Week 3: SVM 算法基本原理

滕明卓

2021 年 4 月 2 日

1 SVM 基础思想

间隔最大化。

2 优化问题 (原问题)

从简单的二类线性可分的情况下，推导出 SVM 的优化问题形式。

点 $\boldsymbol{x_i}$ 到超平面 $\boldsymbol{w}^T\boldsymbol{x}+b=0$ 的距离为

$$\frac{|\boldsymbol{w}^T\boldsymbol{x_i}+b|}{\|\boldsymbol{w}\|}$$

同时缩放 w 和 b，不改变这个距离。

假设超平面 $\boldsymbol{w}^T\boldsymbol{x}+b=0$ 可将数据样本分开。即对于正类样本 (标记 $y=+1$) 和负类样本 (标记 $y=-1$)，有

$$\begin{cases} \boldsymbol{w}^T\boldsymbol{x_i}+b>0, y_i=1 \\ \boldsymbol{w}^T\boldsymbol{x_i}+b<0, y_i=-1 \end{cases} \tag{1}$$

同时缩放 w 和 b, 使得

$$\begin{cases} \boldsymbol{w}^T\boldsymbol{x_i}+b\geq 1, y_i=1 \\ \boldsymbol{w}^T\boldsymbol{x_i}+b\leq -1, y_i=-1 \end{cases} \tag{2}$$

这样，正负样本点到超平面的距离之和就是 $\frac{2}{\|\boldsymbol{w}\|}$，这个距离被称为" 间隔"。

因此得到优化问题:

$$
\begin{aligned}
&\max_{\boldsymbol{w},b} \frac{2}{\|\boldsymbol{w}\|} \\
&\text{s.t. } y_i\left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right) \geq 1, i = 1, 2, \ldots, m.
\end{aligned} \tag{3}
$$

转化为等价形式：

$$
\begin{aligned}
&\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \\
&\text{s.t. } y_i\left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right) \geq 1, i = 1, 2, \ldots, m.
\end{aligned} \tag{4}
$$

这个优化问题就是 SVM 的原问题，这是一个凸优化问题。

3 对偶

使用拉格朗日乘子法得到 SVM 的对偶问题。原问题满足 slater 条件，满足强对偶性，即对偶问题的最优值与原问题最优质相等。因此可以通过求解对偶问题来找到最优值，求最优解。

拉格朗日函数：

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

令其对优化变量 $\boldsymbol{w}, b$ 求导等于 0，得：

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

回代到拉格朗日函数中，消除 $\boldsymbol{w}, b$，得到对偶问题:

$$
\begin{aligned}
&\min_{\boldsymbol{\alpha}} \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j - \sum_{i=1}^{m} \alpha_i \\
&\text{s.t. } \sum_{i=1}^{m} \alpha_i y_i = 0 \\
&\qquad \alpha_i \geq 0, i = 1, 2, \ldots, m.
\end{aligned} \tag{5}
$$

需满足 KKT 条件:

$$\left\{\begin{array}{l}\alpha_i \geq 0 \\ y_i f\left(\boldsymbol{x}_i\right) \geq 1 \\ \alpha_i\left(y_i f\left(\boldsymbol{x}_i\right)-1\right)=0 \\ \boldsymbol{w}=\sum_{i=1}^m \alpha_i y_i \boldsymbol{x}_i \\ \sum_{i=1}^m \alpha_i y_i=0\end{array}\right. \tag{6}$$

KKT 条件其中第三条被称为互补松驰性，如果拉格朗日乘子 α_i 不为 0，那么必有 $y_i f\left(\boldsymbol{x}_i\right)=1$，也就是这个样本点在间隔的边界上。这些样本被称为“支持向量”。

4 SVM 算法过程

- 输入：线性可分训练数据集 $T=\left\{\left(x_1, y_1\right),\left(x_2, y_2\right), \cdots,\left(x_m, y_m\right)\right\}$，其中，$x_i \in \mathcal{X}=R^n, y_i \in \mathcal{Y}=\{+1,-1\}, i=1,2, \cdots, m$
- 输出：最大间隔分离超平面和分类决策函数.

1. 构建并求解约束最优化问题

$$\begin{aligned} & \min_{\boldsymbol{\alpha}} \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\top} \boldsymbol{x}_j-\sum_{i=1}^m \alpha_i \\ & \text{s.t.} \quad \sum_{i=1}^m \alpha_i y_i=0 \\ & \alpha_i \geq 0, i=1,2, \ldots, m \end{aligned} \tag{7}$$

求得最优解 $\boldsymbol{\alpha}^*=\left(\alpha_1^*, \alpha_2^*, \ldots, \alpha_m^*\right)^T$

2. 计算

$$\boldsymbol{w}^*=\sum_{i=1}^m \alpha_i^* y_i \boldsymbol{x}_i \tag{8}$$

选择 α^* 的一个分量 $\alpha_j^*>0$，计算

$$b^*=y_j-\boldsymbol{w}^{*\top} \boldsymbol{x}_j \tag{9}$$

现实任务中常采用一种更鲁棒的做法：使用所有支持向量求解平均值。

3. 得到分离超平面

$$\boldsymbol{w}^{*\top}\boldsymbol{x}+b^{*}=0 \tag{10}$$

分类决策函数

$$f(x)=\operatorname{sign}\left(\boldsymbol{w}^{*\top}\boldsymbol{x}+b^{*}\right) \tag{11}$$

5 SMO 算法求解对偶问题

1. 选取一对需要更新的变量 α_i,α_j

2. 固定 α_i,α_j 以外的参数，求解对偶问题，更新 α_i,α_j

仅考虑 α_i 和 α_j 时对偶问题的约束变为

$$\alpha_i y_i+\alpha_j y_j=-\sum_{k\neq i,j}\alpha_k y_k,\quad \alpha_i\geq 0,\quad \alpha_j\geq 0$$

用一个变量表示另一个变量, 回代入对偶问题可得一个单变量的二次规划, 该问题具有闭式解.

6 线性到非线性-核技巧

一些分类问题无法用线性超平面划分，比如异或问题。如何解决？将样本从原始空间映射到更高维的特征空间，使得样本在这个特征空间内线性可分。

$$\boldsymbol{x}\rightarrow\phi(\boldsymbol{x})$$

模型表示为

$$f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x})+b$$

按之前的思路，得到对偶问题：

$$\begin{aligned}
&\min_{\boldsymbol{\alpha}}\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j\phi\left(\boldsymbol{x}_i\right)^{\top}\phi\left(\boldsymbol{x}_j\right)-\sum_{i=1}^{m}\alpha_i\\
&\text{s.t. }\sum_{i=1}^{m}\alpha_i y_i=0\\
&\qquad \alpha_i\geq 0, i=1,2,\ldots,m.
\end{aligned} \tag{12}$$

有时计算 $\phi\left(\boldsymbol{x}_i\right)^{\top}\phi\left(\boldsymbol{x}_j\right)$ 是很困难的，甚至不存在显式的映射。

使用核函数

$$\kappa\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)=\phi\left(\boldsymbol{x}_i\right)^{\top} \phi\left(\boldsymbol{x}_j\right)$$

来代替 $\phi\left(\boldsymbol{x}_i\right)^{\top} \phi\left(\boldsymbol{x}_j\right)$

Mercer 定理：只要一个对称函数所对应的它就能作核矩阵半正定, 为核函数来使用

对偶问题重写为：

$$\begin{aligned}
& \min_{\boldsymbol{\alpha}} \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \kappa\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)-\sum_{i=1}^m \alpha_i \\
& \text{s.t. } \sum_{i=1}^m \alpha_i y_i=0 \\
& \qquad \alpha_i \geq 0, i=1,2, \ldots, m .
\end{aligned} \tag{13}$$

预测：

$$\begin{aligned}
f(\boldsymbol{x}) & =\boldsymbol{w}^{\mathrm{T}} \phi(\boldsymbol{x})+b \\
& =\sum_{i=1}^m \alpha_i y_i \phi\left(\boldsymbol{x}_i\right)^{\mathrm{T}} \phi(\boldsymbol{x})+b \\
& =\sum_{i=1}^m \alpha_i y_i \kappa\left(\boldsymbol{x}, \boldsymbol{x}_i\right)+b
\end{aligned} \tag{14}$$

常用核函数：

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

(15)

7 不可分-软间隔

7.1 原始问题

允许支持向量机在一些样本上出错，采用惩罚项的方式来放松约束。

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

采用 hinge loss

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

令 $\xi_i = \max\left(0, 1 - y_i\left(\boldsymbol{w}^\top \phi\left(\boldsymbol{x}_i\right) + b\right)\right)$
则优化问题等价于：

$$\begin{aligned} &\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m} \xi_i \\ &\text{s.t. } y_i\left(\boldsymbol{w}^\top \phi\left(\boldsymbol{x}_i\right) + b\right) \geq 1 - \xi_i \\ &\qquad \xi_i \geq 0, i = 1, 2, \ldots, m. \end{aligned} \tag{16}$$

7.2　对偶问题

同样使用拉格朗日乘子法，得到对偶问题：

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j \phi\left(\boldsymbol{x}_i\right)^\top \phi\left(\boldsymbol{x}_j\right) - \sum_{i=1}^{m} \alpha_i \\ &\text{s.t. } \sum_{i=1}^{m} \alpha_i y_i = 0 \\ &\qquad 0 \leq \alpha_i \leq C, i = 1, 2, \ldots, m. \end{aligned} \tag{17}$$

比硬间隔只多了 $\alpha \leq C$ 的约束