

LECTURE 6: Support Vector Machine

1. 线性可分支持向量机

线性可分问题：可在特征空间中找到一个分离超平面 $w^\top x + b = 0$ ，将特征空间划分成正例和负例。通过分类决策函数 $f(x) = \text{sign}(w^\top x + b)$ 可以完美划分正负例。

1.1. 函数间隔与几何间隔

复习：空间中任意一个点 x_i 到平面的距离为：

$$\frac{|w^\top x_i + b|}{\|w\|}.$$

这里 $|w^\top x_i + b|$ 可以相对表示不同点到分离平面的距离。可以用 $y_i(w^\top x_i + b)$ 代表分类的正确性及确信度，这就是函数间隔。

Definition 1. (函数间隔) 对于给定的训练数据和超平面 (w, b) ，定义超平面关于样本点 (x_i, y_i) 的函数间隔为：

$$\hat{\gamma}_i = y_i(w^\top x_i + b).$$

超平面关于训练数据集的函数间隔为所有样本点的函数间隔最小值：

$$\hat{\gamma} = \min_i \hat{\gamma}_i.$$

这里如果将 w 和 b 进行同比例变化，则超平面不变，但是函数间隔却变化。可对 w 加约束 $\|w\| = 1$ ，此时的函数间隔对应的是几何间隔。定义样本点 (x_i, y_i) 与超平面的距离为

$$\gamma_i = y_i \left(\frac{w^\top}{\|w\|} x_i + \frac{b}{\|w\|} \right) \quad (1.1)$$

与样本数据集的几何间隔为

$$\gamma = \min_i \gamma_i.$$

1.2. 间隔最大化

间隔最大化：对训练数据找到几何间隔最大的超平面对训练数据进行分类。不仅将正负例分开，而且是将最难区分的正负例分开，几何间隔代表这种分类的“确信度”。

用数学语言叙述，可以转化为如下约束最优化问题：

$$\begin{aligned} \max_{w,b} \quad & \gamma, \\ \text{s.t.} \quad & y_i \left(\frac{w^\top}{\|w\|} x_i + \frac{b}{\|w\|} \right) \geq \gamma, \quad i = 1, \dots, N. \end{aligned}$$

考虑几何间隔和函数间隔的等价性，可将这个问题改写成：

$$\begin{aligned} \max_{w,b} \quad & \frac{\hat{\gamma}}{\|w\|} \\ \text{s.t.} \quad & y_i (w^\top x_i + b) \geq \hat{\gamma}, \quad i = 1, \dots, N \end{aligned}$$

函数间隔的取值不影响以上最优化问题的解（可以将 w 和 b 同比例变化，此时函数间隔也会同比例变化）。因此可以固定函数间隔 $\hat{\gamma} = 1$ ，代入以上最优化问题。则以上问题可转化为如下最优化问题，

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} \|w\|^2, \\ \text{s.t.} \quad & y_i (w^\top x_i + b) - 1 \geq 0, \quad i = 1, \dots, N \end{aligned}$$

因此，以上问题可以转化为一个凸二次规划问题。由此而得分离超平面，

$$f(x) = \text{sign}(w^* \cdot x + b^*).$$

支持向量（support vector）：在超平面 $w^\top x + b = 1$ & -1 上的点。

2. 学习的对偶算法

2.1. 拉格朗日对偶性

1. 原问题:

考虑有约束的最优化问题

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x), \\ \text{s.t.} \quad & c_i(x) \leq 0, \quad i = 1, \dots, k, \\ & h_j(x) = 0, \quad j = 1, \dots, l. \end{aligned}$$

引入拉格朗日函数,

$$L(x, \alpha, \beta) = f(x) + \sum_{i=1}^k \alpha_i c_i(x) + \sum_{j=1}^l \beta_j h_j(x).$$

这里 α_i 及 β_j 为拉格朗日乘子 ($\alpha_i \geq 0$). 考虑 x 的函数:

$$\theta_P(x) = \max_{\alpha, \beta: \alpha_i \geq 0} L(x, \alpha, \beta).$$

下标 P 表示这是原问题。

解读: 如果 x 违反了约束条件, 例如, $c_i(x) > 0$ 或者 $h_j(x) \neq 0$, 则可以使 $\alpha_i \rightarrow +\infty$ 以及 $\beta_j h_j(x) \rightarrow +\infty$, 从而 $\theta_P(x) = +\infty$. 如果 x 满足条件, 那么有 $\theta_P(x) = f(x)$. 因此我们有

$$\begin{cases} f(x) & c_i(x) \leq 0, \quad h_j(x) = 0 \\ +\infty & \text{others} \end{cases}$$

因此, 考虑极小化问题 $\min_x \theta_P(x)$, 则与原始问题是等价的。

2. 对偶问题

注意这里的原始问题为极小极大问题。对应的对偶问题则是极大极小问题：

$$\begin{aligned} \max_{\alpha, \beta} \min_x L(x, \alpha, \beta), \\ s.t. \quad \alpha_i \geq 0, \quad i = 1, \dots, k. \end{aligned}$$

3. 原问题和对偶问题的关系

(1) 若原始问题和对偶问题都有最优解，则对偶问题的最优解小于等于原问题的最优解：

$$\max_{\alpha, \beta: \alpha_i \geq 0} \min_x L(x, \alpha, \beta) \leq \min_x \max_{\alpha, \beta: \alpha_i \geq 0} L(x, \alpha, \beta). \quad (2.1)$$

(2) 在满足一定条件下，原始问题的解与对偶问题的解相同。

假设 $f(x)$ 与 $c_i(x)$ 是凸函数， $h_j(x)$ 是仿射函数，且不等式 $c_i(x)$ 是严格可行的（即存在 x 使得 $c_i(x) < 0$ ），那么 x^*, α^*, β^* 是原问题和对偶问题的最优解的充分必要条件是 KKT 条件：

$$\begin{aligned} \nabla_x L(x^*, \alpha^*, \beta^*) &= 0, \\ \alpha_i^* c_i(x^*) &= 0, \quad i = 1, \dots, k, \\ c_i(x^*) &\leq 0, \quad i = 1, \dots, k, \\ \alpha_i^* &\geq 0, \quad i = 1, \dots, k, \\ h_j(x^*) &= 0, \quad j = 1, \dots, l. \end{aligned}$$

由以上条件可知，若 $\alpha_i^* > 0$ ，则 $c_i(x) = 0$ 。

2.2. 支持向量机求解

对于支持向量机模型，它的拉格朗日函数如下：

$$L(w, \alpha, \beta) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^N \alpha_i y_i (w^\top x_i + b) + \sum_i \alpha_i.$$

它的对偶问题:

$$\max_{\alpha} \min_{w,b} L(w, b, \alpha).$$

(1) 求 $\min_{w,b} L(w, b, \alpha)$. 求偏导:

$$\begin{aligned}\nabla_w L(w, b, \alpha) &= w - \sum_{i=1}^N \alpha_i y_i x_i = 0, \\ \nabla_b L(w, b, \alpha) &= \sum_i \alpha_i y_i = 0\end{aligned}$$

代入拉格朗日函数(请自己演算一下), 即得:

$$L(w, b, \alpha) = -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) + \sum_i \alpha_i \quad (2.2)$$

则对偶问题可以表示为:

$$\begin{aligned}\max_{\alpha} \quad & -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^N \alpha_i, \\ \text{s.t.} \quad & \sum_{i=1}^N \alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad i = 1, \dots, N\end{aligned}$$

将最大转化为最小可得:

$$\begin{aligned}\min_{\alpha} \quad & \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^N \alpha_i, \\ \text{s.t.} \quad & \sum_{i=1}^N \alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad i = 1, \dots, N\end{aligned}$$

这里原始问题满足条件, 则原始问题的最优解也是对偶问题的最优解。假设 α^* ,

β^* 是对偶问题的最优解，那么KKT条件如下：

$$\nabla_w L(w^*, b^*, \alpha^*) = w^* - \sum_i \alpha_i^* y_i x_i = 0,$$

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

$$\alpha_i^* \{y_i(w^* \cdot x_i + b^*) - 1\} = 0,$$

$$y_i(w^* \cdot x_i + b^*) - 1 \geq 0, \quad i = 1, \dots, N,$$

$$\alpha_i^* \geq 0, \quad i = 1, \dots, N.$$

3. 线性支持向量机与软间隔最大化

3.1. 线性支持向量机

现实中的问题不一定能够完全线性可分。则可以引入一个松弛变量 $\xi_i \geq 0$ ，将约束条件变为：

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i \tag{3.1}$$

目标函数变为

$$\frac{1}{2} \|w\|^2 + C \sum_i \xi_i. \tag{3.2}$$

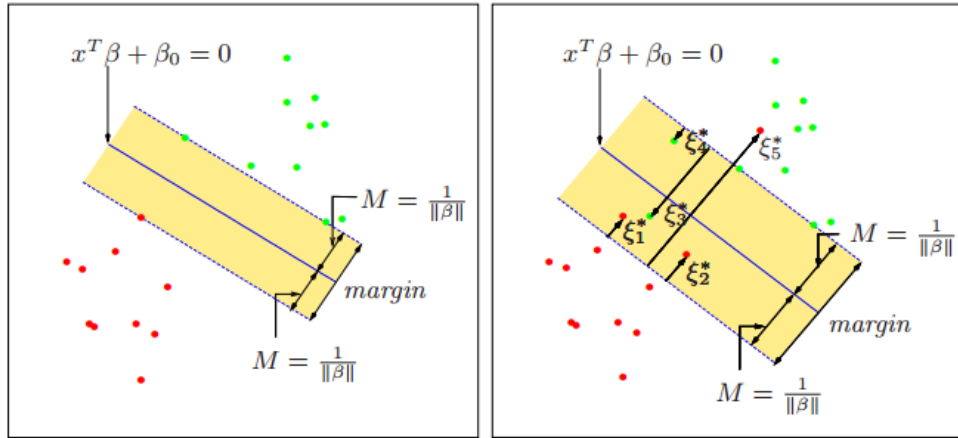


FIGURE 12.1. Support vector classifiers. The left panel shows the separable case. The decision boundary is the solid line, while broken lines bound the shaded maximal margin of width $2M = 2/\|\beta\|$. The right panel shows the nonseparable (overlap) case. The points labeled ξ_j^* are on the wrong side of their margin by an amount $\xi_j^* = M\xi_j$; points on the correct side have $\xi_j^* = 0$. The margin is maximized subject to a total budget $\sum \xi_i \leq \text{constant}$. Hence $\sum \xi_j^*$ is the total distance of points on the wrong side of their margin.

以上问题称为软间隔最大化问题，可以写成如下凸二次规划问题（原始问题）：

$$\begin{aligned} \min_{w, b, \xi} \quad & \frac{1}{2} \|w\|^2 + C \sum_i \xi_i, \\ \text{s.t.} \quad & y_i(w \cdot x_i + b) \geq 1 - \xi_i, i = 1, \dots, N, \\ & \xi_i \geq 0 \end{aligned}$$

3.2. 学习的对偶算法

以上问题的对偶问题是什么？

Step 1. 写出拉格朗日函数：

$$L(w, b, \xi, \alpha, \mu) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i - \sum_i \alpha_i \{y_i(w \cdot x_i + b) - 1 + \xi_i\} - \sum_i \mu_i \xi_i, \quad (3.3)$$

其中 $\alpha_i \geq 0, \mu_i \geq 0$. 转化为极大极小问题, 首先求解极小问题:

$$\nabla_w L(w, b, \xi, \alpha, \mu) = w - \sum_i \alpha_i y_i x_i = 0,$$

$$\nabla_b L(w, b, \xi, \alpha, \mu) = - \sum_i \alpha_i y_i = 0,$$

$$\nabla_{\xi_i} L(w, b, \xi, \alpha, \mu) = C - \alpha_i - \mu_i.$$

则可得:

$$\min_{w, b, \xi} L(w, b, \xi, \alpha, \mu) = ??? \quad (3.4)$$

则对偶问题为:

$$\begin{aligned} & \max_{\alpha} ??? \\ & s.t. \quad \sum_i \alpha_i y_i = 0, \\ & \quad C - \alpha_i - \mu_i = 0, \\ & \quad \alpha_i \geq 0, \quad \mu_i \geq 0. \end{aligned}$$

通过最后三个式子关系, 可得约束:

$$0 \leq \alpha_i \leq C \quad (3.5)$$

以上问题的KKT条件是什么？

$$\nabla_w L(w^*, b^*, \xi^*, \alpha^*, \mu^*) = w^* - \sum_i \alpha_i^* y_i x_i = 0,$$

$$\nabla_b L(w^*, b^*, \xi^*, \alpha^*, \mu^*) = - \sum_i \alpha_i^* y_i = 0,$$

$$\nabla_\xi L(w^*, b^*, \xi^*, \alpha^*, \mu^*) = C - \alpha^* - \mu^* = 0,$$

$$\alpha_i^* \{y_i(w^* \cdot x_i + b^*) - 1 + \xi_i^*\} = 0,$$

$$\mu_i^* \xi_i^* = 0,$$

$$y_i(w^* \cdot x_i + b^*) - 1 + \xi_i^* \geq 0,$$

$$\xi_i^* \geq 0, \quad \alpha_i^* \geq 0, \quad \mu_i^* \geq 0, \quad i = 1, \dots, N.$$

3.3. 合页损失函数(Hinge Loss Function)

线性支持向量机等价于最小化以下目标函数：

$$\sum_i [1 - y_i(w \cdot x_i + b)]_+ + \lambda \|w\|^2. \quad (3.6)$$

其中 $[x]_+ = \max(x, 0)$.

如下图所示，合页损失函数是0-1损失的上界，由于0-1损失不可导，直接优化比较困难，可以认为合页损失函数是一种代理损失函数。

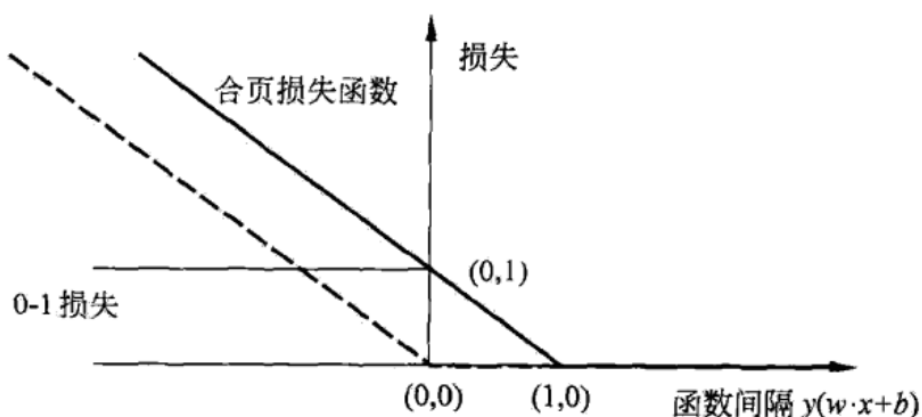


图 7.6 合页损失函数

4. 非线性支持向量机与核函数

4.1 核技巧

1. 非线性分类问题

如果能用 \mathbb{R}^n 中的一个超曲面将正负例分开，则称这个问题为非线性可分问题。

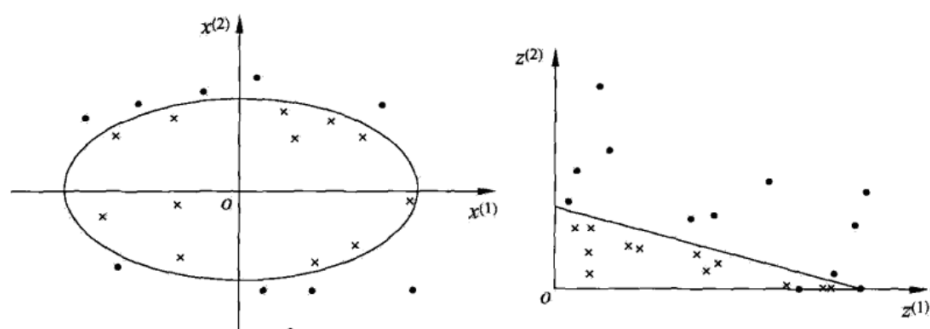


图 7.7 非线性分类问题与核技巧示例

如上图所示，可以将非线性变为线性分类问题。设原空间为 $\mathcal{X} \in \mathbb{R}^2$, $x = (x^{(1)}, x^{(2)})$; 新空间为 $\mathcal{Z} \in \mathbb{R}^2$, $z = (z^{(1)}, z^{(2)}) \in \mathcal{Z}$. 定义从原空间到新空间的变换：
 $z = \phi(x) = (x^{(1)2}, x^{(2)2})^\top$.

核技巧：基本想法是通过非线性变换，将输入空间对应于一个特征空间（希尔伯特空间 \mathcal{H} ），使得在输入空间 \mathbb{R}^n 中的超曲面模型对应于特征空间 \mathcal{H} 中的超平面模型，在新的空间中求解线性支持向量机就可以完成。

2. 核函数定义

Definition 2. (核函数) 设 \mathcal{X} 是输入空间（欧式空间 \mathbb{R}^n 的子集或离散集合），又设 \mathcal{H} 为特征空间（希尔伯特空间），如果存在一个从 \mathcal{X} 到 \mathcal{H} 的映射：

$$\phi(x) : \mathcal{X} \rightarrow \mathcal{H}$$

使得对于所有的 $x, z \in \mathcal{X}$, 函数 $K(x, z)$ 满足条件:

$$K(x, z) = \phi(x) \cdot \phi(z)$$

则称 $K(x, z)$ 为核函数, $\phi(x)$ 为映射函数, $\phi(x) \cdot \phi(z)$ 为 $\phi(x)$ 和 $\phi(z)$ 的内积。

3. 核技巧在支持向量机中的应用

在对偶问题的目标函数中, 内积 $x_i \cdot x_j$ 可以用核函数 $K(x_i, x_j)$ 代替, 此时对偶问题的目标函数为

$$W(\alpha) = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^N \alpha_i \quad (4.1)$$

同时, 分类决策函数中的内积也用核函数代替:

$$f(x) = \text{sign} \left(\sum_i \alpha_i^* y_i \phi(x_i) \cdot \phi(x) + b^* \right) = \text{sign} \left(\sum_i \alpha_i^* y_i K(x_i, x) + b^* \right) \quad (4.2)$$

因此, 不需要显式定义特征空间和映射函数。

4.2 正定核

1. 正定核充要条件

函数 $K(x, z)$ 满足什么条件才能成为核函数? 通常所说的核函数就是正定核函数 (positive definite kernel function)。

Theorem 1. (正定核的充要条件) 设 $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ 是对称函数, 则 $K(x, z)$ 为正定核函数的充要条件是对任意的 $x_i \in \mathcal{X}$ ($i = 1, \dots, m$), $K(x, z)$ 对应的 *Gram* 矩阵: $K = [K(x_i, x_j)]_{m \times m}$ 是半正定矩阵。

以上充要条件也可以成为正定核的等价定义。

2. 常用核函数

多项式核函数(polynomial kernel function):

$$K(x, z) = (x \cdot z + 1)^p \quad (4.3)$$

高斯核函数(Gaussian kernel function):

$$K(x, z) = \exp \left(- \frac{\|x - z\|^2}{2\sigma^2} \right) \quad (4.4)$$