机器学习白板推导系列笔记 (6.1~6.4)

Dexing Huang
dxhuang@bupt.edu.cn
Beijing University of Posts and Telecommunications

日期：2021 年 11 月 7 日

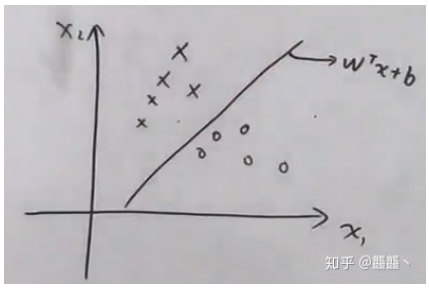
1 支持向量机 (SVM) 简介

支持向量机器是一种二分类模型, 它跟感知机的区别是数据点跟分类超平面的间隔最大. 感知力利用错误驱动的思想求解超平面, 但是这个解有无穷多; SVM 利用最大化间隔求解超平面, 此时只有唯一解.

SVM 有三宝, 间隔，对偶，核技巧. 从类别上看分为三类: 硬间隔 SVM, 软间隔 SVM 以及核技巧 SVM.

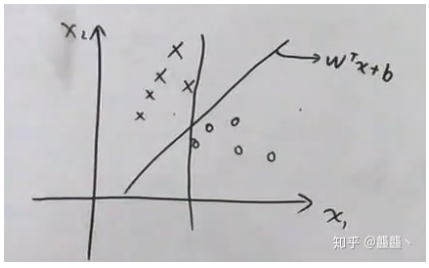
2 硬间隔 SVM

2.1 模型定义



SVM 解决的是二分类问题, 设数据集 $\mathcal{D}=\{(\mathbf{x}_i,y_i)\}_{i=1}^N$, 其中 $\mathbf{x}_i\in\mathbb{R}^p, y_i\in\{+1,-1\}$, 目的是在输入空间内找到一个超平面$\mathbf{w}^T\mathbf{x}+b$ 将两类数据分开, 因此模型为

$$f(\mathbf{x})=sign(\mathbf{w}^T\mathbf{x}+b) \tag{1}$$



可以看到, 有很多超平面可以满足将两类数据分开这个要求, 那么哪一个超平面才是最好的呢, 从距离的角度来看, 距离超平面最远的点将其预测正确确信度是比较高的, 但相对来说, 距离超平面越近的点预测结果的确信度越低. 因此, 若令距离超平面最近的点的距离尽可能大, 即间隔最大化, 那么所有点的预测结果确信度便是最高的.

那么就引出了一个问题, 如何定义间隔呢? 扩充高中学过的知识, 点到直线的距离, 一个点到超平面的距离可以表示为

$$\frac{1}{||\mathbf{w}||}|\mathbf{w}^T\mathbf{x}_i + b| \tag{2}$$

间隔就是所有点到超平面的最小距离, 即

$$\min \frac{1}{||\mathbf{w}||}|\mathbf{w}^T\mathbf{x}_i + b| \tag{3}$$

因此间隔最大化表述成数学语言是

$$\max_{\mathbf{w},b} \min_{\mathbf{x}_i} \frac{1}{||\mathbf{w}||}|\mathbf{w}^T\mathbf{x}_i + b| \tag{4}$$

此外还需要添加一个约束, 因为我们希望的是所有点均被正确分类, 即

$$\mathbf{w}^T\mathbf{x}_i + b > 0 \quad y_i = +1$$

$$\mathbf{w}^T\mathbf{x}_i + b < 0 \quad y_i = -1$$

$$\Rightarrow y_i(\mathbf{w}^T\mathbf{x}_i + b) > 0$$

因此完整的间隔最大化的数学表示为

$$\begin{cases} \max\limits_{\mathbf{w},b} \min\limits_{\mathbf{x}_i} \frac{1}{||\mathbf{w}||}|\mathbf{w}^T\mathbf{x}_i + b| \\ s.t. \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) > 0 \end{cases}$$

由约束条件可以将极大极小化表示为

$$\begin{cases} \max\limits_{\mathbf{w},b} \frac{1}{||\mathbf{w}||} \min\limits_{\mathbf{x}_i} y_i(\mathbf{w}^T\mathbf{x}_i + b) \\ s.t. \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) > 0 \end{cases}$$

由 $y_i(\mathbf{w}^T\mathbf{x}_i + b)$ 可知, 一定存在 $\gamma > 0$ 使得 $\min y_i(\mathbf{w}^T\mathbf{x}_i + b) = \gamma$, 又因为只需要求解参数 $\mathbf{w}, b$ 即可, 因此不妨设 $\gamma = 1$, 那么上式化为

$$\begin{cases} \max\limits_{\mathbf{w},b} \frac{1}{||\mathbf{w}||} \\ s.t. \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \end{cases}$$

又因为极大化 $\frac{1}{||\mathbf{w}||}$ 和极小化 $||\mathbf{w}||^2$ 是等价的, 因此

$$\begin{cases} \min\limits_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 \\ s.t. \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \end{cases}$$

注. $\frac{1}{2}$ 只是为了求导简便, 对结果无影响

可以看到, 间隔最大化被表示成了上述形式, 这是一个凸二次规划问题, 下节对其进行求解.

2.2 模型求解

重新观察一下最优化问题

$$\begin{cases} \min\limits_{\mathbf{w},b} \dfrac{1}{2}\mathbf{w}^T\mathbf{w} \\ s.t. \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \end{cases} \tag{5}$$

它是一个带约束最优化问题, 一个自然而然的解决思路就是利用拉格朗日乘子法进行解决, 构造拉格朗日函数

$$\mathcal{L}(\mathbf{w}, b, \lambda) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{N} \lambda_i(1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)) \tag{6}$$

其中 $\lambda_i \geq 0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b) \leq 0$, 为什么 $\lambda_i \geq 0$ 呢, 在**附录**中进行解释. 将优化问题式 (5) 转化为无约束优化问题 (对于参数 $\mathbf{w}, b$ 来说)

$$\begin{cases} \min\limits_{\mathbf{w},b} \max\limits_{\lambda} \mathcal{L}(\mathbf{w}, b, \lambda) \\ s.t. \quad \lambda_i \geq 0 \end{cases} \tag{7}$$

为什么优化问题 (5) 和 (7) 等价, 在**附录**中进行定量的解释. 下面引出该问题的对偶问题

$$\begin{cases} \max\limits_{\lambda} \min\limits_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \lambda) \\ s.t. \quad \lambda_i \geq 0 \end{cases} \tag{8}$$

此对偶问题 (8) 与 (7) 等价, 又称为强对偶问题. 下面先介绍对偶问题的概念, 对于优化问题 $\mathcal{L}$ 有

$$\min \max \mathcal{L} \geq \max \min \mathcal{L} \tag{9}$$

此关系称为弱对偶关系, 在这里不对其进行数学证明, 采用直观的方式理解, $\min \max \mathcal{L}$ 相当于在“凤”里选最小 (凤尾), $\max \min \mathcal{L}$ 相当于在“鸡”里选最大 (鸡头), 显然凤尾是大于鸡头的 (笑).

当等号成立时, 弱对偶关系转化为强对偶关系, 又因为凸优化问题都是强对偶的 (不做证明), 因此上述转化成立.

综上, 我们就把原始的最优化问题 (5) 转化为了 (8), 下面来求解优化问题 (8).

2.2.1 w 和 b 的求解

先对 b 求导, 令导数为 0

$$\frac{\partial}{\partial b}\mathcal{L}(\mathbf{w}, b, \lambda) = \sum_{i=1}^{N} -\lambda_i y_i = 0$$

解得

$$\sum_{i=1}^{N} \lambda_i y_i = 0 \tag{10}$$

将式 (10) 代入 (6) 得

$$\begin{aligned}\mathcal{L}(\mathbf{w},b,\lambda) &= \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{N}\lambda_i - \sum_{i=1}^{N}\lambda_i y_i(\mathbf{w}^T\mathbf{x}_i + b) \\ &= \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{N}\lambda_i - \sum_{i=1}^{N}\lambda_i y_i \mathbf{w}^T\mathbf{x}_i \end{aligned} \tag{11}$$

对 $\mathbf{w}$ 求导, 令导数为 0

$$\frac{\partial}{\partial \mathbf{w}}\mathcal{L}(\mathbf{w},b,\lambda) = \mathbf{w} - \sum_{i=1}^{N}\lambda_i y_i \mathbf{x}_i$$

解得

$$\mathbf{w} = \sum_{i=1}^{N}\lambda_i y_i \mathbf{x}_i \tag{12}$$

将上式代入式 (11) 得

$$\mathcal{L}(\mathbf{w},b,\lambda) = \frac{1}{2}(\sum_{i=1}^{N}\lambda_i y_i \mathbf{x}_i)^T(\sum_{j=1}^{N}\lambda_j y_j \mathbf{x}_j) - \sum_{i=1}^{N}\lambda_i y_i(\sum_{j=1}^{N}\lambda_j y_j \mathbf{x}_j)^T\mathbf{x}_i + \sum_{i=1}^{N}\lambda_i$$

由于 $\lambda_i y_i$ 为常数, x_i 为向量, 因此上式可以改写为

$$\begin{aligned}\mathcal{L}(\mathbf{w},b,\lambda) &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j - \sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j \mathbf{x}_j^T\mathbf{x}_i + \sum_{i=1}^{N}\lambda_i \\ &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j - \sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j \mathbf{x}_i\mathbf{x}_i^T + \sum_{i=1}^{N}\lambda_i \\ &= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j + \sum_{i=1}^{N}\lambda_i \end{aligned} \tag{13}$$

因此问题 (8) 被转化为了最大值问题

$$\begin{cases} \max\limits_{\lambda} -\frac{1}{2}\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{N}\lambda_i\lambda_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j + \sum\limits_{i=1}^{N}\lambda_i \\ s.t. \quad \lambda_i \geq 0, \sum\limits_{i=1}^{N}\lambda_i y_i = 0 \end{cases} \tag{14}$$

2.2.2 KKT 条件

式 (14) 的等价表示

$$\begin{cases} \min\limits_{\lambda} \frac{1}{2}\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{N}\lambda_i\lambda_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j - \sum\limits_{i=1}^{N}\lambda_i \\ s.t. \quad \lambda_i \geq 0, \sum\limits_{i=1}^{N}\lambda_i y_i = 0 \end{cases} \tag{15}$$

最终的目的是求出最优参数 $\mathbf{w}^\star, b,$ 仍然可以通过拉格朗日乘子法来求解, 但这里有一种更加简便的求解方式, 先介绍一下KKT 条件, 强对偶的充要条件是满足 KKT 条件, 因此对于本问题来

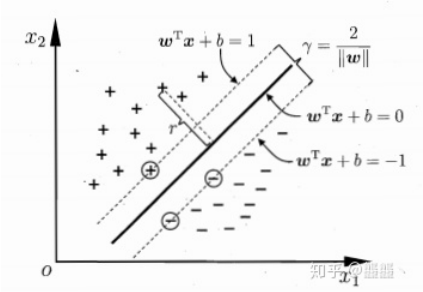
说是满足 KKT 条件的, 即

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0, \dfrac{\partial \mathcal{L}}{\partial b} = 0, \dfrac{\partial \mathcal{L}}{\partial \lambda} = 0 \\ \lambda_i (1 - y_i(\mathbf{w}^T \mathbf{x}_i + b)) = 0 \\ \lambda_i \geq 0 \\ 1 - y_i(\mathbf{w}^T \mathbf{x}_i + b) \leq 0 \end{cases} \tag{16}$$

根据上一部分的讨论可知

$$\mathbf{w}^\star = \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i \tag{17}$$

只需要求解出 $b^\star$ 即可, 需要用到 KKT 条件中的 $\lambda_i(1 - y_i(\mathbf{w}^T \mathbf{x}_i + b)) = 0$



这一条件也称为松弛互补条件, 其中距离超平面最近的数据点正好在 $\mathbf{w}^T \mathbf{x} + b = 1$ 或 $\mathbf{w}^T \mathbf{x} + b = -1$ 上, 这些点也称为支持向量.

现在来看条件 $\lambda_i(1 - y_i(\mathbf{w}^T \mathbf{x}_i + b)) = 0$, 对于支持向量来说, $1 - y_i(\mathbf{w}^T \mathbf{x}_i + b) = 0$ 成立, 对应的 λ_i 为任意值, 但是对于其他点来说, $1 - y_i(\mathbf{w}^T \mathbf{x}_i + b) \neq 0$, 那么这些点对应的 $\lambda_i = 0$.

因此, 存在 $(\mathbf{x}_k, y_k)$ 是支持向量, 使得 $1 - y_k(\mathbf{w}^T \mathbf{x}_k + b) = 0$, 那么

$$\begin{aligned} y_k(\mathbf{w}^{\star T} \mathbf{x}_k + b^\star) &= 1 \\ y_k^2(\mathbf{w}^{\star T} \mathbf{x}_k + b^\star) &= y_k \\ \mathbf{w}^{\star T} \mathbf{x}_k + b^\star &= y_k \\ b^\star &= y_k - \mathbf{w}^{\star T} \mathbf{x}_k \\ b^\star &= y_k - \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i^T \mathbf{x}_k \end{aligned}$$

那么就得到了最优参数

$$\begin{cases} \mathbf{w}^\star = \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i \\ b^\star = y_k - \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i^T \mathbf{x}_k \end{cases} \tag{18}$$

由于只有在支持向量上才有 $\lambda_i \neq 0$, 因此可以看出, 式 (18) 的值只与支持向量有关, 得到最优分类决策函数 $f(\mathbf{x}) = sign(\mathbf{w}^{\star T} \mathbf{x} + b^\star)$ 和超平面 $\mathbf{w}^{\star T} \mathbf{x} + b^\star$.

对于如何找到支持向量, 视频中并没有讲, 下面举一个李航老师《统计学习方法》中的例子来说明.

例. 假设正例点为 $\mathbf{x}_1 = (3,3)^T$, $\mathbf{x}_2 = (4,3)^T$, 负例点为 $\mathbf{x}_3 = (1,1)^T$, 求解线性可分的支持向量机.

根据数据, 对偶问题是

$$\begin{cases} \min\limits_{\lambda} \dfrac{1}{2}\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{N}\lambda_i\lambda_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j - \sum\limits_{i=1}^{N}\lambda_i \\ = \dfrac{1}{2}(18\lambda_1^2 + 25\lambda_2^2 + 2\lambda_3^2 + 42\lambda_1\lambda_2 - 12\lambda_1\lambda_3 - 14\lambda_2\lambda_3) - \lambda_1 - \lambda_2 - \lambda_3 \\ s.t. \quad \lambda_i \geq 0,\ \lambda_1 + \lambda_2 - \lambda_3 = 0 \end{cases}$$

解这一问题, $\lambda_3 = \lambda_1 + \lambda_2$

$$s(\lambda_1, \lambda_2) = 4\lambda_1^2 + \frac{13}{2}\lambda_2^2 + 10\lambda_1\lambda_2 - 2\lambda_1 - 2\lambda_2 \tag{19}$$

分别对 λ_1, λ_2 求偏导

$$\frac{\partial s(\lambda_1,\lambda_2)}{\partial \lambda_1} = 8\lambda_1 + 10\lambda_2 - 2$$
$$\frac{\partial s(\lambda_1,\lambda_2)}{\partial \lambda_2} = 10\lambda_1 + 13\lambda_2 - 2$$

令偏导数为 0 得 $(\frac{3}{2}, -1)^T$ 取得极值, 但是不满足 $\lambda_i \geq 0$ 的条件, 因此最小值在边界处.

- 当 $\lambda_1 = 0$ 时 $s(0, \frac{2}{13}) = -\frac{2}{13}$
- 当 $\lambda_2 = 0$ 时 $s(\frac{1}{4}, 0) = -\frac{1}{4}$

于是 $s(\lambda_1, \lambda_2)$ 在 $(\frac{1}{4}, 0)$ 处取得最小, 此时 $\lambda_3 = \frac{1}{4} + 0 = \frac{1}{4}$, 因此可知 λ_1, λ_3 对应的点 $\mathbf{x}_1, \mathbf{x}_3$ 为支持向量. 于是

$$\mathbf{w}^\star = \frac{1}{4}(3,3)^T - \frac{1}{4}(1,1)^T = (\frac{1}{2}, \frac{1}{2})^T$$
$$b^\star = 1 - (\frac{1}{4}(3,3)(3,3)^T - \frac{1}{4}(1,1)(3,3)^T) = -2$$

因此分离超平面为

$$\frac{1}{2}x_1 + \frac{1}{2}x_2 - 2 = 0 \tag{20}$$

分类决策函数为

$$f(\mathbf{x}) = sign(\frac{1}{2}x_1 + \frac{1}{2}x_2 - 2) \tag{21}$$

3 软间隔 SVM

3.1 模型定义

硬间隔 SVM 对线性不可分的数据不适用, 可以通过软间隔 SVM 和核函数 SVM 来解决, 本节主要讲述软间隔 SVM. 造成训练数据线性不可分的一种情况是训练数据有一些异常值 (outlier), 因此需要引入软间隔 SVM, 其主要思想是允许分类时存在一点错误, 即在 $\frac{1}{2}\mathbf{w}^T\mathbf{w}$ 后再加上一项损失

$$\frac{1}{2}\mathbf{w}^T\mathbf{w} + loss \tag{22}$$

定义损失有两种方式, 在硬间隔 SVM 中, 所有点满足约束条件 $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$, 但对于异常

值来说往往不满足, 因此可以使用不满足的数据点的个数来表示, 即

$$loss = \sum_{i=1}^{N} \mathbb{I}\{y_i(\mathbf{w}^T\mathbf{x}_i + b) < 1\} \tag{23}$$

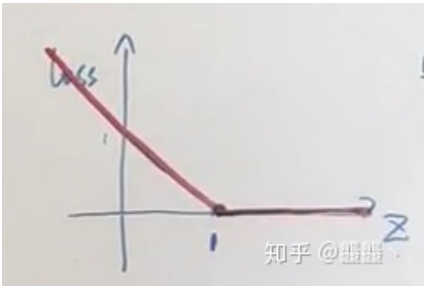
但就像在感知机中讨论的那样, 该函数不连续, 无法进行求导, 因此一般不用. 个数不行那就使用距离来表示

- 若 $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$ $loss = 0$
- 若 $y_i(\mathbf{w}^T\mathbf{x}_i + b) < 1$ $loss = 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)$

综上 loss 可以表示为如下所示函数

$$loss = \max\{0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)\} \tag{24}$$

若令 $z = y_i(\mathbf{w}^T\mathbf{x}_i + b)$, 则 loss 的图像如下



其连续 (在数学上严格来说并不连续), 也称为 hinge loss(合页函数). 此时优化问题可以写为

$$\begin{cases} \min\limits_{\mathbf{w},b} \dfrac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum\limits_{i=1}^{N} \max\{0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)\} \\ s.t. \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \text{(此处这么写感觉不严谨)} \end{cases} \tag{25}$$

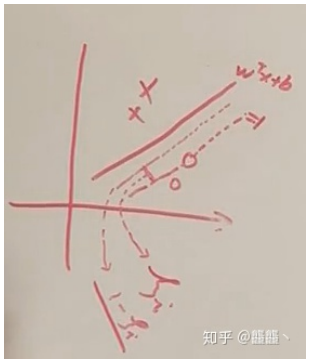
注. C 为超参数

为了简化起见, 令 $\xi_i = 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b), \xi_i \geq 0$, 因此上式化为

$$\begin{cases} \min\limits_{\mathbf{w},b,\xi} \dfrac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum\limits_{i=1}^{N} \xi_i \\ s.t. \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \ \xi_i \geq 0 \end{cases} \tag{26}$$

为何将约束条件中的 1 改为 $1 - \xi_i$? 如下图所示, 假设圆圈这类的支持向量的平面是 $\mathbf{w}^T\mathbf{x} + b = 1$, 对于大多数点来说 $y_i(\mathbf{w}^T\mathbf{x} + b) \geq 1$ 是成立的, 它们对应的 $\xi_i = 0$, 但是对于异常点来说并不满足, 因此设定了一定的容忍距离 ξ_i, 使得异常点也能满足约束条件.

对于 $\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N} \xi_i$ 笔者的理解是在硬间隔最大化的基础上加入了容错系数 ξ, 因为在满足 $y_i(\mathbf{w}^T\mathbf{x} + b) \geq 1$ 的时候 ξ=0, 因此目标还是企图找到一个间隔最大超平面同时也要保证异常点的个数最少.



3.2 模型求解

对于模型求解, 跟硬间隔最大化基本一个思路, 视频中并没有给出求解过程, 笔者参考《统计学习方法》进行求解. 首先构造拉格朗日函数

$$\mathcal{L}(\mathbf{w}, b, \xi, \lambda, \eta) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\lambda_i(1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b)) - \sum_{i=1}^{N}\eta_i\xi_i \tag{27}$$

注. $\lambda_i \geq 0, \eta_i \geq 0$

将问题 (26) 转化为

$$\begin{cases} \min\limits_{\mathbf{w},b,\xi}\max\limits_{\lambda,\eta}\mathcal{L}(\mathbf{w}, b, \xi, \lambda, \eta) \\ s.t. \quad \lambda_i \geq 0, \eta_i \geq 0 \end{cases} \tag{28}$$

上式为原始问题的等价表示, 现在将其转化为对偶问题

$$\begin{cases} \max\limits_{\lambda,\eta}\min\limits_{\mathbf{w},b,\xi}\mathcal{L}(\mathbf{w}, b, \xi, \lambda, \eta) \\ s.t. \quad \lambda_i \geq 0, \eta_i \geq 0 \end{cases} \tag{29}$$

下面对 $\mathbf{w}, b, \xi$ 分别求导并令其为 0

$$\frac{\partial}{\partial \mathbf{w}}\mathcal{L}(\mathbf{w}, b, \xi, \lambda, \eta) = \mathbf{w} - \sum_{i=1}^{N}\lambda_i y_i \mathbf{x}_i = 0$$

$$\frac{\partial}{\partial b}\mathcal{L}(\mathbf{w}, b, \xi, \lambda, \eta) = \sum_{i=1}^{N} -\lambda_i y_i = 0$$

$$\frac{\partial}{\partial \xi_i}\mathcal{L}(\mathbf{w}, b, \xi, \lambda, \eta) = C - \lambda_i - \eta_i = 0$$

代入式 (27) 得

$$\begin{aligned}
\mathcal{L}(\mathbf{w},b,\xi,\lambda,\eta) &= \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\lambda_i - \sum_{i=1}^{N}\lambda_i\xi_i - \sum_{i=1}^{N}\lambda_i y_i(\mathbf{w}^T\mathbf{x}_i + b) - \sum_{i=1}^{N}\eta_i\xi_i \\
&= \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\lambda_i - \sum_{i=1}^{N}\lambda_i\xi_i - \sum_{i=1}^{N}\lambda_i y_i\mathbf{w}^T\mathbf{x}_i - \sum_{i=1}^{N}\eta_i\xi_i \\
&= \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{N}\lambda_i - \sum_{i=1}^{N}\lambda_i y_i\mathbf{w}^T\mathbf{x}_i + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}(\lambda_i\xi_i + \eta_i\xi_i) \\
&= \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{N}\lambda_i - \sum_{i=1}^{N}\lambda_i y_i\mathbf{w}^T\mathbf{x}_i \\
&= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j\mathbf{x}_i^T\mathbf{x}_j + \sum_{i=1}^{N}\lambda_i \quad \text{(过程同式 (13))}
\end{aligned}$$

因此对偶问题 (29) 化为

$$\begin{cases}
\max\limits_{\lambda,\eta} -\dfrac{1}{2}\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j\mathbf{x}_i^T\mathbf{x}_j + \sum_{i=1}^{N}\lambda_i \\
s.t. \quad \displaystyle\sum_{i=1}^{N}\lambda_i y_i = 0 \\
\qquad C - \lambda_i - \eta_i = 0 \\
\qquad \lambda_i \geq 0, \eta_i \geq 0
\end{cases} \tag{30}$$

利用等式约束消去 η, 只留下 λ

$$0 \leq \lambda_i \leq C \tag{31}$$

于是得到

$$\begin{cases}
\min\limits_{\lambda} \dfrac{1}{2}\displaystyle\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j\mathbf{x}_i^T\mathbf{x}_j - \sum_{i=1}^{N}\lambda_i \\
s.t. \quad \displaystyle\sum_{i=1}^{N}\lambda_i y_i = 0 \\
\qquad 0 \leq \lambda_i \leq C
\end{cases} \tag{32}$$

这里又要用到 KKT 条件, 因为原始问题时凸二次规划问题, 解满足 KKT 条件

$$\begin{cases}
\dfrac{\partial}{\partial\mathbf{w}}\mathcal{L}(\mathbf{w},b,\xi,\lambda,\eta) = 0, \dfrac{\partial}{\partial b}\mathcal{L}(\mathbf{w},b,\xi,\lambda,\eta) = 0, \dfrac{\partial}{\partial\eta}\mathcal{L}(\mathbf{w},b,\xi,\lambda,\eta) = 0 \\
\lambda_i(y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i) = 0 \\
\eta_i\xi_i = 0 \\
y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i \geq 0 \\
\xi_i \geq 0, \lambda_i \geq 0, \eta_i \geq 0
\end{cases} \tag{33}$$

同样, $\mathbf{w}^\star$ 的值已知

$$\mathbf{w}^\star = \sum_{i=1}^{N}\lambda_i y_i\mathbf{x}_i \tag{34}$$

根据约束条件 $\lambda_i(y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i) = 0$ 和 $\eta_i\xi_i = 0$ 可知, 若存在 $\lambda_j \in (0, C)$, 那么

$y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 = 0$, 可得

$$b^\star = y_j - \sum_{i=1}^{N} y_i \lambda_i \mathbf{x}_i^T \mathbf{x}_j \tag{35}$$

那么就得到了分离超平面

$$\mathbf{w}^{\star T}\mathbf{x} + b^\star = 0 \tag{36}$$

以及分类决策函数

$$f(\mathbf{x}) = sign(\mathbf{w}^{\star T}\mathbf{x} + b^\star) \tag{37}$$

4 附录

4.1 式 (6) 拉格朗日乘子法 $\lambda_i \geq 0$

拉格朗日系数大于等于 0 可以分开讨论, 等于 0 时这个 constraint 不发挥任何作用, 问题变成了直接取最小值; 而大于 0 的时候, 最小值发生在了 feasible region 和 isocontour(等高线) 交界处.[1]

假设问题是

$$\begin{cases} \min\limits_{\mathbf{x}} f(\mathbf{x}) \\ s.t. \quad g(\mathbf{x}) \leq 0 \end{cases}$$

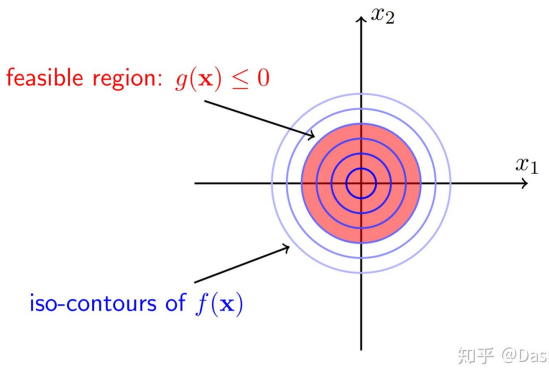
通过拉格朗日乘子法构造为

$$\mathcal{L}(\mathbf{w}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x}) \tag{38}$$

将问题转化为

$$\begin{cases} \min\limits_{\mathbf{x}} \max\limits_{\lambda} f(\mathbf{x}) + \lambda g(\mathbf{x}) \\ s.t. \quad \lambda_i \geq 0 \end{cases}$$

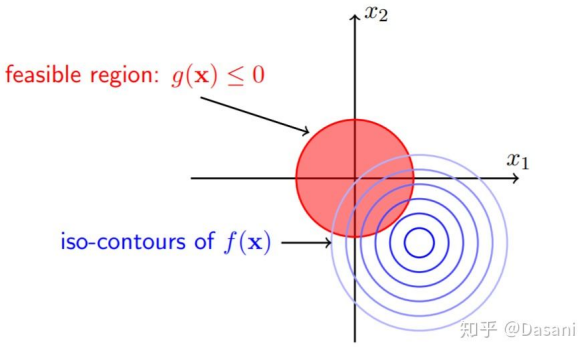
假设 $f(\mathbf{x}) = x_1^2 + x_2^2$, $g(\mathbf{x}) = x_1^2 + x_2^2 - 1$, 那这个时候可以很容易想到 f 的 global minimum 就是在 $g(\mathbf{x})$ 的限制里, 如图在图的圆心, 这个时候调用拉格朗日函数, 系数就是等于 0



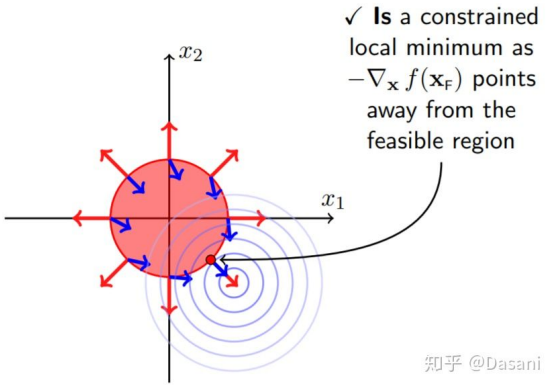
假设 $f(\mathbf{x}) = (x_1 - 1.1)^2 + (x_2 - 1.1)^2$, $g(\mathbf{x}) = x_1^2 + x_2^2 - 1$, f 的最小值就被 feasible region 限制了, kkt condition 下面有一个步骤是需要对 g 和 f 取导数并且让二者平行

$$-\nabla_{\mathbf{x}} f(\mathbf{x}^\star) = \lambda \nabla_{\mathbf{x}} g(\mathbf{x}^\star), \quad \lambda > 0 \tag{39}$$

[1] https://www.zhihu.com/question/375106014/answer/1042717952



为什么 λ 大于 0, 如下图所示, 最小值会发生在 feasible region 的边缘, 而且会发生在 $-\nabla_{\mathbf{x}} f$ 和 $-\nabla_{\mathbf{x}} g$ 方向相同的点上



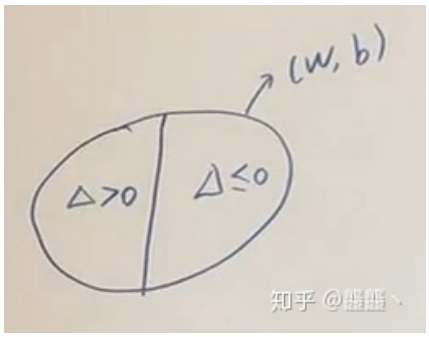
4.2 式 (5) 和式 (7) 等价性讨论

仅做定性的分析, 不做严格的数学证明, 让我们再重新看一下这两个问题

$$\begin{cases} \min\limits_{\mathbf{w},b} \dfrac{1}{2}\mathbf{w}^T\mathbf{w} \\ s.t. \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \end{cases} \qquad \begin{cases} \min\limits_{\mathbf{w},b} \max\limits_{\lambda} \mathcal{L}(\mathbf{w}, b, \lambda) \\ s.t. \quad \lambda_i \geq 0 \end{cases}$$

注. $\mathcal{L}(\mathbf{w}, b, \lambda) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{N} \lambda_i(1 - y_i(\mathbf{w}^T\mathbf{x}_i + b))$

令 $\Delta = 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)$, 对于 $(\mathbf{w}, b)$ 二元组来说, 可以将 Δ 分为大于 0 和小等 0 两类来看



- 若 $\Delta > 0$, 又因为 $\lambda_i \geq 0$, 那么 $\max_{\lambda} \mathcal{L}(\mathbf{w}, b, \lambda) \to \infty$

- 若 $\Delta \leq 0$, $\lambda_i \geq 0$, 那么当 $\lambda_i = 0$ 时 $\sum_{i=1}^{N} \lambda_i(1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)$ 最大, 即 $\max_\lambda \mathcal{L}(\mathbf{w}, b, \lambda) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$

  将上述讨论代入 $\min_{\mathbf{w},b} \max_\lambda \mathcal{L}(\mathbf{w}, b, \lambda)$

$$\min_{\mathbf{w},b} \max_{\lambda} \mathcal{L}(\mathbf{w}, b, \lambda) = \min_{\mathbf{w},b} \{\infty, \frac{1}{2}\mathbf{w}^T\mathbf{w}\} = \min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

  且此时 $\Delta \leq 0$, 因此二者的等价性得到了解释.