

Support Vector Machine

支持向量机能够通过核函数非常方便的拓展到非线性分类的领域，为方便计算，这里假设 $y = \{1, -1\}$ ，**假设数据是线性可分的**，可以使用超平面将数据点分类 $h_{\omega,b}(X_i) = g(\omega X_i + b)$ 。

函数间距与几何间距

定义函数间距：

$$\tilde{\gamma}^i = y_i(\omega X_i + b), \quad \tilde{\gamma} = \min \tilde{\gamma}^i$$

这里用 y^i 乘是确保分类结果正确，即 $y^i(\omega X_i + b) > 0$ 。

定义几何距离：

$$\gamma^i = y_i \left(\frac{\omega X_i + b}{\|\omega\|} \right)$$

模型

通过最大化上述的函数间距，我们分别得到第一种模型，

$$\theta^* = \max_{\theta} \tilde{\gamma}, \quad s.t. y_i(\omega X_i + b) \geq \tilde{\gamma}, \|\omega\| = 1$$

第一个条件保证所有数据点到超平面的距离都大于函数间距，第二个条件将系数归一化。类似的，最大化几何间距可以得到第二种模型，

$$\theta^* = \max_{\theta} \frac{\tilde{\gamma}}{\|\omega\|}, \quad s.t. y_i(\omega X_i + b) \geq \tilde{\gamma}$$

不幸的是以上两种模型都属于非凸优化模型，难以求解。

为了转化问题，我们考虑最糟糕的情况，即只考虑距离超平面最近的数据点，这些数据点我们称之为支持向量。我们总可以选择合适的归一化条件 $\|\omega\|$ 使得这些支持向量的函数间距 $\tilde{\gamma} = 1$ ，所以第二种模型可以写成，

$$\theta^* = \max_{\omega} \frac{2}{\|\omega\|} = \min_{\omega} \frac{1}{2} \|\omega\|^2, \quad s.t. y_i(\omega X_i + b) \geq 1$$

这就是支持向量机模型。

拉格朗日对偶

拉格朗日乘数法一般用于求解带等式约束条件的函数极值问题。拉格朗日对偶是将一个带等式约束和不等式约束的优化问题转化成另一个更容易求解的问题。假设有

$$\begin{aligned} \min_x & f(x) \\ g_i(x) & \leq 0 \quad i = 1, \dots, m \\ h_i(x) & = 0 \quad i = 1, \dots, p \end{aligned}$$

构造广义拉格朗日乘子函数

$$L(x, \lambda, \nu) = f(x) + \sum_{i=1}^m \lambda_i g_i(x) + \sum_{i=1}^p \nu_i h_i(x)$$

其中 $\lambda_i \geq 0$ ，原问题的最优解为

$$p^* = \min_x \max_{\lambda, \nu} L(x, \lambda, \nu) = \min_x \theta_P(x)$$

$$\theta_P(x) = \max_{\lambda, \nu} L(x, \lambda, \nu)$$

其对偶问题以及最优解可以写成,

$$d^* = \max_{\lambda, \nu} \min_x L(x, \lambda, \nu) = \max_{\lambda, \nu} \theta_D(x)$$

$$\theta_D(x) = \min_x L(x, \lambda, \nu)$$

原问题与对偶问题的最优解之间满足弱对偶定理, 即 $d^* \leq p^*$

KKT条件

KKT条件是广义拉格朗日乘子法用于求解带有不等式约束条件优化问题的一阶必要条件。上述原问题应满足

$$\begin{cases} \nabla_x L(x, \lambda, \nu) = 0 \\ h_i(x) = 0 \\ g_i(x) \leq 0 \\ \lambda_i \geq 0 \\ \lambda_i g_i(x) = 0 \end{cases}$$

对于后面两个条件讨论, 当 $g_i(x) < 0$, 即数据点在不等式约束区域内部, 则 $\lambda_i = 0$, 即约束条件不起作用; 当 $g_i(x) = 0$, 即数据点在不等式约束区域边界, 则 $\lambda_i \geq 0$ 。

目标函数

首先, 支持向量机模型的拉格朗日函数可以写成

$$L(\omega, b, \alpha) = \frac{1}{2} \|\omega\|^2 + \sum_{i=1}^m \alpha_i g_i(\omega, b), \quad g_i(\omega, b) = (1 - y_i(\omega X_i + b))$$

支持向量机模型的原问题

$$p^* = \min_{\omega, b} \max_{\alpha} L(\omega, b, \alpha) = \min_{\omega, b} \theta_P(\alpha)$$

$$\theta_P(\alpha) = \max_{\alpha} L(\omega, b, \alpha)$$

其对偶问题以及最优解可以写成,

$$d^* = \max_{\alpha} \min_{\omega, b} L(\omega, b, \alpha) = \max_{\alpha} \theta_D(\alpha)$$

$$\theta_D(\alpha) = \min_{\omega, b} L(\omega, b, \alpha)$$

首先我们求解极值函数 $\theta_D(\omega, b)$, 应该满足KKT条件

$$\begin{cases} \nabla_{\omega} L(\omega, b, \alpha) = \omega - \sum_{i=1}^m \alpha_i y_i X_i \\ \nabla_b L(\omega, b, \alpha) = \sum_{i=1}^m \alpha_i y_i \\ g_i(\omega, b) \leq 0 \\ \alpha_i \geq 0 \\ \alpha_i g_i(\omega, b) = 0 \end{cases}$$

通过前两个条件, 我们可以得到

$$\omega^* = \sum_{i=1}^m \alpha_i y_i X_i, \quad \sum_{i=1}^m \alpha_i y_i = 0$$

代入向量机对偶问题

$$d^* = \max_{\alpha} \theta_D(\alpha) = \max_{\alpha} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j X_i^T X_j \quad \sum_{i=1}^m \alpha_i y_i = 0, \alpha_i \geq 0$$

考虑后面三个KKT条件，我们发现 $\alpha_i > 0$ 时，即 $g_i(\omega, b) = 0, y_i(\omega X_i + b) = 1$ 时约束条件起作用，对 ω^* 的取值有影响。

算法优化

核函数

支持向量机模型为了能够处理非线性分类的问题，引入了核函数的方法。这个方法将样本的原始空间映射到一个更高维度的特征空间，使得样本在这个高维空间内线性可分。在高维空间又可以利用核函数避免计算高维的向量内积。

低维到高维的映射： $x^i \rightarrow \phi(x^i)$

向量内积到核函数的映射： $\langle x_i, x_j \rangle = K(x_i, x_j) = K_{ij}$

常用的核函数有高斯核（无穷维），多项式核

软间隔

引入软间隔的概念，来允许支持向量机在一些样本上出错。在目标函数上加入松弛变量 ξ

$$\theta^* = \min_{\omega} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^m \xi_i, \quad s.t. y_i(\omega X_i + b) \geq 1 - \xi_i, \xi_i \geq 0$$

同样的，利用拉格朗日对偶方法，并考虑KKT条件可以得到

$$d^* = \max_{\alpha} \theta_D(\alpha) = \max_{\alpha} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j K_{ij} \quad \sum_{i=1}^m \alpha_i y_i = 0, C \geq \alpha_i \geq 0$$

算法实现：SMO算法

主要思想，将多变量问题转化为二变量优化问题。

坐标上升法

对多元函数，坐标上升法每次只对一个变量优化，其他变量固定不动，直至收敛。假设求解的优化问题为

$$\max_x f(x), \quad x = (x_1, x_2, \dots, x_d)$$

算法依次对每个 x_i 求函数的极值

$$x_i = \arg \max_{x_i} f(x)$$

缺点：对有些不可导的多元函数无法有效处理

SMO算法

使用核函数后，向量机对偶问题变成

$$d^* = \max_{\alpha} \theta_D(\alpha) = \min_{\alpha} - \sum_{i=1}^m \alpha_i + \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j K_{ij} \quad \sum_{i=1}^m \alpha_i y_i = 0, C \geq \alpha_i \geq 0$$

不失一般性的，我们先选择 α_1, α_2 作为优化变量，则 $y_1^2 = y_2^2 = 1, y_1 y_2 = s, \nu_i = \sum_{j=3}^m \alpha_j K_{ij} y_j$

$$\theta_D(\alpha) = \frac{1}{2} (\alpha_1^2 K_{11} + \alpha_2^2 K_{22}) + \alpha_1 \alpha_2 s K_{12} + \alpha_1 y_1 \nu_1 + \alpha_2 y_2 \nu_2 + C_1$$

约束条件

$$\begin{cases} \alpha_1 y_1 + \alpha_2 y_2 = C_2 \\ \alpha_1, \alpha_2 \geq 0 \end{cases}$$

代入求极值

$$\frac{\partial \theta_D(\alpha)}{\partial \alpha} = 0 \Rightarrow \alpha_2 = \frac{(K_{11} - K_{12})y_2 C_2 + y_2(\nu_1 - \nu_2) + 1 - s}{K_{11} + K_{22} - 2K_{12}}$$

已经知道 $h_\theta(X_j) = \sum_{i=1}^m \alpha_i y_i K_{ij} + b$, 容易推到

$$\nu_1 = h_\theta(x_1) - \sum_{j=1}^2 \alpha_j K_{1j} y_j - b, \quad \nu_2 = h_\theta(x_2) - \sum_{j=1}^2 \alpha_j K_{2j} y_j - b$$

代入上式, 令 $E_1 = h_\theta(x_1) - y_1$

$$\begin{aligned} \alpha_2 &= \frac{y_2((K_{11} - K_{12})C_2 + (E_1 - E_2) - y_1 \alpha'_1 K_{11} - y_2 \alpha'_2 K_{12} + y_1 \alpha'_1 K_{21} + y_2 \alpha'_2 K_{22})}{K_{11} + K_{22} - 2K_{12}} \\ &= \alpha'_2 + \frac{y_2(E_1 - E_2)}{K_{11} + K_{22} - 2K_{12}} \end{aligned}$$

这里使用了关系 $\alpha'_1 y_1 + \alpha'_2 y_2 = \alpha_1 y_1 + \alpha_2 y_2$, 由此得到 $\alpha_1 = \alpha'_1 + s(\alpha'_2 - \alpha'_1)$ 。在软间隔支持向量机中, 还应满足条件 $0 < \alpha_1, \alpha_2 \leq C$ 。我们可以通过 $y_1, y_2, \alpha'_1 y_1 + \alpha'_2 y_2, C$ 的取值判断直线与正方形的相对位置, 进而推断 α_2 的取值。

最后, 更新变量 b

$$\begin{cases} b_1 = y_1 - \omega^* x_1 \\ b_2 = y_2 - \omega^* x_2 \end{cases} \quad b = \begin{cases} \frac{b_1 + b_2}{2} & \text{if}(x_2, y_2) = SupportVector \\ b_1 & \text{if not} \end{cases}$$