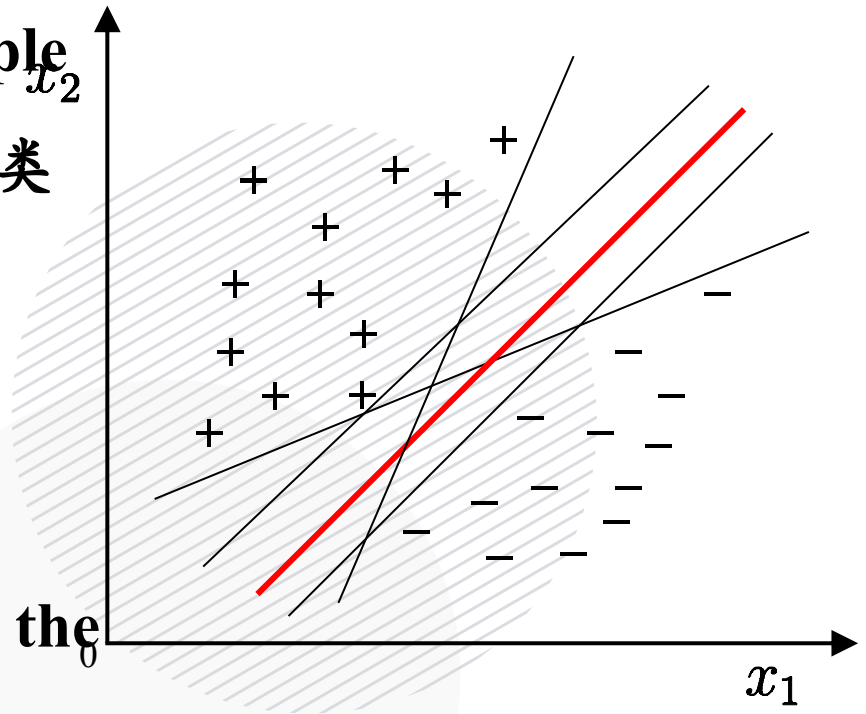


The most "central" hyperplane is least sensitive to perturbations and most robust.

最“居中”的超面对扰动最不敏感，最鲁棒

➢ How to Quantify "Centrality"?如何量化“居中”程度?

- **Margin（间隔）: Distance from the closest sample to the classification hyperplane.最近样本点到分类超平面的距离**
- **Maximum-margin hyperplane: The hyperplane with the largest minimum distance.**
- **Support vectors（支持向量）: Samples lying on the margin boundaries that determine the hyperplane.**
  位于间隔边界上的样本点，决定超平面位置



➢ Building a Mathematical Model for Margin Maximization建立间隔最大化的数学模型

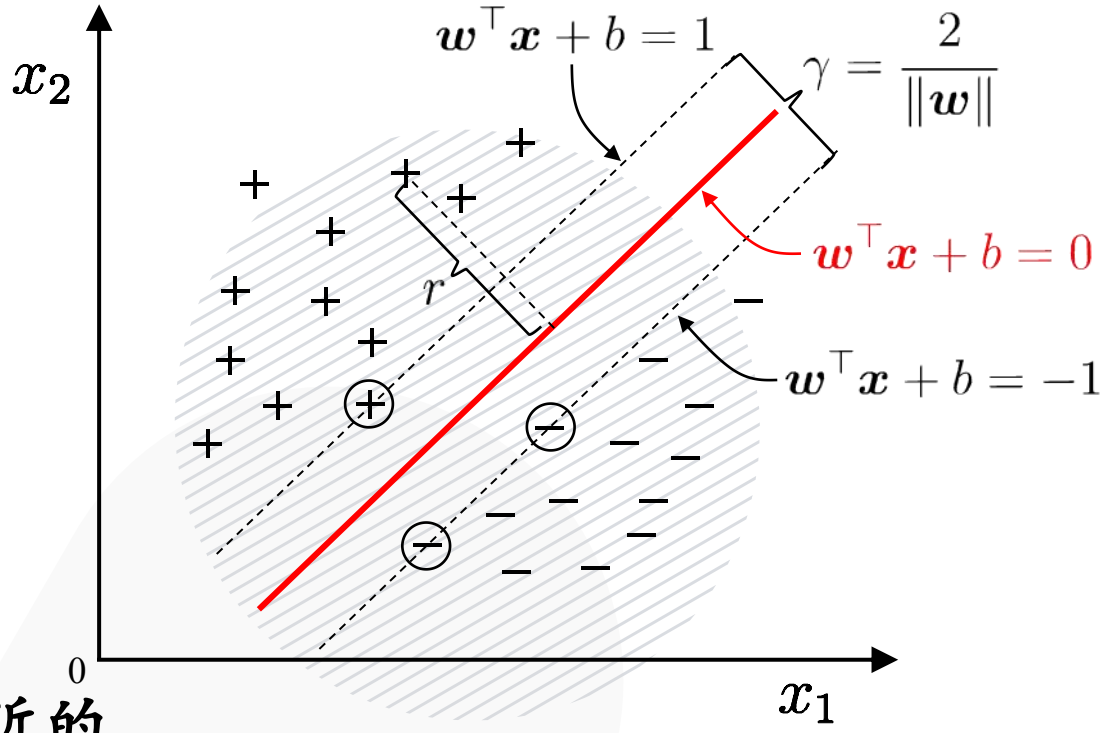
- **Hyperplane equation（超平面方程）：**

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

- **Distance from sample to hyperplane:**

$$\boldsymbol{\gamma_i = \frac{|w^T x + b|}{\|w\|}}$$

- 我们不仅要分类正确，还要让离超平面最近的样本点尽可能远。



正类样本: $\boldsymbol{y_i = +1}$

负类样本: $\boldsymbol{y_i = -1}$

➢ Building a Mathematical Model for Margin Maximization建立间隔最大化的数学模型

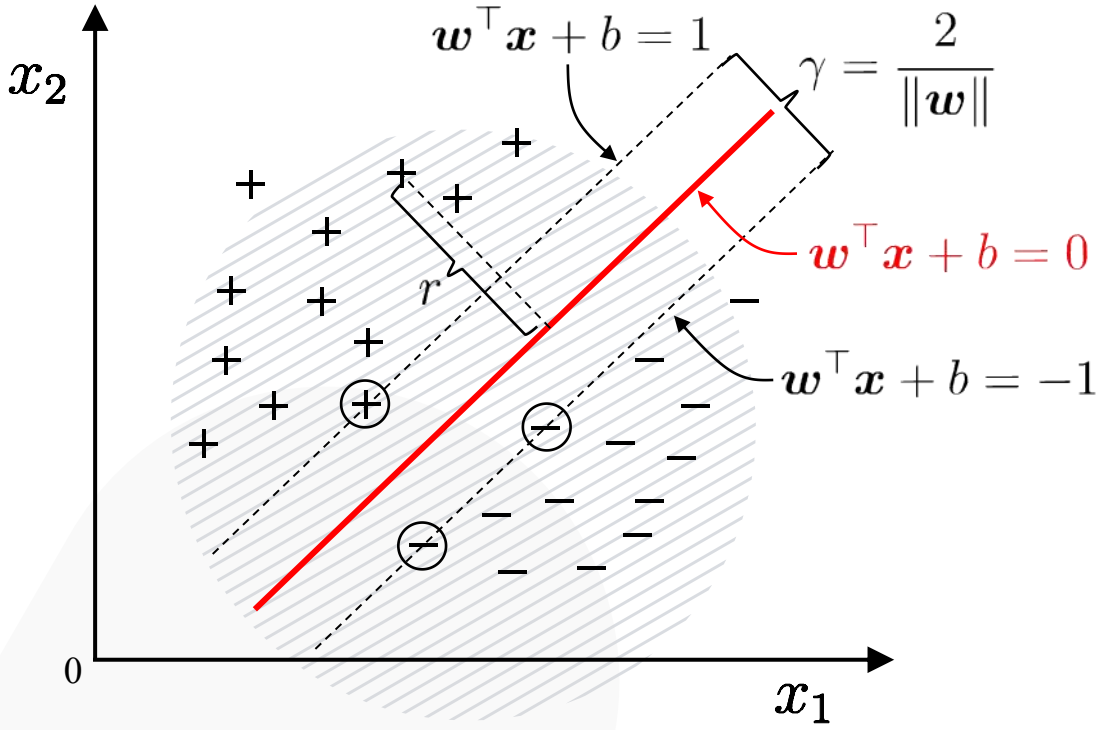
- 分类约束条件：

$$y_i(w^T x_i + b) > 0$$

- 当分类正确时，y_i和 $w^T x_i + b$ 同号，

  所以 $y_i(w^T x_i + b) > 0$

- 函数间隔的值越大，说明点离超平面越远



正类样本：$y_i = +1$

负类样本：$y_i = -1$

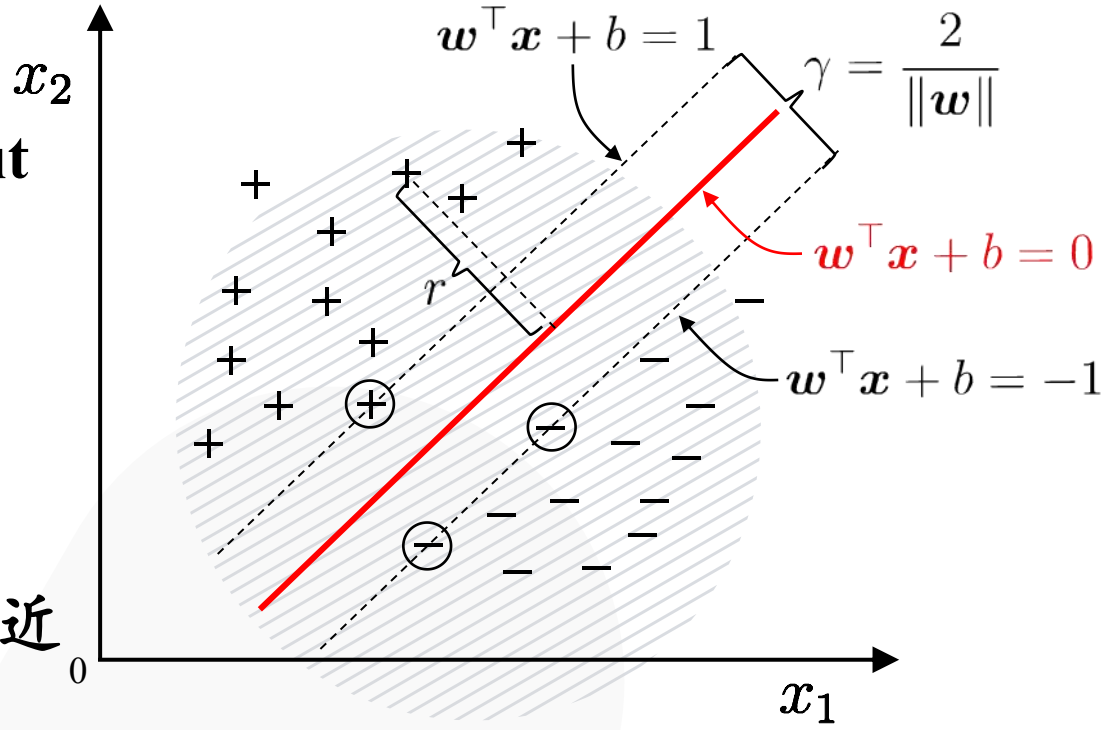
➢ Building a Mathematical Model for Margin Maximization建立间隔最大化的数学模型

- 归一化处理：w and b can be scaled without changing the hyperplane.

$$w^T x_i + b = 0$$

$$3w^T x_i + 3b = 0$$

- 调整 w 和 b 的尺度，使得对于离超平面最近的那个样本点，有：
  - $y_i(w^T x_i + b) = 1$
- 对于其他样本有：$y_i(w^T x_i + b) \geq 1$



正类样本：$y_i = +1$
负类样本：$y_i = -1$

➢ Building a Mathematical Model for Margin Maximization建立间隔最大化的数学模型

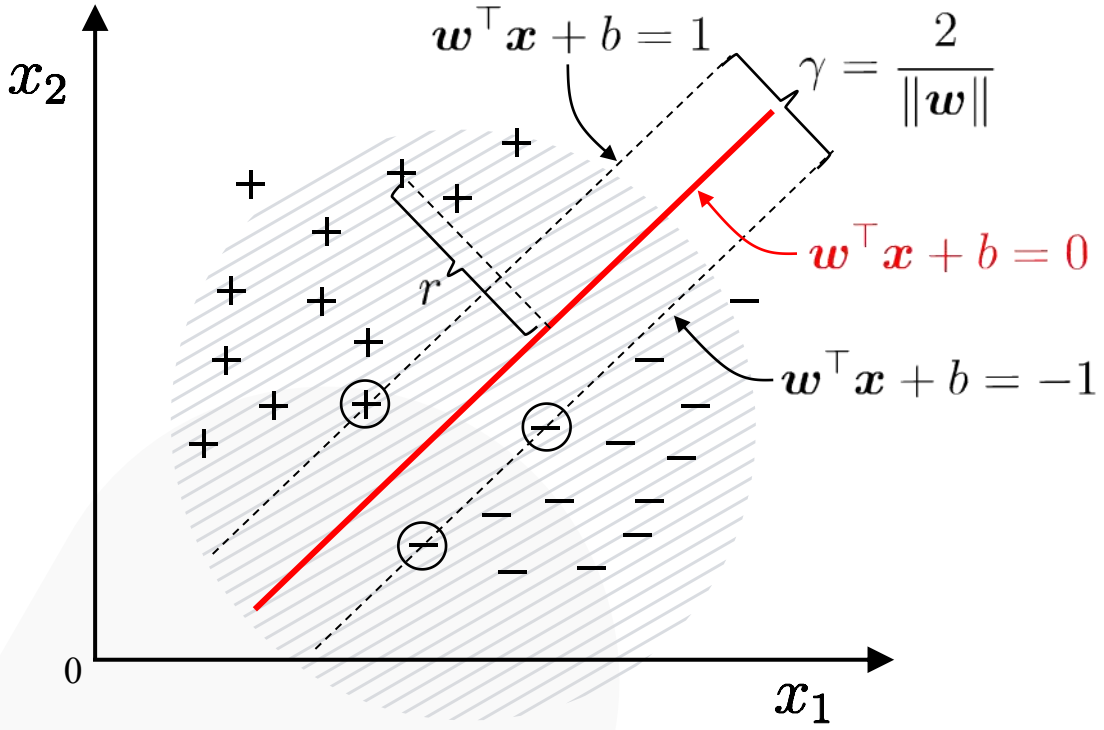
- **Optimization problem:**

$$\max_{w,b} \frac{1}{\|w\|} \text{ s.t. } y_i(w^T x_i + b) \geq 1$$

- **Equivalent form:**

$$\min_{w,b} \frac{1}{2}\|w\|^2 \text{ s.t. } y_i(w^T x_i + b) \geq 1$$

- 这是一个凸二次规划问题，存在唯一全局最优解



正类样本: $y_i = +1$
负类样本: $y_i = -1$

10.2 Dual Problem
对偶问题

➢ The Primal Optimization Problem

- **Target:**

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

- ***Constraints:***

$$y_i(w^T x_i + b) \geq 1$$

- 直接求解涉及高维变量 w 和大量约束

➢ 拉格朗日函数：将约束融入目标（将约束转化为惩罚项）

- 引入拉格朗日乘子$\alpha_i > 0$
- **Lagrangian function:**

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i [y_i(w^T x_i + b) - 1]$$

- 对 w 和 b 求偏导并令为零:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \quad \Rightarrow \quad w = \sum_{i=1}^{m} \alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{m} \alpha_i y_i = 0 \quad \Rightarrow \quad \sum_{i=1}^{m} \alpha_i y_i = 0$$

- 代入拉格朗日函数，消去 w 和 b

➢ Final Form of the Dual Problem

- **After Simplification, Obtain Standard Dual Problem**

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$\text{s.t.} \sum_{i=1}^{m} \alpha_i y_i = 0,$$

$$\alpha_i \geq 0, \quad i = 1, \ldots, m$$

- 这是一个凸二次规划问题，变量为$\boldsymbol{\alpha}$
- 约束数量等于样本数，但通常只有少量$\boldsymbol{\alpha_i > 0}$
- **Decision function:** $\boldsymbol{f(x) = w^T x_i + b = \sum_{i=1}^{m} \alpha_i y_i x^T x + b}$

➢ Final Form of the Dual Problem

- 最优解必须满足的条件 | Conditions that Optimal Solution Must Satisfy
- $f(x) = w^T x_i + b = \sum_{i=1}^{m} \alpha_i y_i x^T x + b$ （K-T-T条件）
  - 对偶约束：$\alpha_i > 0$
  - 原始约束：$y_i(w^T x_i + b) \geq 1$
  - 互补条件：$\alpha_i[y_i(w^T x_i + b) - 1]=0$
- 支持向量的判定：
  - 若 $\alpha_i > 0$，则对应样本为支持向量
  - 此时必有$y_i(w^T x_i + b) = 1$（位于间隔边界上）
  - 非支持向量对应 $\alpha_i = 0$，不影响模型

➢ Why is the Dual Problem Preferred?

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$\text{s.t.} \sum_{i=1}^{m} \alpha_i y_i = 0,$$

$$\alpha_i \geq 0, \quad i = 1, \ldots, m$$

- 计算简化：从优化w（d维）转为优化α（m维）
- 解的结构：$\sum_{i=1}^{m} \alpha_i y_i x_i$ → w是样本的线性组合
- 稀疏性：大多数αi=0，只有少数αi>0（支持向量）

10.3 Kernel Functions
核函数

➢ What Did the Dual Problem Leave Us?

- 对偶问题的核心形式

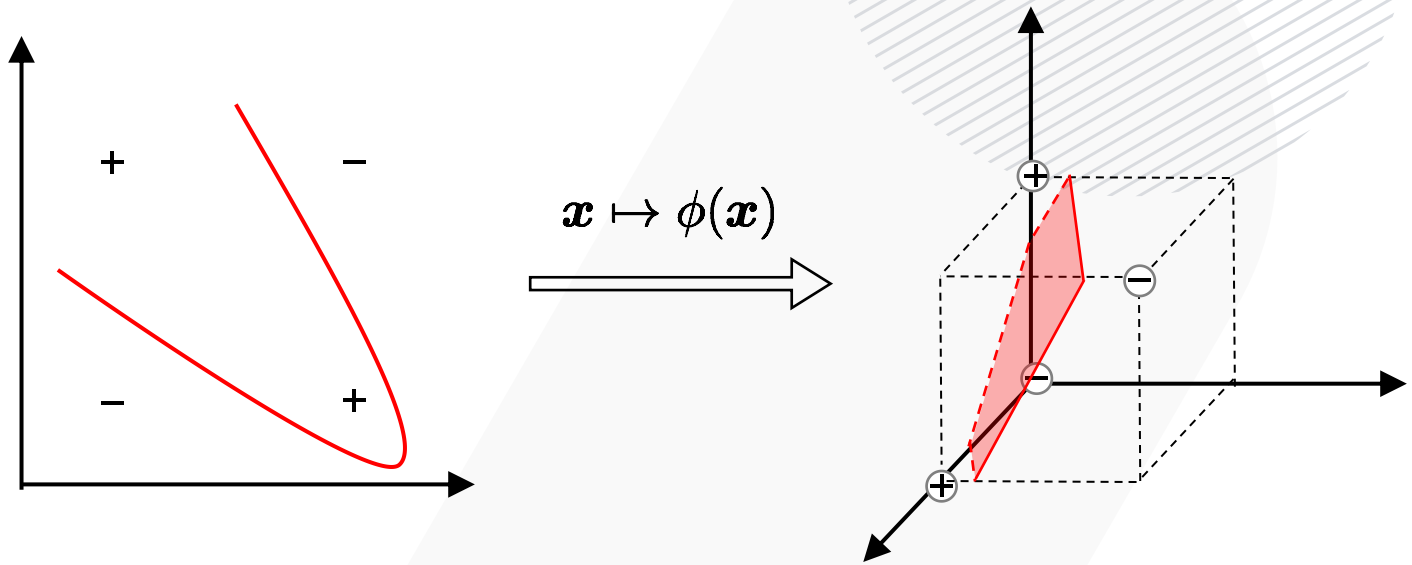
$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

  - 样本仅以内积 $x_i{}^T x_j$ 的形式出现
  - **Decision function:** $f(x) = w^T x_i + b = \sum_{i=1}^{m} \alpha_i y_i x^T x + b$

- **Practical Challenge: Linearly Inseparable Problems 线性不可分问题**
  - 许多真实数据集无法用直线/超平面完美分开
  - **Many real-world datasets cannot be perfectly separated by a line/hyperplane**

➢ Mapping to Higher Dimensions

- 在低维空间中非线性可分的数据，在高维空间中可能变得线性可分
  - 原始特征： $\boldsymbol{x} = (x_1, x_2)$
  - 映射到三维： $\boldsymbol{\Phi}(\boldsymbol{x}) = (x_1{}^2, x_2{}^2, \sqrt{2}x_1x_2)$
  - 在新的三维空间中，原本非线性可分的数据可能变得线性可分



➢ 直接映射的困境

- 假设映射函数：$\phi:\mathbb{R}^{d} \rightarrow \mathbb{R}^{D}$，其中D≫d
- 计算内积$\boldsymbol{\Phi}(\boldsymbol{x}_i)^{T}\boldsymbol{\Phi}(\boldsymbol{x}_j)$，需要 O(D) 时间，可能非常高
- 存储高维向量需要大量内存

- 能否间接的完成计算（在低维空间完成高维计算过程），避免实际操作高维向量?

➢ kernel function

- **We do not need to know the explicit form of the mapping function ϕ!**
- **我们只需要一个函数 K，能够直接计算高维空间中的内积：**
- **We only need a function K that can directly compute the inner product in the high-dimensional space:**

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle = \phi(x_i)^T \phi(x_j)$$

- **这个函数 K 就是核函数**

➢ 为什么核函数在低维空间的计算等价于高维空间的内积?

- 考虑一个二维空间中的点：$\boldsymbol{x} = (\boldsymbol{x_1}, \boldsymbol{x_2})$ $\boldsymbol{z} = (\boldsymbol{z_1}, \boldsymbol{z_2})$
- 使用多项式核（d=2）：

$$K(x, z) = (x \cdot z + 1)^2$$

- 展开计算:

$$\begin{aligned} K(x, z) &= (x_1 z_1 + x_2 z_2 + 1)^2 \\ &= x_1^2 z_1^2 + x_2^2 z_2^2 + 2 x_1 z_1 x_2 z_2 + 2 x_1 z_1 + 2 x_2 z_2 + 1 \end{aligned}$$

➢ 为什么核函数在低维空间的计算等价于高维空间的内积?

- 显式映射到高维空间
- 定义一个映射函数 φ，将二维点映射到六维空间：

$$\varphi(x) = (x_1^2,\ x_2^2,\ \sqrt{2}x_1x_2,\ \sqrt{2}x_1,\ \sqrt{2}x_2,\ 1)$$

- 在六维空间中计算内积：

$$\begin{aligned}\varphi(x)\cdot\varphi(z) &= x_1^2z_1^2 + x_2^2z_2^2 + 2x_1z_1x_2z_2 + 2x_1z_1 + 2x_2z_2 + 1 \\ &= K(x,z)\end{aligned}$$

➢ 为什么核函数在低维空间的计算等价于高维空间的内积?

- 在二维空间：我们只计算了$K(x,z)=(x\cdot z+1)^2$，这是简单的操作

- 在六维空间：我们显式计算了φ(x)和φ(z)，然后求内积

- 我们在低维空间（2维）做了一个简单的计算，却得到了与在高维空间（6维）计算内积相同的结果。

➢ What is kernel function?

- 对于函数K:X × X→R，如果存在一个从输入空间X到特征空间H的映射φ，使得对于所有x, z ∈ X:

$$K(x,z) = \langle \varphi(x), \varphi(z) \rangle_{\mathcal{H}}$$

- 其中 $\langle \cdot,\cdot \rangle_{\mathcal{H}}$ 是特征空间H中的内积，那么K就是一个核函数。

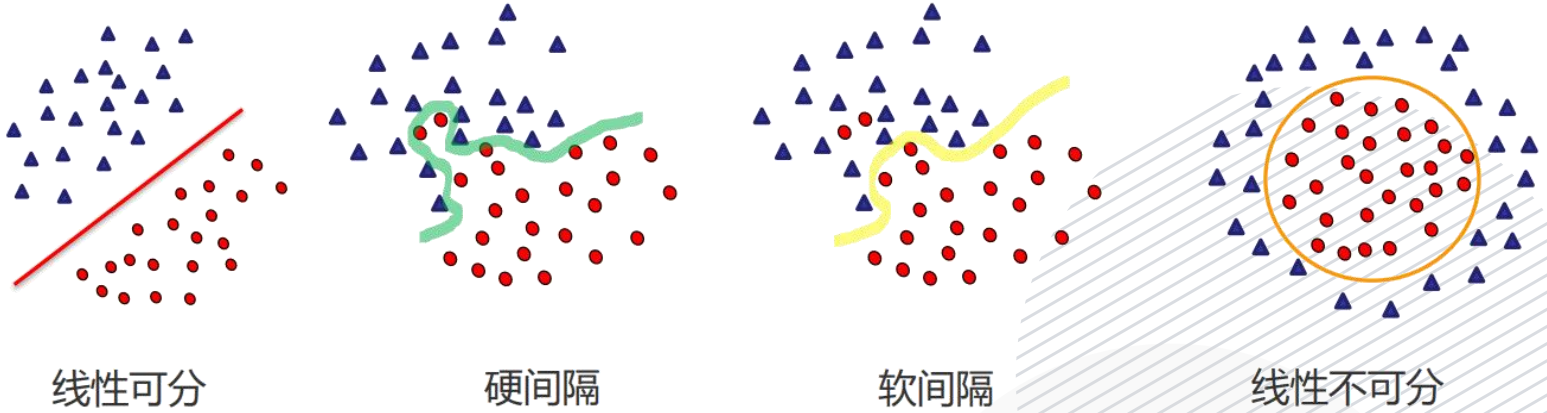
名称	表达式	参数
线性核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^\top \boldsymbol{x}_j$	
多项式核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^\top \boldsymbol{x}_j)^d$	$d \geq 1$为多项式的次数
高斯核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\delta^2}\right)$	$\delta > 0$为高斯核的带宽(width)
拉普拉斯核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|}{\delta}\right)$	$\delta > 0$
Sigmoid核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(\beta \boldsymbol{x}_i^\top \boldsymbol{x}_j + \theta)$	tanh为双曲正切函数, $\beta > 0, \theta < 0$

10.4 Soft Margins and Regularization
软间隔与正则化

➢ Limitations of Hard Margin

- **Hard margin（硬间隔） requires: All samples must be correctly classified and outside the margin boundaries.所有样本必须正确分类，且位于间隔边界之外**
- **现实数据常含有：**
  - **Noise（噪声）: Incorrect labels or measurement errors.错误的标签或测量误差**
  - **Outliers（异常值）: Data points significantly different from the majority.与大多数样本明显不同的数据点**
  - **Overlapping regions（重叠区域）: Samples from different classes partially overlap in feature space.不同类别的样本在特征空间中有部分重叠**

➢ From "Must Be Perfect" to "Allow Approximation"

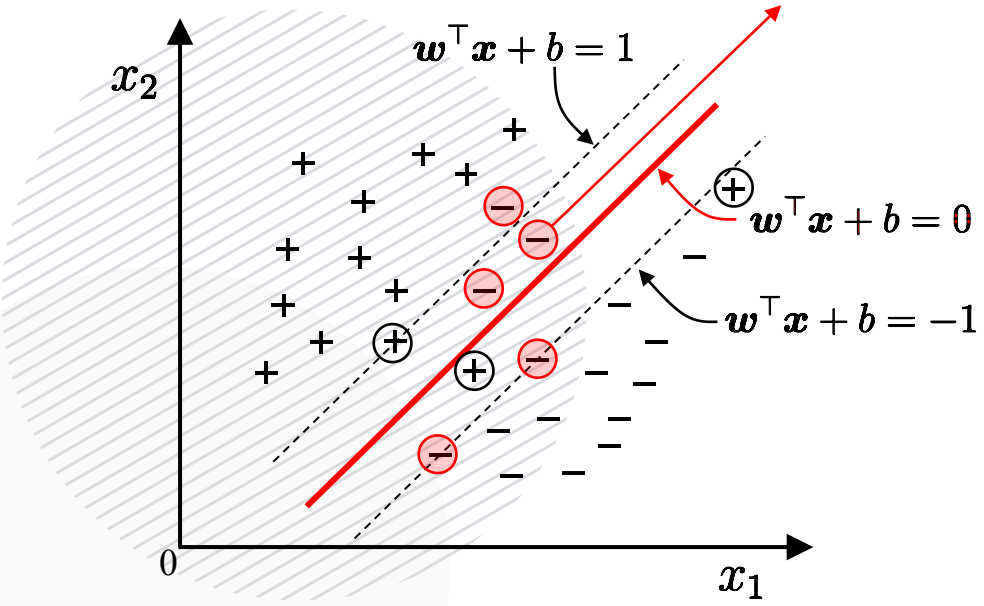


- 基本理念：为了获得更宽间隔、更鲁棒的模型，可以容忍一些样本被错分或落入间隔内
- 引入松弛变量 ξ_i 量化每个样本的违规程度

➢ 新目标函数：权衡间隔宽度与分类错误

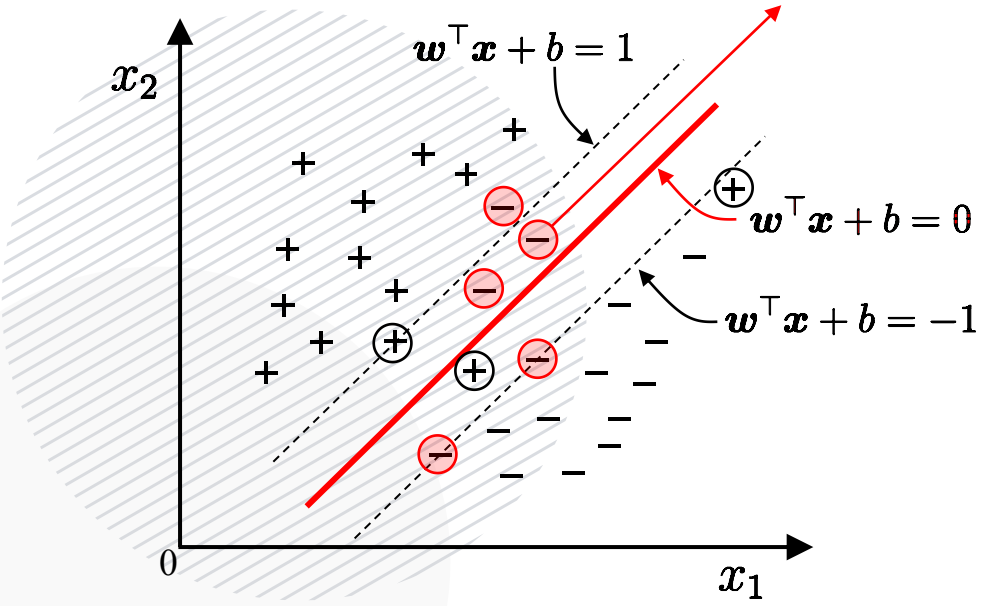
- 原始目标：最小化$\frac{1}{2}\|\boldsymbol{w}\|^2$（最大化间隔）
- 新增目标：最小化所有样本的违规程度总和
- 组合目标函数：

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m} l_{0/1}\left(y_i(\boldsymbol{w}^\top\phi(\boldsymbol{x}_i)+b)-1\right)$$



➢ 新目标函数：权衡间隔宽度与分类错误

- C > 0，称为正则化参数或惩罚参数
- 物理意义：
- C 很大：对错误容忍度低，接近硬间隔（可能过拟合）
- C 很小：对错误容忍度高，间隔可能很宽（可能欠拟合）



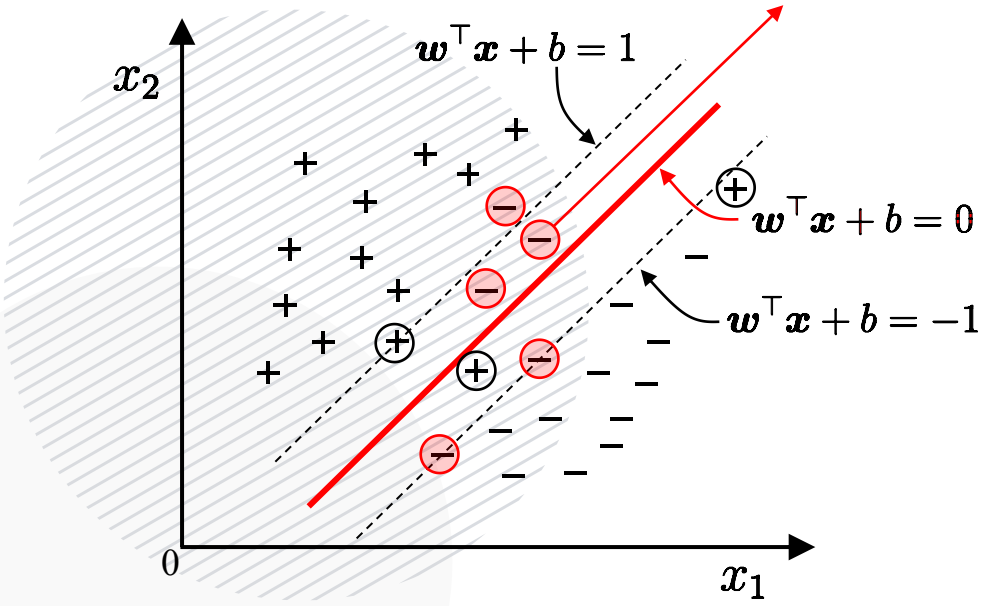
$$\min_{\boldsymbol{w},b} \ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m} l_{0/1}\left(y_i(\boldsymbol{w}^\top\phi(\boldsymbol{x}_i)+b)-1\right)$$

➢ New Form of the Dual Problem

$$\min_{\boldsymbol{w},b} \ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m} l_{0/1}\left(y_i(\boldsymbol{w}^\top\phi(\boldsymbol{x}_i)+b)-1\right)$$

Where $l_{0/1}$ is the "0/1 loss function"

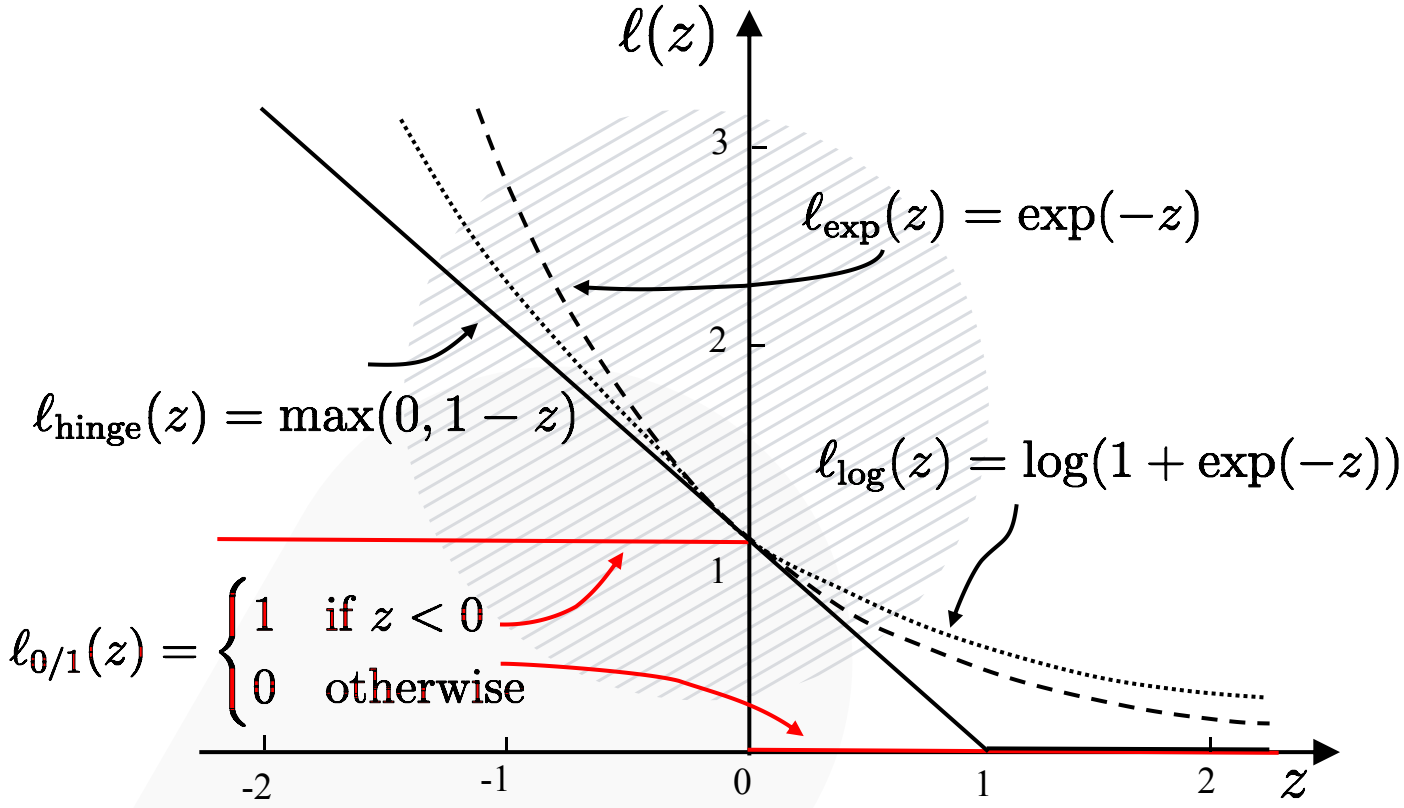
$$l_{0/1} = \begin{cases} 1 & z < 0 \\ 0 & otherwise \end{cases}$$



Existing problem: The 0/1 loss function is non-convex（非凸） and non-continuous, which is not easy to optimize!

➢ New Form of the Dual Problem

三种常见的替代损失函数 hinge 损失、指数损失、 对率损失



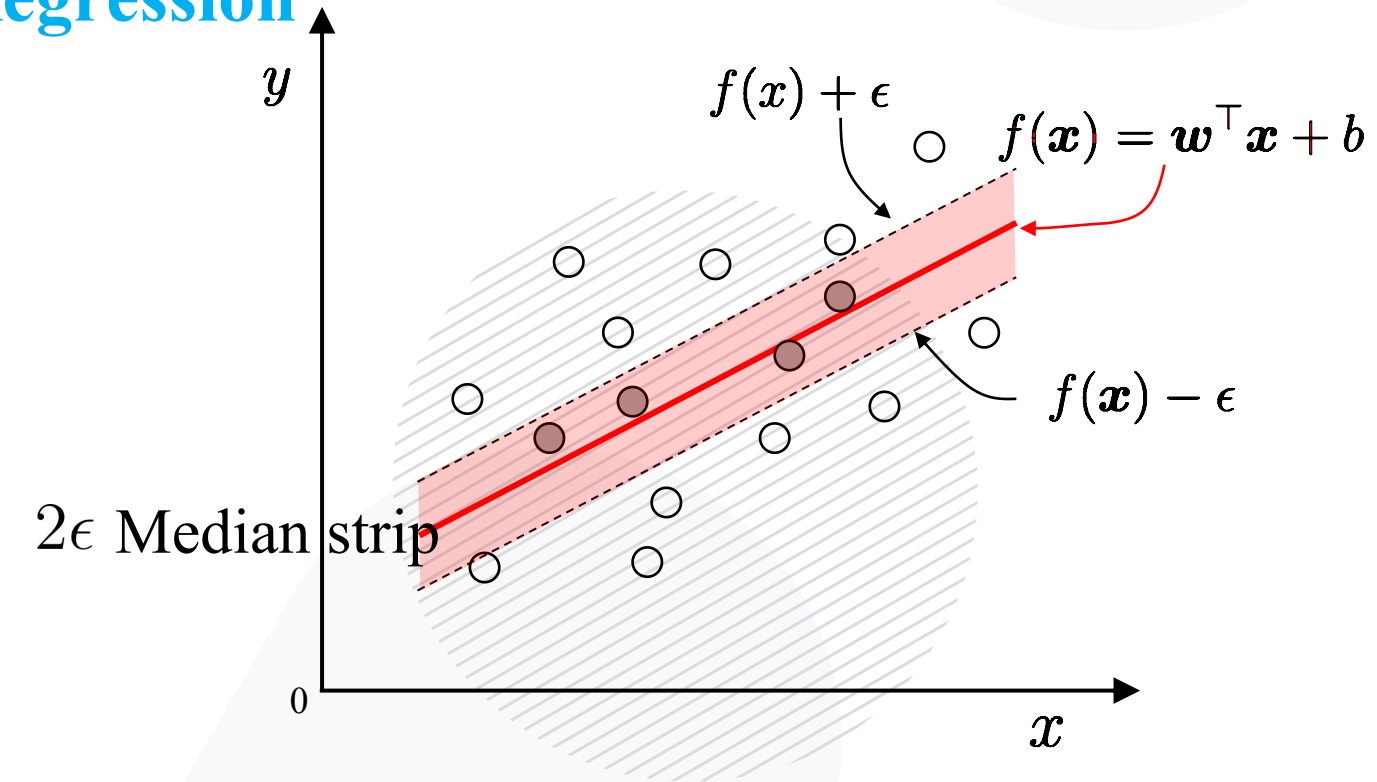
10.5 Support Vector Regression
支持向量回归

➢ Limitations of Support Vector Classification

- **SVM focuses on binary classification problems, outputting discrete class labels.SVM专注于二分类问题，输出为离散类别标签**
- **Real-world problems require predicting continuous values: price prediction, temperature forecasting, stock trends, etc.现实问题需要预测连续值：价格预测、温度预报、股票趋势等**

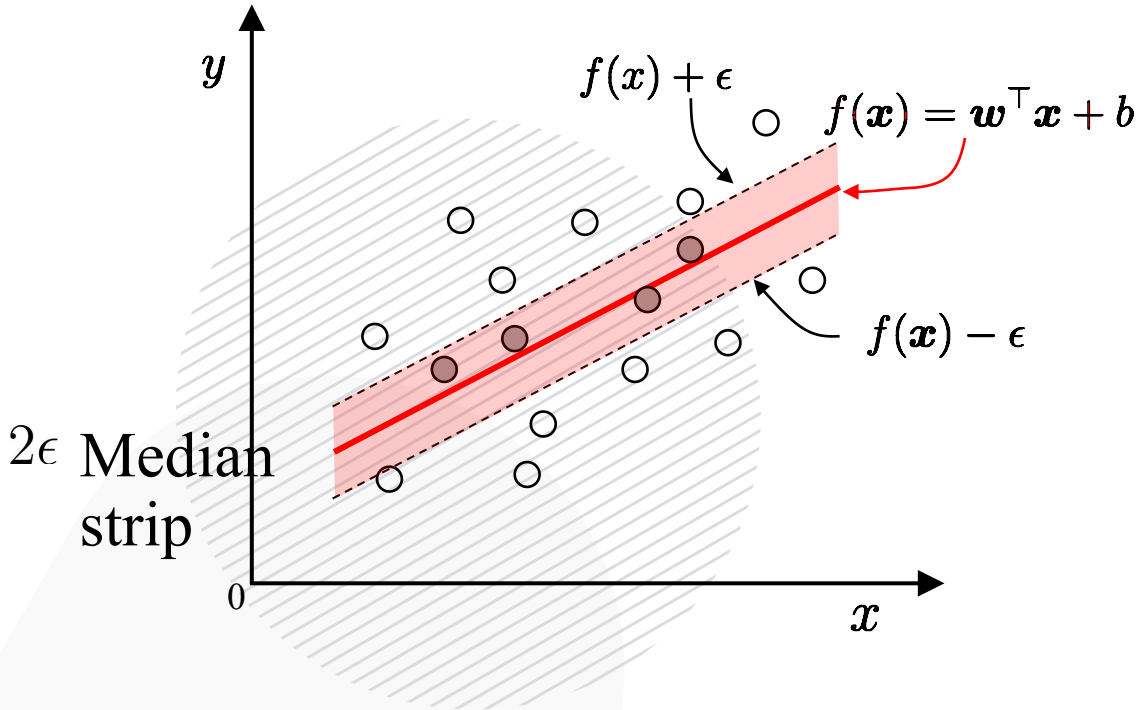
➢ Core Idea of Support Vector Regression

- 类似SVM的"间隔"概念，在回归中引入"间隔带"
- 目标：找到一个函数，使得大部分训练样本落在以该函数为中心、宽度为2ε的间隔带内



➢ Margin Band Concept in Support Vector Regression

- **ε-不敏感区域 | ε-Insensitive Region**
  - **定义：以回归函数f(x)为中心，上下各扩展ε形成的区域**
  - **落入该区域的样本被视为预测正确，不计损失**

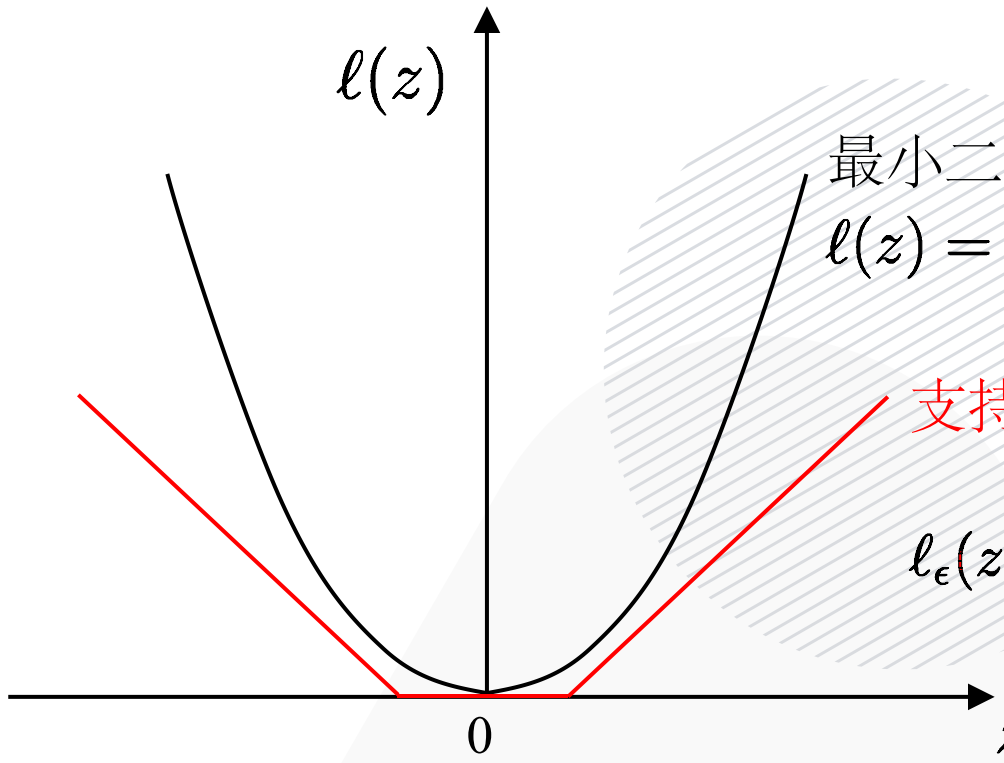


- **损失函数：**
  - **对于样本(x_i, y_i)，如果$|f(x_i) - y_i| \leq \varepsilon$，损失为0**
  - **如果$|f(x_i) - y_i| > \varepsilon$，损失为$|f(x_i) - y_i| - \varepsilon$**

➢ Comparison with Least Squares Regression

- 最小二乘：对所有偏差进行平方惩罚

- SVR：仅对超出ε的偏差进行线性惩罚



最小二乘损失函数

$$\ell(z) = z^2$$

支持向量回归损失函数

$$\ell_\epsilon(z) = \begin{cases} 0 & \text{if } |z| \leq \epsilon \\ |z| - \epsilon & \text{otherwise} \end{cases}$$

➢ SVR Optimization Problem

Original question

$$\min_{w,b,\xi_i,\hat{\xi}_i} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\left(\xi_i + \hat{\xi}_i\right)$$

$$\text{s.t. } f(\boldsymbol{x}_i) - y_i \leqslant \epsilon + \xi_i$$

$$y_i - f(\boldsymbol{x}_i) \leqslant \epsilon + \hat{\xi}_i$$

$$\xi_i \geqslant 0, \hat{\xi}_i \geqslant 0, i = 1, 2, \ldots, m$$

Dual problem

$$\max_{\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}}} \sum_{i=1}^{m} y_i\left(\hat{\alpha}_i - \alpha_i\right) - \epsilon\left(\hat{\alpha}_i + \alpha_i\right)$$

$$-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\hat{\alpha}_i - \alpha_i\right)\left(\hat{\alpha}_j - \alpha_j\right)\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j$$

$$\text{s.t. } \sum_{i=1}^{m}\left(\hat{\alpha}_i - \alpha_i\right) = 0$$

$$0 \leqslant \alpha_i, \hat{\alpha}_i \leqslant C$$

Predictions

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\top}\phi(\boldsymbol{x}) + b = \sum_{i=1}^{m}(\hat{\alpha}_i - \alpha_i)y_i\kappa(\boldsymbol{x}_i, \boldsymbol{x}) + b$$