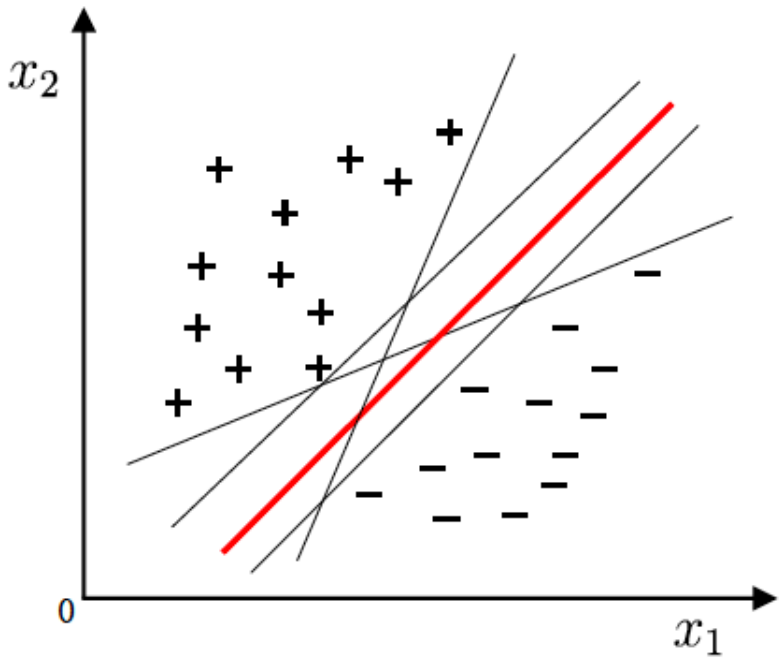
第四讲 支持向量机

高级机器学习

线性模型

将训练样本分开的超平面可能有很多，哪一个更好呢?



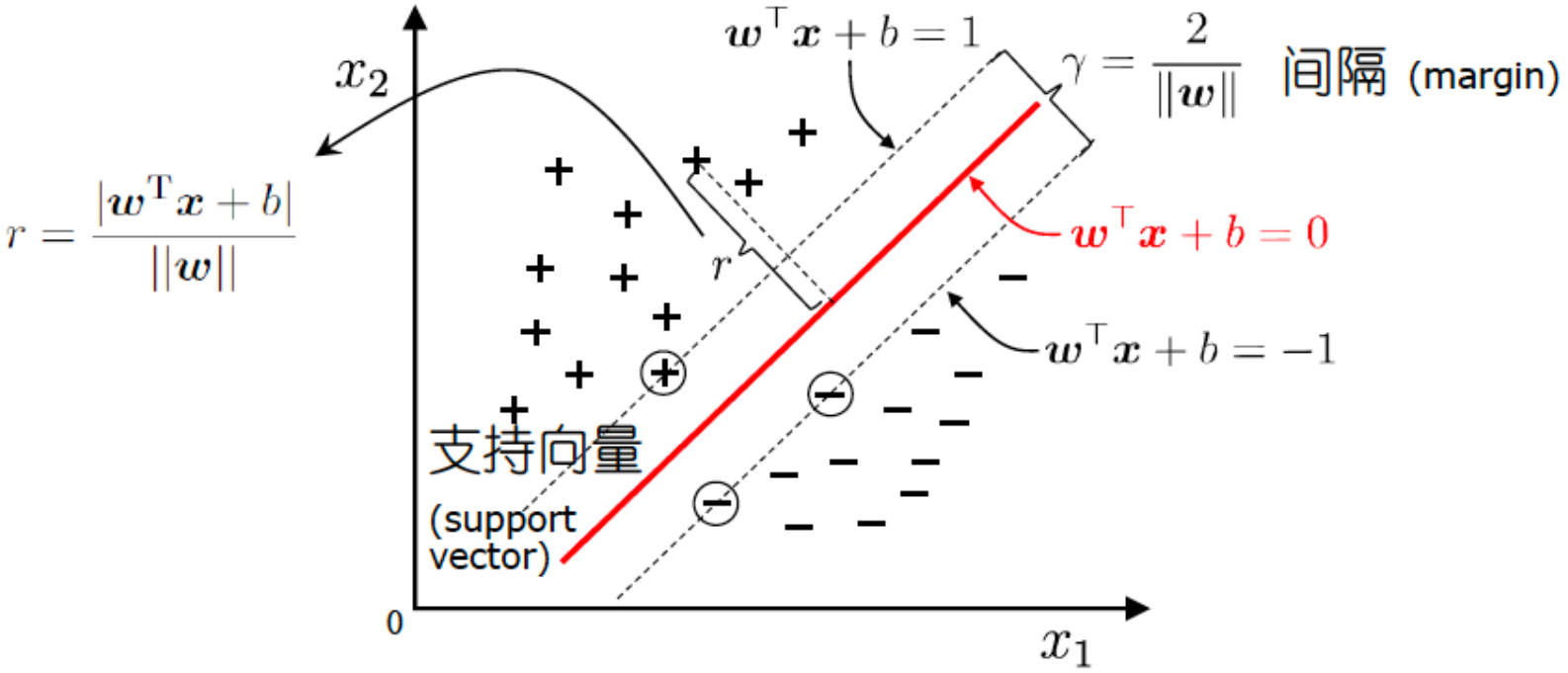
Vapnik的统计学习理论给出了答案：

"正中间"的：鲁棒性最好，泛化能力最强

间隔和支持向量

数据线性可分的情形

超平面方程：$\boldsymbol{w}^{\top}\boldsymbol{x}+b=0$



支持向量机 也被称为 “大间隔分类器”

支持向量机-基本型

最大间隔：寻找参数 $\boldsymbol{w}$ 和 b，使得 γ 最大

优化问题：

$$\begin{aligned} \arg\max_{\boldsymbol{w},b} \quad & \frac{2}{\|\boldsymbol{w}\|} \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1, \quad i = 1, 2, \ldots, m. \end{aligned}$$

⇩ 等价形式

$$\begin{aligned} \arg\min_{\boldsymbol{w},b} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1, \quad i = 1, 2, \ldots, m. \end{aligned}$$

凸二次规划问题

支持向量机-对偶型

- 引入拉格朗日乘子 $\alpha_i \geq 0$ 得到拉格朗日函数

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

- 令 $L(\boldsymbol{w}, b, \boldsymbol{\alpha})$ 对 $\boldsymbol{w}$ 和 b 的偏导为零可得

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

- 回代可得

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j \\ \text{s.t.} \quad & \sum_{i=1}^{m} \alpha_i y_i = 0 \ , \quad \alpha_i \geqslant 0 \ , \quad i = 1, 2, \ldots, m \end{aligned}$$

基本型和对偶型的比较

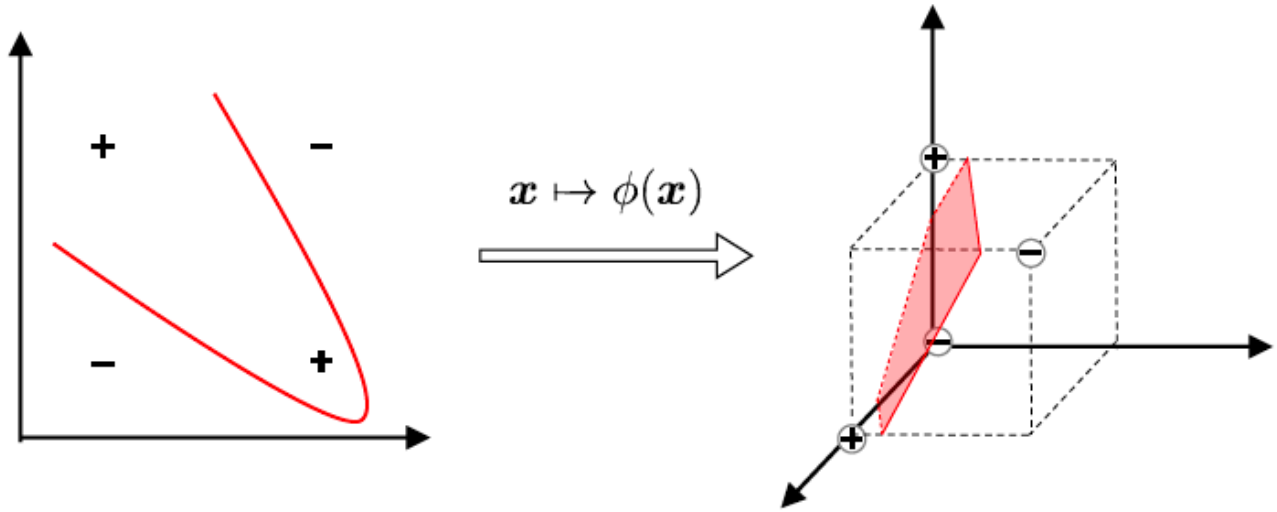
- 基本型和对偶型都是二次凸规划问题，根据凸优化理论，都可以利用成熟数值优化软件包得到全局最优解。
- 基本型和对偶型各有千秋，主要取决于效率和效果的需要
- 一般来说，基本型比较善于处理 1）低维数据；2）高维稀疏数据（所谓‘稀疏’指样本大部分属性的值为零）
- 对偶型比较善于处理高维稠密数据，此外对偶型容易吸收核函数处理非线性分类

特征空间映射

数据线性不可分的情形

若不存在一个能正确划分两类样本的超平面，怎么办?

将样本从原始空间映射到一个更高维的特征空间，使样本在这个特征空间内线性可分



如果原始空间是有限维(属性数有限)，那么一定存在一个高维特征空间使样本可分

在特征空间中

设样本 $\boldsymbol{x}$ 映射后的向量为 $\phi(\boldsymbol{x})$，划分超平面为 $f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}) + b$

原始问题

$$\begin{aligned} \min_{\boldsymbol{w},b} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i) + b) \geq 1, \quad i = 1, 2, \ldots, m. \end{aligned}$$

对偶问题

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \phi(\boldsymbol{x}_i)^{\mathrm{T}} \phi(\boldsymbol{x}_j) \\ \text{s.t.} \quad & \sum_{i=1}^{m} \alpha_i y_i = 0 \,, \quad \alpha_i \geqslant 0 \,, \quad i = 1, 2, \ldots, m \end{aligned}$$

预测

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}) + b = \sum_{i=1}^{m} \alpha_i y_i \phi(\boldsymbol{x}_i)^{\mathrm{T}} \phi(\boldsymbol{x}) + b$$

只以内积形式出现

核函数 (kernel function)

基本思路：设计核函数

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^{\mathrm{T}} \phi(\boldsymbol{x}_j)$$

绕过显式考虑特征映射、以及计算高维内积的困难

Mercer 定理：若一个对称函数所对应的核矩阵半正定，则它就能作为核函数来使用

“核函数选择”成为决定支持向量机性能的关键！

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, $\beta > 0, \theta < 0$

软间隔支持向量机

- 软间隔SVM是一类更为实用的模型
- 之前SVM都假定所有的样本要么线性可分，要么利用核函数达到线性可分。现实情况下，这两个“假定”都可能“理想化”，因此必须要思考样本不可分时，该如何抉择？软间隔SVM就是用于克服这类问题。
- 软间隔SVM不再假设所有样本都可分，而是引入损失函数，计算每个样本的“损失”，然后在最大化间隔和最小化整体损失之间做个合理的折衷。

理想型软间隔SVM

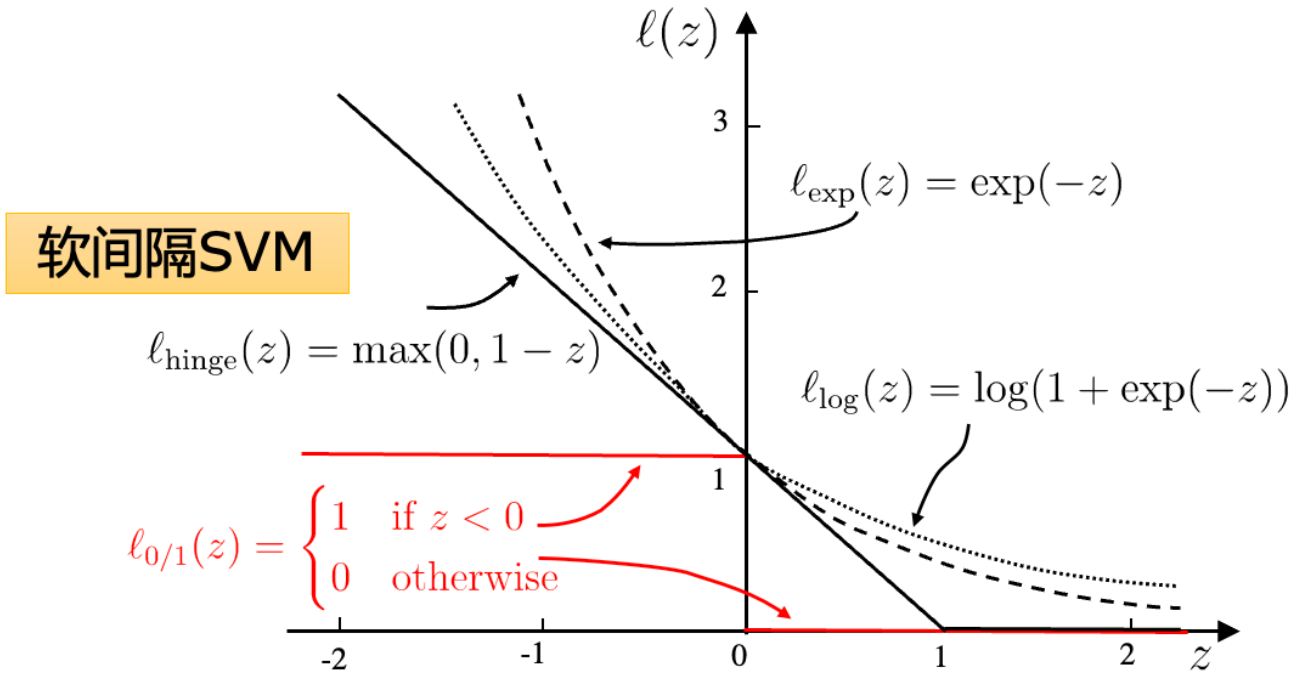
$$\min_{\boldsymbol{w},b} \ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m} l_{0/1}\left(y_i(\boldsymbol{w}^\top\phi(\boldsymbol{x}_i)+b)-1\right)$$

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

软间隔支持向量机

- 存在的问题：0/1损失函数非凸、非连续，不易优化！

- 替换方法：采用hinge损失函数



替代损失函数数学性质较好，一般是0/1损失函数的上界

软间隔支持向量机-基本型和对偶型

原始问题

$$\min_{\boldsymbol{w},b} \ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\max\left(0, 1 - y_i(\boldsymbol{w}$$

$$\max_{\boldsymbol{\alpha}} \ \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}$$
$$\text{s.t.} \ \sum_{i=1}^{m}\alpha_i y_i = 0 \ ,$$

对偶问题

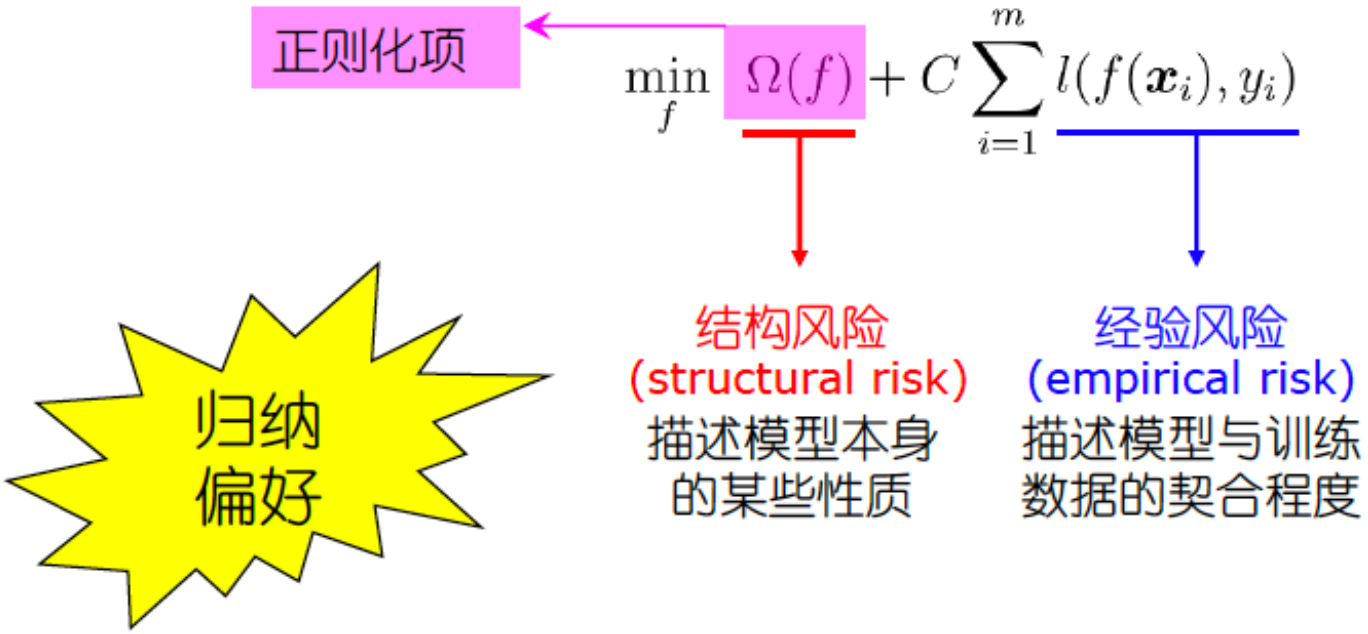
$$\min_{\boldsymbol{\alpha}} \ \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \phi(\boldsymbol{x}_i)^{\top}\phi(\boldsymbol{x}_j) - \sum_{i=1}^{m}\alpha_i$$

$$\text{s.t.} \ \sum_{i=1}^{m}\alpha_i y_i = 0, \ 0 \le \alpha_i \le C, \ i = 1, 2, \ldots, m.$$

- 仍是二次凸规划问题，根据凸优化理论，都可以利用成熟数值优化软件包得到全局最优解。

SVM拓展 - 正则化（regularization）

统计学习模型（例如 SVM）的更一般形式



- 正则化可理解为“罚函数法”
  通过对不希望的结果施以惩罚，使得优化过程趋向于希望目标
- 从贝叶斯估计的角度，则可认为是提供了模型的先验概率

小结

- 间隔和支持向量
- 支持向量机基本型
- 支持向量机对偶型
- 特征空间映射
- 核函数
- 软间隔支持向量机
- 正则化