

# Quadratic Programming: SVMs as Examples

Apple Zhang

2022 年 10 月 23 日

## 1 上下界约束的二次规划问题

对于支持向量机

$$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|_2^2 + c \sum_{i=1}^n (1 - y_i \mathbf{w}^T \mathbf{x}_i)_+^p \quad (1)$$

其中  $(\cdot)_+ = \max(0, \cdot)$ . 对于  $p = 1, 2$ , 上述问题均可以转化为以下形式的优化问题.

$$\tilde{\alpha} = \arg \min_{\alpha} f(\alpha) = \frac{1}{2} \alpha^T \mathbf{Q} \alpha + \mathbf{h}^T \alpha, \quad 0 \leq \alpha_i \leq c. \quad (2)$$

其中  $f(\alpha)$  为优化目标函数. 则 (1) 的最优解为

$$\mathbf{w} = \sum_{i=1}^n y_i \tilde{\alpha}_i \mathbf{x}_i. \quad (3)$$

### 1.1 坐标下降法

文献 [1] 提出了采用坐标下降 (Coordinate Descent, CD) 的方法求解问题 (2). 在每个 epoch 中将更新  $\alpha$  的每个元素且每个元素只更新一次. 当更新第  $i$  个元素, 即  $\alpha_i$  时, 固定其它的元素不变. 假设  $\alpha_i$  将被更新为  $\alpha_i + d$ , 则问题转化为

$$\min_d f(\alpha + d\mathbf{e}_i), \quad \text{s.t. } 0 \leq \alpha_i + d \leq c. \quad (4)$$

其中  $\mathbf{e}_i$  表示只有第  $i$  个元素为 1, 其余元素均为 0 的列向量. 化简:

$$\begin{aligned} & f(\alpha + d\mathbf{e}_i), \\ &= \frac{1}{2} (\alpha + d\mathbf{e}_i)^T \mathbf{Q} (\alpha + d\mathbf{e}_i) + \mathbf{h}^T (\alpha + d\mathbf{e}_i), \\ &= \frac{1}{2} \mathbf{e}_i^T \mathbf{Q} \mathbf{e}_i d^2 + (\mathbf{Q} \alpha)_i d + h_i d + f(\alpha), \\ &= \frac{1}{2} Q_{ii} d^2 + [(\mathbf{Q} \alpha)_i + h_i] d + \text{const} \\ &= \frac{1}{2} Q_{ii} d^2 + \nabla_i f(\alpha) d + \text{const} \end{aligned}$$

其中  $\nabla_i f(\boldsymbol{\alpha})$  表示  $f(\boldsymbol{\alpha})$  对  $\alpha_i$  的梯度. 即 (4) 等价于

$$\min_d \frac{1}{2} Q_{ii} d^2 + \nabla_i f(\boldsymbol{\alpha}) d, \quad \text{s.t. } 0 \leq \alpha_i + d \leq c. \quad (5)$$

若不考虑约束, 则该问题最优解为

$$d = -\frac{\nabla_i f(\boldsymbol{\alpha})}{Q_{ii}}. \quad (6)$$

在考虑约束时,  $\alpha_i$  的更新不能逃离  $[0, c]$  的区间, 因此得到以下的更新公式:

$$\alpha_i := \pi_{[0, c]} \left( \alpha_i - \frac{\nabla_i f(\boldsymbol{\alpha})}{Q_{ii}} \right). \quad (7)$$

其中  $\pi_{[0, c]}(\cdot)$  表示到区间  $[0, c]$  上的投影, 即  $\pi_{[0, c]}(\cdot) = \min(\max(\cdot, 0), c)$ .

在以上更新中, 每次计算  $\nabla_i f(\boldsymbol{\alpha})$  都涉及到矩阵  $\mathbf{Q}$  的计算, 为了降低计算量, [1] 采用了 “ $\mathbf{w}$  技巧”, 注意到<sup>1</sup>

$$(\mathbf{Q}\boldsymbol{\alpha})_i = \sum_{j=1}^n y_i y_j \mathbf{x}_j^T \mathbf{x}_i \alpha_j = \sum_{j=1}^n (\alpha_j y_j \mathbf{x}_j)^T \mathbf{x}_i = \mathbf{w}^T \mathbf{x}_i. \quad (8)$$

因此  $\nabla_i f(\boldsymbol{\alpha}) = \mathbf{w}^T \mathbf{x}_i + h_i$ . 从而不需要存储整个矩阵  $\mathbf{Q}$ , 只需存储其所有对角线元素即可进行计算. 而在每一次更新  $\alpha_i$  后, 我们也需要更新  $\mathbf{w}$  为

$$\mathbf{w} := \mathbf{w} + (\alpha_i^{new} - \alpha_i^{old}) y_i \mathbf{x}_i. \quad (9)$$

#### Example: Inverse Free TWSVM.

考虑孪生支持向量机,

$$\min_{\mathbf{w}_+, \boldsymbol{\xi}} \frac{1}{2} \|\mathbf{A}^T \mathbf{w}_+\|_2^2 + c \mathbf{1}^T \boldsymbol{\xi} + \frac{r}{2} \|\mathbf{w}_+\|_2^2, \quad \text{s.t. } \mathbf{1} - \boldsymbol{\xi} + \mathbf{B}^T \mathbf{w}_+ \leq \mathbf{0}, \quad \boldsymbol{\xi} \geq \mathbf{0}. \quad (10)$$

$$\min_{\mathbf{w}_-, \boldsymbol{\eta}} \frac{1}{2} \|\mathbf{B}^T \mathbf{w}_-\|_2^2 + c \mathbf{1}^T \boldsymbol{\eta} + \frac{r}{2} \|\mathbf{w}_-\|_2^2, \quad \text{s.t. } \mathbf{1} - \boldsymbol{\eta} - \mathbf{A}^T \mathbf{w}_- \leq \mathbf{0}, \quad \boldsymbol{\eta} \geq \mathbf{0}. \quad (11)$$

其中  $\mathbf{A}, \mathbf{B}$  为正负类矩阵:  $\mathbf{A} = [\mathbf{x}_1^+, \dots, \mathbf{x}_{n_A}^+]$ ,  $\mathbf{B} = [\mathbf{x}_1^-, \dots, \mathbf{x}_{n_B}^-]$ . 两个优化问题的解法相同, 所以下面只讨论第一个问题的求解. IF-TWSVM (Inverse Free TWSVM) 通过改写目标函数, 使得求解过程不会涉及求逆:

$$\min_{\mathbf{w}_+} \frac{1}{2} \|\mathbf{t}\|_2^2 + c \mathbf{1}^T \boldsymbol{\xi} + \frac{r}{2} \|\mathbf{w}_+\|_2^2, \quad \text{s.t. } \mathbf{1} - \boldsymbol{\xi} + \mathbf{B}^T \mathbf{w}_+ \leq \mathbf{0}, \quad \boldsymbol{\xi} \geq \mathbf{0}, \quad \mathbf{A}^T \mathbf{w}_+ = \mathbf{t}. \quad (12)$$

拉格朗日函数为

$$L(\mathbf{w}_+, \mathbf{t}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2} \|\mathbf{t}\|_2^2 + c \mathbf{1}^T \boldsymbol{\xi} + \frac{r}{2} \|\mathbf{w}_+\|_2^2 + \boldsymbol{\alpha}^T (\mathbf{1} - \boldsymbol{\xi} + \mathbf{B}^T \mathbf{w}_+) - \boldsymbol{\beta}^T \boldsymbol{\xi} + \boldsymbol{\lambda}^T (\mathbf{t} - \mathbf{A}^T \mathbf{w}_+). \quad (13)$$

根据 KKT 条件, 令  $\nabla_{\mathbf{w}_+} L = 0, \nabla_{\mathbf{t}} L = 0, \nabla_{\boldsymbol{\xi}} L = 0$  可以得到

$$\mathbf{w}_+ = \frac{1}{r} (\mathbf{B}\boldsymbol{\alpha} - \mathbf{A}\boldsymbol{\lambda}), \quad \mathbf{t} = \boldsymbol{\lambda}, \quad c \mathbf{1} = \boldsymbol{\alpha} + \boldsymbol{\beta}. \quad (14)$$

<sup>1</sup>这里指的是  $p = 1$ , 即常规 SVM 的情况, 对于  $p = 2$ , 即采用平方铰链损失的 SVM, 有  $\nabla_i f(\boldsymbol{\alpha}) = \mathbf{w}^T \mathbf{x}_i + h - \alpha_i / (2c)$ .

则可以得到对偶问题

$$\max_{\alpha, \lambda} \frac{1}{2} \begin{bmatrix} \lambda^T & \alpha^T \end{bmatrix} \begin{bmatrix} \mathbf{A}^T \mathbf{A} + r \mathbf{I} & -\mathbf{A}^T \mathbf{B} \\ -\mathbf{B}^T \mathbf{A} & \mathbf{B}^T \mathbf{B} \end{bmatrix} \begin{bmatrix} \lambda \\ \alpha \end{bmatrix} - r \begin{bmatrix} \mathbf{0}^T & \mathbf{1}^T \end{bmatrix} \begin{bmatrix} \lambda \\ \alpha \end{bmatrix}, \quad \text{s.t. } 0 \leq \alpha_i \leq c. \quad (15)$$

其是规模为  $n$  的二次规划问题. 设其目标函数为  $f(\mathbf{s})$ , 其中

$$\mathbf{s} = \begin{bmatrix} \lambda \\ \alpha \end{bmatrix} \quad (16)$$

采用坐标下降法进行求解, 首先注意到

$$\nabla_{\mathbf{s}} f = \left( \begin{bmatrix} r \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{A}^T \mathbf{A} & -\mathbf{A}^T \mathbf{B} \\ -\mathbf{B}^T \mathbf{A} & \mathbf{B}^T \mathbf{B} \end{bmatrix} \right) \begin{bmatrix} \lambda \\ \alpha \end{bmatrix} - r \begin{bmatrix} \mathbf{0} \\ \mathbf{1} \end{bmatrix} \quad (17)$$

$$= r \left( \begin{bmatrix} \lambda \\ -1 \end{bmatrix} + \begin{bmatrix} -\mathbf{A}^T \\ \mathbf{B}^T \end{bmatrix} \mathbf{w}_+ \right) \quad (18)$$

根据前面坐标下降法的原理, 可以将更新分为两个部分:

**$\lambda$  part:** 更新  $\lambda_i$ , 由于  $\lambda_i$  是无约束的, 因此有

$$\lambda_i := \lambda_i - \frac{\nabla_{\lambda_i} f(\mathbf{s})}{\|\mathbf{x}_i^+\|_2^2}. \quad (19)$$

此处  $\nabla_{\lambda_i} f(\mathbf{s}) = r(\lambda_i - \mathbf{w}_+^T \mathbf{x}_i^+)$ . 并更新

$$\mathbf{w}_+ := \mathbf{w}_+ + \frac{\nabla_{\lambda_i} f(\mathbf{s})}{r \|\mathbf{x}_i^+\|_2^2} \mathbf{x}_i^+. \quad (20)$$

**$\alpha$  part:** 根据前面分析, 可以直接得到

$$\alpha_i := \pi_{[0, c]} \left( \alpha_i - \frac{\nabla_{\alpha_i} f(\mathbf{s})}{\|\mathbf{x}_i^-\|_2^2} \right). \quad (21)$$

而这里  $\nabla_{\alpha_i} f(\mathbf{s}) = r(-1 + \mathbf{w}_+^T \mathbf{x}_i^-)$ . 并更新

$$\mathbf{w}_+ := \mathbf{w}_+ + \frac{1}{r} (\alpha_i^{new} - \alpha_i^{old}) \mathbf{x}_i^-. \quad (22)$$

## 1.2 超松弛迭代法

TODO!

## 参考文献

- [1] C.-J. Hsieh, K.-W. Chang, C.-J. Lin, S. S. Keerthi, and S. Sundararajan, "A dual coordinate descent method for large-scale linear svm," in *Proceedings of the 25th international conference on Machine learning*, pp. 408–415, 2008.