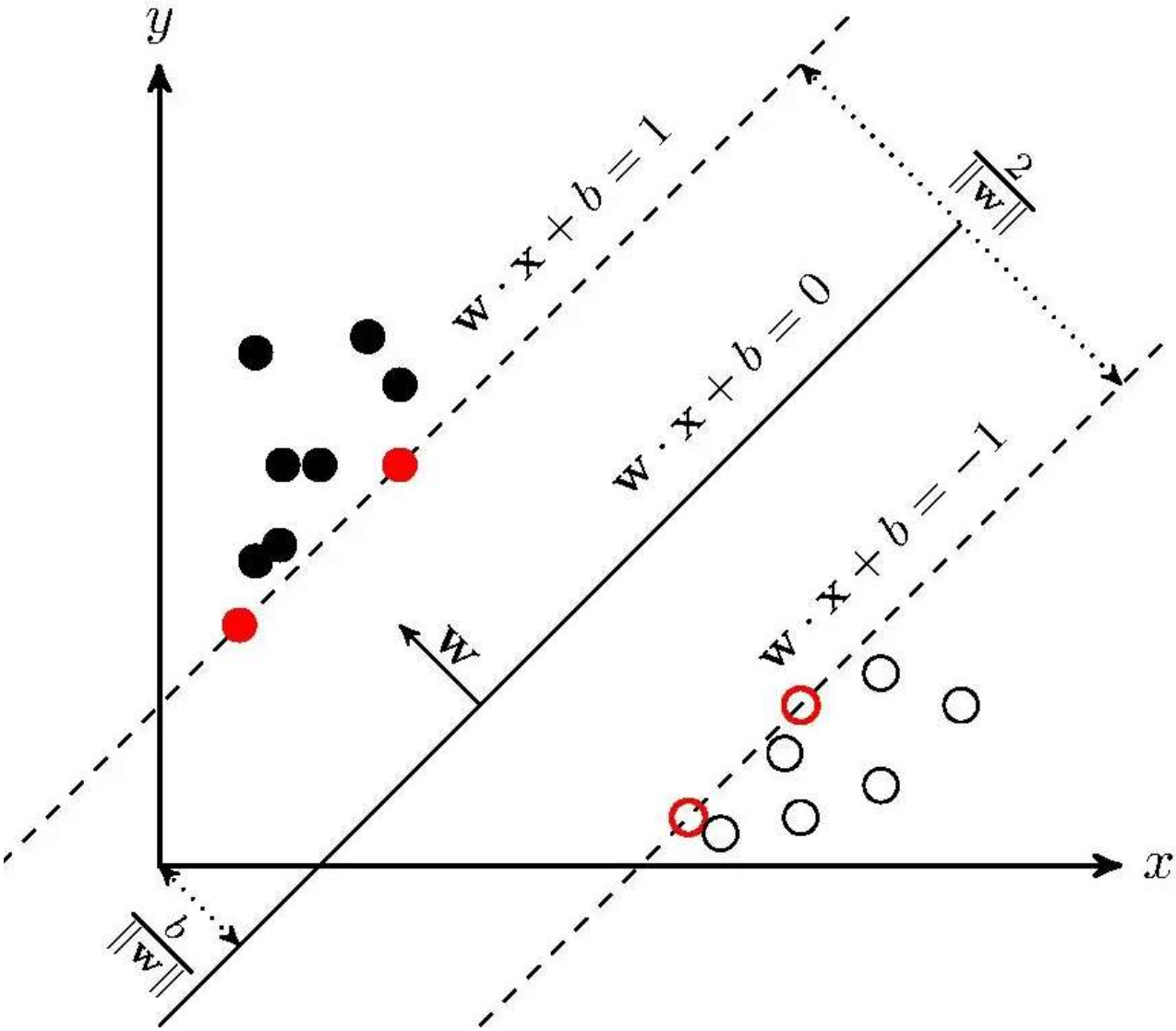
SVM和LSVM

硬间隔

对于任意的数据集我们要进行分类的一个方法就是找到一个平面可以将不同的类别进行划分。那么一种可行的就是找到一个超平面将空间划分为不相交的两个部分。那么对于二维的数据集而言，超平面就是一条直线。



那么就是如图所示找到一个超平面使得该平面离不同类别的距离最大。 此时假设该直线为:

$$w \cdot x + b = 0 \tag{1.1}$$

而该空间内任意一点到该直线的距离为:

$$\frac{w^T \cdot x + b}{||w||} \tag{1.2}$$

那么假设存在一个最大距离d，对于这个数据集而言需要满足:

$$\begin{cases} \frac{w^T \cdot x+b}{||w||} \geq d, y=1 \\ \frac{w^T \cdot x+b}{||w||} \leq -d, y=-1 \end{cases} \tag{1.3}$$

所以就可以统一成一个公式:

$$y(\frac{w^T \cdot x+b}{||w||}) \geq d \tag{1.4}$$

那么支持向量[支持平面就是数据集中离该超平面最近的点]到该超平面的距离就是:

$$y(\frac{w^T \cdot x+b}{||w||}) = d \tag{1.5}$$

那么问题就转变成为了:

$$argmax_{w,b}(\frac{1}{||w||}min_i[y_i(w^T \cdot x+b)]) \tag{1.6}$$

这里是找到距离超平面最近的样本点，然后找到使得该样本点离超平面最远的w,b
那么根据公式(1.4)可以知道$y(w^T \cdot x+b) > 0$,从而可以对其进行缩放，这里假设$min_i y(w^T \cdot x+b) = 1$因为同时调节$w,b$对结果$d$是没有任何影响的，所以可以通过调节比例使得$y(w^T \cdot x+b) = 1$。 那么经过上面的处理就变成了:

$$\begin{gathered} argmax_{w,b}(\frac{1}{||w||}) \\ s.t.\ \ y(w^T \cdot x+b) \geq 1 \end{gathered} \tag{1.7}$$

可以将其转换为:

$$\begin{gathered} argmin_{w,b}(\frac{||w||^2}{2}) \\ s.t.\ \ 1-y(w^T \cdot x+b) \leq 0 \end{gathered} \tag{1.8}$$

这样就可以使用拉格朗日乘子法进行求解:

$$\begin{gathered} L(w,b,\lambda) = \frac{||w||^2}{2} + \sum_{i=1}^{n} \lambda_i[1-y_i(w^T \cdot x_i+b)] \\ s.t. \quad \lambda \geq 0 \end{gathered} \tag{1.9}$$

那么问题就变成了求解:

$$min_{w,b}max_{\lambda}L(w,b,\lambda) \tag{1.10}$$

那么此时可以转化成为拉格朗日对偶问题

$$max_{\lambda}min_{w,b}L(w,b,\lambda) \tag{1.11}$$

此时计算原式关于w,b的偏导，偏导等于0，求得此时的最优解之一，再计算λ的最优值。

$$\frac{\partial L(w,b,\lambda)}{\partial w}=0\Rightarrow w-\sum_{i=1}^{n}\lambda_i y_i x_i=0\Rightarrow w=\sum_{i=1}^{n}\lambda_i y_i x_i$$
$$\frac{\partial L(w,b,\lambda)}{\partial b}=0\Rightarrow \sum_{i=1}^{n}\lambda_i y_i=0 \tag{1.12}$$

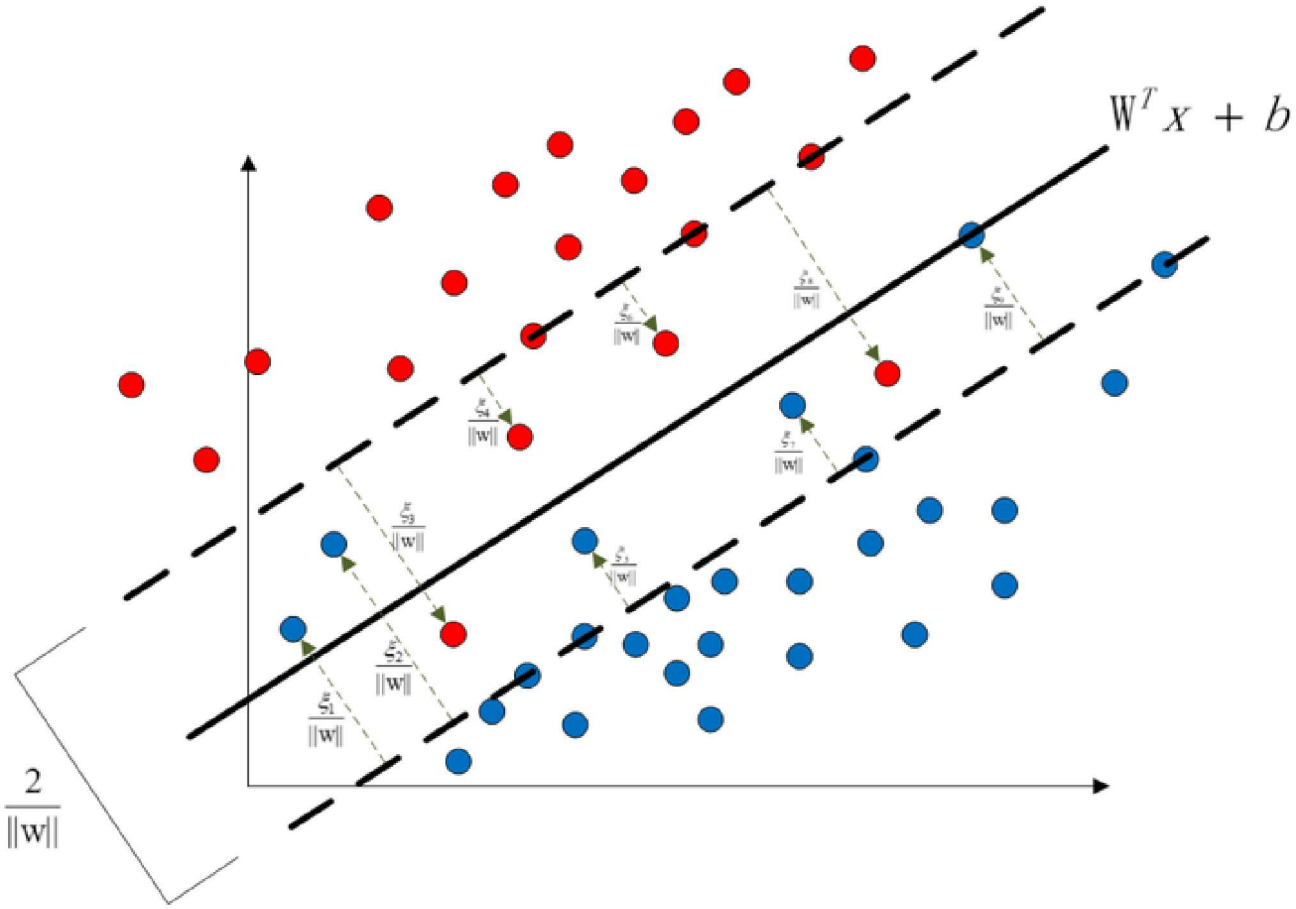
将其带入拉格朗日目标式中，消去w,b:

$$\begin{aligned} max_{\lambda}L(w,b,\lambda)&=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j y_i y_j x_j x_j+\sum_{i=1}^{n}\lambda_i \\ &\quad-\sum_{i=1}^{n}\lambda_i y_i\sum_{j=1}^{n}\lambda_j y_j x_j x_i \\ &=-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j y_i y_j x_i x_j+\sum_{i=1}^{n}\lambda_i \\ &\quad s.t.\ \lambda\geq 0 \\ &\quad\sum_{i=1}^{n}\lambda_i y_i=0 \end{aligned} \tag{1.13}$$

可以看到目标函数是一个二次型函数且不等式为仿射函数，该问题为一个凸二次规划问题存在解。可以使用cvxop包进行求解，也可以用下面的SMO进行求解。
得到最优的λ^*后可以利用公式(1.12)求得w^*。而求解b^*可以通过对在支持向量上的数据点进行求解$b^*=y_i-w^{*T}x_i$求得。

软间隔

硬间隔保证了线性可分情况下的支持向量的求解。然而数据是近似线性可分的时候，则需要给予一个松弛变量ϵ使得尽可能将样本分类正确的情况下保证支持向量距离超平面的距离尽可能的大。



由此上面硬间隔的公式转变为:

$$
\begin{aligned}
&argmin_{w,b,\epsilon} \frac{||w||^2}{2} + C\sum_{i=1}^{n}\epsilon_i \\
&s.t.\ \ y(w^T x + b) \geq 1 - \epsilon \\
&\epsilon \geq 0
\end{aligned}
\tag{2.1}
$$

该公式的理解可以看作当数据集中的点位于支持向量之内的时候，给予一个松弛向量使得该点可以被正确分类。如果是支持向量之外或者支持向量上面的点，则松弛向量的值为0。而C类似与正则化的作用。使得在求得最大的距离d和有多少个处于间隔内部的点之前取平衡。C越大，后半部分权重越高，优化更侧重于后半部分，所以位于间隔内部的点越少，当$C->\infty$时，则是硬间隔。

那么根据公式(2.1)仍然可以使用拉格朗日乘子法进行求解:

$$
\begin{aligned}
L(w,b,\epsilon,\alpha,\beta) = &\frac{||w||^2}{2} + C\sum_{i=1}^{n}\epsilon_i + \sum_{i=1}^{n}\alpha_i[1-\epsilon_i - y_i(w^T x_i + b)] \\
&- \sum_{i=1}^{n}\beta_i\epsilon_i \\
&s.t.\ \ \alpha \geq 0 \\
&\beta \geq 0
\end{aligned}
$$

同理转换为拉格朗日对偶式：

$$min_{w,b,\epsilon}max_{\alpha,\beta}L(w,b,\epsilon,\alpha,\beta) \tag{2.2}$$

$$\Rightarrow max_{\alpha,\beta}min_{w,b,\epsilon}L(w,b,\epsilon,\alpha,\beta) \tag{2.3}$$

同样对w,b,ϵ求偏导:

$$\begin{aligned}\frac{\partial L(w,b,\epsilon,\alpha,\beta)}{\partial w}=0\Rightarrow w=\sum_{i=1}^{n}\alpha_i y_i x_i\\ \frac{\partial L(w,b,\epsilon,\alpha,\beta)}{\partial b}=0\Rightarrow \alpha_i y_i=0\\ \frac{\partial L(w,b,\epsilon,\alpha,\beta)}{\partial \epsilon_i}=0\Rightarrow \alpha_i+\beta_i=C\end{aligned} \tag{2.4}$$

将公式(2.4)带入公式(2.3)对偶格式可得：

$$\begin{aligned}max_{\alpha,\beta}L(w,b,\epsilon,\alpha,\beta)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j\\ s.t.\ \sum_{i=1}^{n}\alpha_i y_i=0\\ 0\le\alpha\le C\end{aligned} \tag{2.5}$$

上式中的$\alpha\le C$是因为约束$\alpha+\beta=C$和约束$\beta\ge 0$。
而后使用SMO(Sequential Minimal Optimization)方式进行求解，每次固定一对α_i,α_j进行分析，取一对的原因在于需要满足公式(2.5)的约束条件。具体算法步骤如下:

- 选取一对需要更新的变量α_i,α_j
- 固定其他参数，求解公式(2.5)的最大值并更新α_i,α_j

假设选取α_1,α_2作为固定的参数，原式可以化为:

$$\begin{aligned}max_{\alpha,\beta}L(w,b,\epsilon,\alpha,\beta)=\alpha_1+\alpha_2-\frac{1}{2}\alpha_1^2y_1^2x_1^Tx_1\\ -\frac{1}{2}\alpha_2^2y_2^2x_2^Tx_2-\alpha_1\alpha_2y_1y_2x_1^Tx_2\\ -y_1\alpha_1\sum_{i=3}^{n}\alpha_iy_ix_i^Tx_1-y_2\alpha_2\sum_{i=3}^{n}\alpha_iy_ix_i^Tx_2\\ +Constant\\ s.t.\quad \alpha_1y_1+\alpha_2y_2=-\sum_{i=3}^{n}\alpha_iy_i=\eta\\ 0\le\alpha\le C\end{aligned} \tag{2.6}$$

其中Constant是与α_1,α_2无关的变量，可以当成常数项处理。
此时我们可以将α_2用α_1代替,并且两边同时乘y_2,由于$y_2y_2=1$,从而避免了除y_2带来的分数项。可得:

$$\alpha_2 = \eta y_2 - \alpha_1 y_1 y_2 \tag{2.7}$$

带入公式(2.6)可得:

$$\begin{gathered} max_{\alpha,\beta} L(w,b,\epsilon,\alpha,\beta) = \alpha_1 + \eta y_2 - \alpha_1 y_1 y_2 - \frac{1}{2}\alpha_1^2 y_1^2 x_1^T x_1 \\ -\frac{1}{2}(\eta y_2 - \alpha_1 y_1 y_2)^2 y_2^2 x_2^T x_2 - \alpha_1(\eta y_2 - \alpha_1 y_1 y_2) y_1 y_2 x_1^T x_2 \\ -y_1\alpha_1 \sum_{i=3}^{n} \alpha_i y_i x_i^T x_1 - y_2(\eta y_2 - \alpha_1 y_1 y_2) \sum_{i=3}^{n} \alpha_i y_i x_i^T x_2 \\ s.t.\ \ 0 \le \alpha_i \le C \end{gathered} \tag{2.8}$$

将公式(2.8)对α_1求偏导可得[$y_i^2 = 1$]:

$$\begin{gathered} \frac{\partial(L(w,b,\epsilon,\alpha,\beta))}{\partial(\alpha_1)} = (2x_1^T x_2 - x_1^T x_1 - x_2^T x_2)\alpha_1 \\ +1 - y_1 y_2 + \eta y_1 x_2^T x_2 - y_1 y_2 x_1^T x_2 \\ -y_1 \sum_{i=3}^{n} \alpha_i y_i x_i^T x_1 + y_1 \sum_{i=3}^{n} \alpha_i y_i x_i^T x_2 = 0 \\ s.t.\ \ 0 \le \alpha_i \le C \end{gathered} \tag{2.9}$$

而根据SVM的预测公式和公式(2.4)可得:

$$\begin{gathered} \hat{y}_i = w^T x_i + b \\ \Rightarrow \hat{y}_i = \sum_{j=1}^{n} \alpha_j y_j x_j^T x_i + b \end{gathered} \tag{2.10}$$

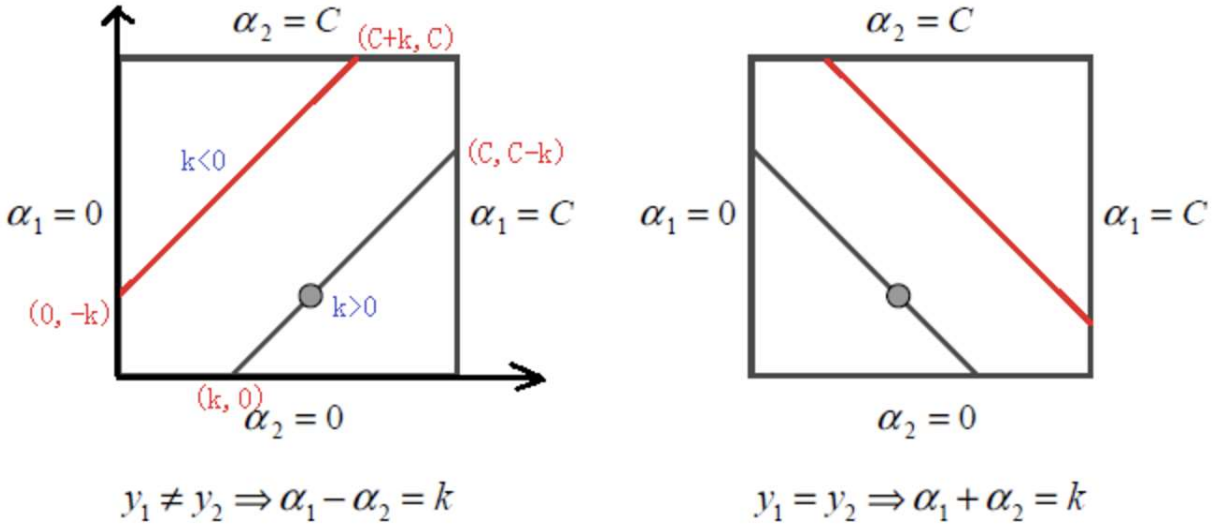
从而计算出公式(2.9)的求和项:

$$\begin{aligned} \sum_{i=3}^{n} \alpha_i y_i x_i^T x_1 &= \hat{y}_1 - b - \alpha_1 y_1 x_1^T x_1 - \alpha_2 y_2 x_2^T x_1 \\ \sum_{i=3}^{n} \alpha_i y_i x_i^T x_2 &= \hat{y}_2 - b - \alpha_1 y_1 x_1^T x_2 - \alpha_2 y_2 x_2^T x_2 \end{aligned} \tag{2.11}$$

将其带入公式(2.9)[此时的参数α_1, α_2为上一次迭代的参数，非本次迭代参数,所以分为old和new参数],并且将公式(2.9)中的1看成$y_1 y_1$可得:

$$\begin{gathered} \frac{\partial(L(w,b,\alpha))}{\partial(\alpha_1)} = (2x_1^T x_2 - x_1^T x_1 - x_2^T x_2)\alpha_1^{new} \\ -(2x_1^T x_2 - x_1^T x_1 - x_2^T x_2)\alpha_1^{old} + y_1[(y_1 - \hat{y}_1) - (y_2 - \hat{y}_2)] = 0 \\ \Rightarrow \alpha_1^{new} = \alpha_1^{old} - \frac{y_1[(y_1 - \hat{y}_1) - (y_2 - \hat{y}_2)]}{2x_1^T x_2 - x_1^T x_1 - x_2^T x_2} \\ s.t.\ \ 0 \le \alpha_i \le C \end{gathered} \tag{2.12}$$

由此就可以得到迭代表达式，同时根据公式(2.7)求解$\alpha_1^{new}, \alpha_2^{new}$，因为没办法确保$\alpha_1^{new}$满足约束，所以需要对$\alpha_1^{new}$进行修剪。



当y_1, y_2异号时，公式(2.7)可以转换为:

$$\alpha_1 - \alpha_2 = \eta \; or \; \alpha_2 - \alpha_1 = \eta \tag{2.13}$$

而公式(2.13)可以统一看成:

$$\alpha_1 - \alpha_2 = k(k = \eta \; or \; k = -\eta) \tag{2.14}$$

那么第一种$k < 0$的情况下，α_1的可行域为红线下方，第二种$k > 0$的情况下，α_1的可行域为红线上方[k其实可以看成直线的位移]，所以综上可以得出α_1的取值范围:

$$\begin{cases} L = max(0, \alpha_2^{old} - \alpha_1^{old}) \\ H = min(C, C + \alpha_2^{old} - \alpha_1^{old}) \end{cases} \tag{2.15}$$

当y_1, y_2同号时，可以统一看成:

$$\alpha_1 + \alpha_2 = k(k = \eta \; or \; k = -\eta) \tag{2.16}$$

同样可以得到α_2的可行范围:

$$\begin{cases} L = max(0, \alpha_1^{old} + \alpha_2^{old} - C)[k > C \Rightarrow \alpha_1^{old} + \alpha_2^{old} - C] \\ H = min(C, C + \alpha_2^{old} - \alpha_1^{old}) \end{cases} \tag{2.17}$$

由此可以得到α_1的值。

$$\alpha_1^{new} = \begin{cases} H, \alpha_1^{new,unclipped} > H \\ \alpha_1^{new,unclipped}, L \leq \alpha_1^{new,unclipped} \leq H \\ L, \alpha_1^{new,unclipped} < L \end{cases} \tag{2.18}$$

该情况是默认$2x_1^T x_2 - x_1^T x_1 - x_2^T x_2$小于0的情况，即原函数为开口向下的二次函数，从而可以取得最大极值点。而当$2x_1^T x_2 - x_1^T x_1 - x_2^T x_2 = 0$时，原函数为一次函数，则极值在边界选择；当$2x_1^T x_2 - x_1^T x_1 - x_2^T x_2 < 0$时，原函数为开口向上的二次函数，那么极值

点也在边界处取得。
再根据$\alpha_1^{new}+\alpha_2^{new}=\alpha_1^{old}+\alpha_2^{old}$求出$\alpha_2$。 根据KKT条件：

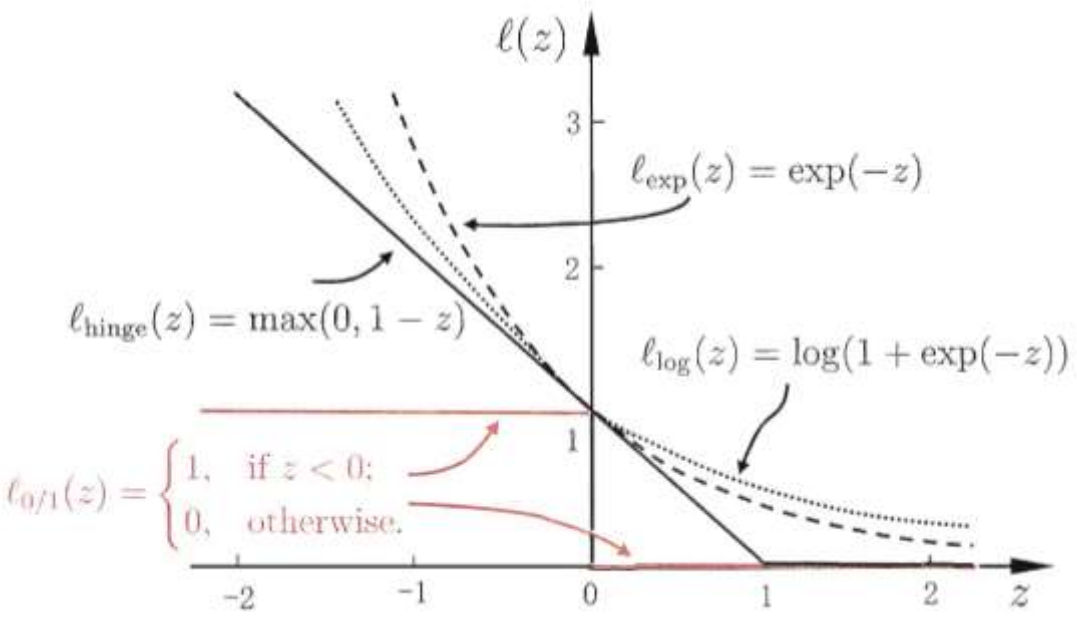
$$\begin{cases}\alpha_i \geq 0\\ \beta_i \geq 0\\ \epsilon_i \geq 0\\ \epsilon_i\beta_i = 0\\ y^i(w^Tx_i+b)-1+\epsilon_i \geq 0\\ \alpha_i[y^i(w^Tx_i+b)-1+\epsilon_i]=0\end{cases} \tag{2.19}$$

和$\alpha+\beta=C$,可以推出。

- 当$0<\alpha_1^{new}<C$时，$0<\beta_1<C$,此时$\epsilon_1=0$,从而$y^1(w^Tx_1+b)=1$，此时该点在支持向量上。
- 当$\alpha_1=0$时,$\beta_1=C$,从而$\epsilon_1=0$,那么此时$y^1(w^Tx_1+b)\geq 1$,此时该点可能是被正确分类或者是在支持向量上。
- 当$\alpha_1=C$时，$y^1(w^Tx_i+b)-1+\epsilon_1=0$,而$\beta_1=0$,从而$\epsilon_1\geq 0$,从而$y^1(w^Tx_1+b)\leq 1$。
  所以当$0<\alpha_1^{new}<C$,一定在支持向量上，那么由此根据$y_1(w^Tx_1+b)=1$可以求得b_1^{new}，同理可以求得b_2^{new}。此时b^{new}更新为:

$$b^{new}=\begin{cases}\frac{b_1^{new}+b_2^{new}}{2}, 0\leq\alpha_1,\alpha_2\leq C\\ b_1^{new}, 0\leq\alpha_1\leq C\\ b_2^{new}, 0\leq\alpha_2\leq C\end{cases} \tag{2.20}$$

软间隔与梯度下降



当考虑到软间隔时，可以将后半部分看作为hinge loss的损失函数$max(0,1-y_i\hat{y}_i)$[更适合

于大规模数据]。得出损失函数的计算公式为:

$$Loss(w,b) = \frac{w^T w}{2} + \frac{\epsilon}{N}\sum_{i=1}^{n} ReLu(1 - y_i(w^T x_i + b)) \tag{3.1}$$

并通过梯度下降法求解最优值。 流程为:

- 求解梯度:

$$\begin{cases} \frac{\partial(Loss(w,b))}{\partial(w)} = w + \frac{\epsilon}{N}\sum_{y_i(w^T x_i+b)<1}(-y_i x_i) \\ \frac{\partial(Loss(w,b))}{\partial(b)} = \frac{\epsilon}{N}\sum_{y_i(w^T x_i+b)<1}(-y_i) \end{cases} \tag{3.2}$$

- 梯度下降:

$$\begin{cases} w = w - \eta\frac{\partial(Loss(w,b))}{\partial(w)} \\ b = b - \eta\frac{\partial(Loss(w,b))}{\partial(b)} \end{cases} \tag{3.3}$$

拉格朗日对偶和KKT

对于有约束的优化问题而言，我们是要找到其在约束范围之内的极值点。那么也就可能是目标函数与约束条件等高线相切的地方，此时两者的梯度相同，也就是$\nabla f = -\lambda\nabla g$,其中$f(x), g(x)$分别为目标函数和约束条件[等式]，那也就是各个维度上的偏导都为0。所以可以写成$\phi = f + \lambda\nabla g$,$\partial\phi = 0$。
但是对于约束不等式，我们需要考虑到边界情况和内部情况，当处于边界时，处理方式与等式方式一致,也即$g = 0$；当在内部时，即为f的极值点，此时令$\lambda = 0$,$\phi = f$，没有g作为约束。所以合并边缘和内部两种情况就可以得到不等式约束情况下的最优解的必要条件为:

$$\begin{cases} \partial\phi = 0 \\ \lambda \geq 0 \\ \lambda g = 0[\text{两种情况必有一个为0}] \\ g \leq 0 \end{cases} \tag{4.1}$$

此时的约束条件即为KKT约束。
针对SVM的软间隔，我们首先进行拉格朗日乘子法操作得到如下:

$$min_{w,b}max_{\alpha,\beta}L(w,b,\epsilon,\alpha,\beta) = \frac{||w||^2}{2} + C\sum_{i=1}^{n}\epsilon_i + \sum_{i=1}^{n}\alpha_i[1 - \epsilon_i - y_i(w^T x_i + b)]$$
$$-\sum_{i=1}^{n}\beta_i\epsilon_i$$
$$s.t.\ \alpha \geq 0$$
$$\beta \geq 0$$
$$\partial(L) = 0$$

_

$$
\begin{aligned}
1-\epsilon_i-y_i(w^Tx_i+b)\leq 0\\
\epsilon\geq 0
\end{aligned}
\tag{4.2}
$$

但是这种情况下对w求解的时候，需要使用w对数据集做内积，会导致速度变慢，并且SVM只依赖与支持向量上的数据点，也就是说α很多都会为0，所以转换成对偶问题更容易求解。
对偶问题也就是从不同的角度来看待该问题的最优值，对于$min_{w,b}max_{\alpha,\beta}$则是将求出$\alpha,\beta$的最优表示，并固定从而再求解$w,b$最优解，而其对偶形式则是反之。
令:

$$
f(w,b,\epsilon)=\frac{||w||^2}{2}+C\sum_{i=1}^{n}\epsilon_i \tag{4.3}
$$

而对于对偶问题的上界：

$$
\begin{aligned}
&max_{\alpha,\beta}min_{w,b}L(w,b,\epsilon,\alpha,\beta)\\
&=max_{\alpha,\beta}inf_{w,b,\epsilon}L(w,b,\epsilon,\alpha,\beta)\\
&\leq f(w^*,b^*,\epsilon^*)+\sum_{i=1}^{n}\alpha_i[1-\epsilon_i^*-y_i(w^Tx_i+b^*)]-\sum_{i=1}^{n}\beta_i\epsilon_i^*\\
&\leq f(w^*,b^*,\epsilon^*)\\
&=p^*
\end{aligned}
\tag{4.4}
$$

其中，w^*,b^*,ϵ^*代表原问题的最优解也就是最小值，在w^*,b^*,ϵ^*符合约束的情况下,α,β的值越大，等式的值越小，所以显然是满足的。
原问题的下界:

$$
\begin{aligned}
&min_{w,b,\epsilon}max_{\alpha,\beta}L(w,b,\epsilon,\alpha,\beta)\\
&=min_{w,b,\epsilon}sup_{\alpha,\beta}L(w,b,\epsilon,\alpha,\beta)\\
&=\begin{cases}f(w,b,\epsilon),\alpha=0,\beta=0\\\infty,otherwise\end{cases}
\end{aligned}
$$

所以显然有:

$$
sup_{\alpha,\beta}inf_{w,b}L(w,b,\epsilon,\alpha,\beta)\leq inf_{w,b,\epsilon}sup_{\alpha,\beta}L(w,b,\epsilon,\alpha,\beta) \tag{4.5}
$$

那么对于强对偶而言，存在:

$$
sup_{\alpha,\beta}inf_{w,b}L(w,b,\epsilon,\alpha,\beta)=f(w^*,b^*,\epsilon^*)=inf_{w,b,\epsilon}sup_{\alpha,\beta}L(w,b,\epsilon,\alpha,\beta)
$$

也就是后面的部分为0，所以也就是满足KKT条件中的$\lambda g=0$。