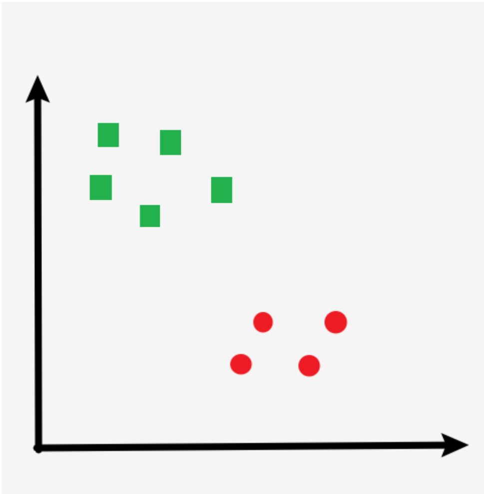
Support Vector Machine

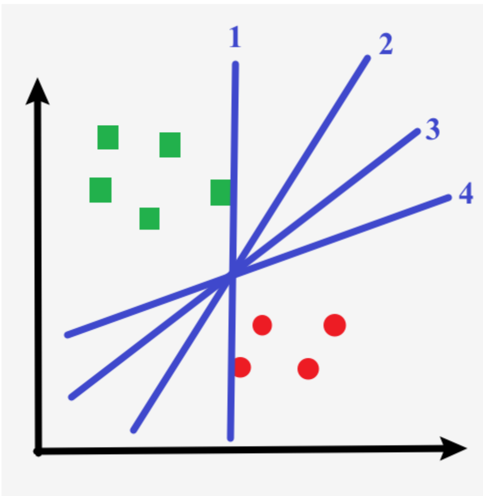
> *Baicheng Chen*

支持向量机是一个经典的二分类模型，由Cortes和Vapnik于1995年提出。它区别于感知机，感知机是通过错误驱动的方式来确定一个可行的决策边界（minimize loss function => misclassification points，有无穷多个），而支持向量机则是选出“最优”的一个决策边界。这里的“最优”如何定义呢？SVM考虑了最大间隔（两个类别之间）。

这里，我们首先考虑一个简单的线性可分的二分类问题。

> 假设有样本点$\{(x_i,y_i)\}_i^N, x_i \in R^p, y \in \{-1,1\}$。
>
> 

现在，若我们想将绿色样本点与红色样本点分开，有无数条决策边界可以做到。下图中的1、2、3、4这四条线都可以完美的将这两种样本点进行分类。但我们可以发现1号线明显鲁棒性较差，如果样本点变多，它的泛化误差较大，并不能算一个足够“好”的边界。而我们需要的边界是一条鲁棒性好，可泛化的决策边界。



于是，SVM变提出了“最大间隔”的概念。从上图中看，最大间隔便是两个类别最“中间”的那条线（即为3号线），它保持了到两个分类区域的距离最大，我们认为“最大间隔”的决策边界就是“最好”的决策边界。而这就引出了这篇文章的第一部分，也就是线性可分支持向量机，也称之为最大间隔分类器（Hard-margin SVM）。

Hard-Margin SVM

现在，我们从数学的角度来定义“最大间隔”这个概念。首先，我们可以将决策边界看作一个超平面，用公式表示为$0=w^Tx+b$，这个超平面将n维空间分割为两半，其中，法向量w指向的那一半定义为正空间 ($\forall x^+, w^Tx^++b>0$)，反之另一半则为负空间，可知：

$$\max\ margin(w,b)$$
$$s.t.\begin{cases}w^Tx_i+b>0 & ,y_i=+1\\ w^Tx_i+b<0 & ,y_i=-1\end{cases}$$

这个公式还是不够简洁，我们尝试将constrains合并为一个条件，即：

$$\max\ margin(w,b)$$
$$s.t.\ \ y_i(w^Tx_i+b)>0\ ,\ \text{for}\ \forall i=1,2,\ldots,N$$

这个条件与上面两个条件是等价的，因为同号相乘一定为正（y_i与w^Tx_i+b始终同号）。

接下来，我们来详细推导$margin(w,b)$这个函数，即解释如何用数学形式来表示我们在前文所提到的“间隔”。简单来讲，“间隔”就是分类区域到决策边界的距离，在这里，我们认为N个样本点到决策边界的距离中最小的那个即为“间隔”。这样，如果我们通过最大化这个最小距离找到了一个最优的决策边界，它对于整个分类区域来说也一定是最优的。

假设这个最小距离由样本点(x_i,y_i)提供，下面计算(x_i,y_i)到决策边界$y=w^Tx+b$的距离：

$$\text{对于点}\boldsymbol{x_0},\text{设其在超平面}0=\boldsymbol{w^Tx}+b\text{的投影为}\boldsymbol{x_1},\text{则有}\boldsymbol{w^Tx_1}+b=0\text{。}$$
$$\text{因为法向量}\boldsymbol{w}\text{垂直于超平面},\vec{\boldsymbol{x_1x_0}}\text{也垂直于超平面},\text{故}\boldsymbol{w}//\vec{\boldsymbol{x_1x_0}},\text{则：}$$
$$|\boldsymbol{w}\cdot\vec{\boldsymbol{x_1x_0}}|=|||\boldsymbol{w}||\cdot\cos\pi\cdot||\vec{\boldsymbol{x_1x_0}}|||=||\boldsymbol{w}||\cdot||\vec{\boldsymbol{x_1x_0}}||=||\boldsymbol{w}||\cdot r,\quad r\text{为距离};$$
$$\begin{aligned}\boldsymbol{w}\cdot\vec{\boldsymbol{x_1x_0}}&=w_1(x_1^0-x_1^1)+w_2(x_2^0-x_2^1)+\ldots+w_n(x_n^0-x_n^1)\\&=w_1x_1^0+w_2x_2^0+\ldots+w_nx_n^0-(w_1x_1^1+w_2x_2^1+\ldots+w_nx_n^1)\\&=\boldsymbol{w^Tx_0}-\boldsymbol{w^Tx_1}\\&=\boldsymbol{w^Tx_0}+b\qquad(\because\ \boldsymbol{w^Tx_1}+b=0).\end{aligned}$$

所以有，

$$\begin{aligned}|\boldsymbol{w^Tx_0}+b|&=||\boldsymbol{w}\cdot\vec{\boldsymbol{x_1x_0}}||\\&=||\boldsymbol{w}||\cdot r\\\Rightarrow r&=\frac{|\boldsymbol{w^Tx_i}+b|}{||\boldsymbol{w}||}\end{aligned}$$

有了这个距离，我们就可以表示$margin(w,b)$函数了，注意：margin是样本点到决策边界的最小距离！

$$\begin{aligned}margin(w,b)&=\min_{w,b,x_i}distance(w,b,x_i)\\&=\min_{w,b,x_i}\frac{|w^Tx_i+b|}{||w||}\end{aligned}$$

将导出的$margin(w,b)$公式带回到我们最开始写出的的优化问题中：

$$\max_{w,b}\ \min_{x_i}\frac{|w^Tx_i+b|}{||w||}$$
$$s.t.\ \ y_i(w^Tx_i+b)>0$$

因为我们有constrain $y_i(w^Tx_i+b)>0$，则可以将y_i看作w^Tx_i+b的绝对值符号，将其替换到objective function中：

$$
\max_{w,b} \ \min_{x_i} \frac{y_i(w^T x_i + b)}{||w||}
$$
$$
s.t. \quad y_i(w^T x_i + b) > 0
$$

因为$y_i(w^T x_i + b)$始终大于0，我们假设$\exists \gamma > 0, s.t. \ \min_{x_i, y_i} y_i(w^T x_i + b) = \gamma$。

则我们可以将objective function转化为：

$$
\begin{aligned}
\max_{w,b} \ \min_{x_i} \frac{y_i(w^T x_i + b)}{||w||} &= \max_{w,b} \frac{1}{||w||} \min_{x_i} y_i(w^T x_i + b) \\
&= \max_{w,b} \frac{1}{||w||} \gamma
\end{aligned}
$$

我们知道，超平面方程不唯一，即当我们等倍缩放w和b时，所得的新超平面与原超平面相同，故而，γ有无数种可能的取值。所以，我们添加一个约束，即令$\gamma = 1$，以此来限制到唯一超平面。优化问题变成了：

$$
\max_{w,b} \ \frac{1}{||w||}
$$
$$
s.t. \quad \min \ y_i(w^T x_i + b) = 1
$$

等价于，

$$
\max_{w,b} \ \frac{1}{||w||}
$$
$$
s.t. \quad y_i(w^T x_i + b) \geq 1
$$

等价于，

$$
\min_{w,b} \ ||w||
$$
$$
s.t. \quad y_i(w^T x_i + b) \geq 1
$$

显然，这个优化问题可以被转化成一个QP问题，即：

$$
\min_{w,b} \ \frac{1}{2} w^T w
$$
$$
s.t. \quad y_i(w^T x_i + b) \geq 1, \ \ for \ \ \forall i = 1, \ldots, N
$$

等价于，

$$
\min_{w,b} \ \frac{1}{2} w^T w
$$
$$
s.t. \quad 1 - y_i(w^T x_i + b) \leq 0, \ \ for \ \ \forall i = 1, \ldots, N
$$

将这个QP问题看作一个约束优化问题，而上式则为该约束优化问题的原问题 (primal problem)。

下面，我们引入拉格朗日函数，

$$
L(w, b, \lambda) = \frac{1}{2} w^T w + \sum_{i=1}^{N} \lambda_i (1 - y_i(w^T x_i + b)), \ \ \lambda_i \geq 0
$$

通过拉格朗日函数的引入，我们可以找出原问题的无约束形式：

$$
\min_{w,b} \ \max_{\lambda} L(w, b, \lambda)
$$
$$
s.t. \quad \lambda_i \geq 0
$$

下面对上述优化问题中的objective function进行解释：

$$
\begin{gathered}
对于L(w,b,\lambda)=\frac{1}{2}w^Tw+\sum_{i=1}^{N}\lambda_i(1-y_i(w^Tx_i+b)),\\
我们在左右两边同时取max, 即：\\
\max_{\lambda} L(w,b,\lambda)=\frac{1}{2}w^Tw+\max_{\lambda}(\sum_{i=1}^{N}\lambda_i(1-y_i(w^Tx_i+b)))\\
在该式中, \lambda_i\geq 0, 1-y_i(w^Tx_i+b)\leq 0, 我们可以得到,\\
\lambda_i(1-y_i(w^Tx_i+b))\leq 0\\
即\max_{\lambda}(\sum_{i=1}^{N}\lambda_i(1-y_i(w^Tx_i+b)))=0\\
由此我们可以推出,\\
\begin{aligned}
\frac{1}{2}w^Tw&=\max_{\lambda} L(w,b,\lambda)-\max_{\lambda}(\sum_{i=1}^{N}\lambda_i(1-y_i(w^Tx_i+b)))\\
&=\max_{\lambda} L(w,b,\lambda)-0\\
&=\max_{\lambda} L(w,b,\lambda)
\end{aligned}
\end{gathered}
$$

于是，我们就可以得出原问题的对偶问题 (dual problem)。因为该优化问题为凸二次优化问题，且约束条件为仿射函数，满足放松Slater条件，故原问题与对偶问题满足强对偶关系（convex + slater ⇒ strong duality）。

$$
\begin{gathered}
\max_{\lambda}\ \min_{w,b} L(w,b,\lambda)\\
s.t.\quad \lambda_i\geq 0
\end{gathered}
$$

下面，我们着手解决这个无约束优化问题。

- 对b求导，

$$
\begin{aligned}
\frac{\partial L}{\partial b}&=\frac{\partial}{\partial b}[\sum_{i=1}^{N}\lambda_i-\sum_{i=1}^{N}\lambda_iy_i(w^Tx_i+b)]\\
&=\frac{\partial}{\partial b}[-\sum_{i=1}^{N}\lambda_iy_ib]\\
&=-\sum_{i=1}^{N}\lambda_iy_i=0
\end{aligned}
$$

- 将上述结果带入原式化简，

$$
\begin{aligned}
L(w,b,\lambda)&=\frac{1}{2}w^Tw+\sum_{i=1}^{N}\lambda_i(1-y_i(w^Tx_i+b))\\
&=\frac{1}{2}w^Tw+\sum_{i=1}^{N}\lambda_i-\sum_{i=1}^{N}\lambda_iy_i(w^Tx_i+b)\\
&=\frac{1}{2}w^Tw+\sum_{i=1}^{N}\lambda_i-\sum_{i=1}^{N}\lambda_iy_iw^Tx_i-\sum_{i=1}^{N}\lambda_iy_ib\\
&=\frac{1}{2}w^Tw+\sum_{i=1}^{N}\lambda_i-\sum_{i=1}^{N}\lambda_iy_iw^Tx_i\quad(\because\ -\sum_{i=1}^{N}\lambda_iy_i=0且b为常数)
\end{aligned}
$$

- 对w求导，

$$\frac{\partial L}{\partial w}=\frac{1}{2}\cdot 2\cdot w-\sum_{i=1}^{N}\lambda_i y_i x_i=0$$

$$\Rightarrow w^*=\sum_{i=1}^{N}\lambda_i y_i x_i$$

- 将w^*代入objective function，

$$\begin{aligned}\min_{w,b} L(w,b,\lambda)&=\frac{1}{2}(\sum_{i=1}^{N}\lambda_i y_i x_i)^T(\sum_{j=1}^{N}\lambda_j y_j x_j)-\sum_{i=1}^{N}\lambda_i y_i(\sum_{j=1}^{N}\lambda_j y_j x_j)^T x_i+\sum_{i=1}^{N}\lambda_i\\&=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j x_i^T x_j-\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j x_j^T x_i+\sum_{i=1}^{N}\lambda_i\\&=-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j x_i^T x_j+\sum_{i=1}^{N}\lambda_i\end{aligned}$$

- 由上述推导，对偶问题可化为，

$$\begin{gathered}\max_{\lambda}\ -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j x_i^T x_j+\sum_{i=1}^{N}\lambda_i\\ s.t.\quad \lambda_i\geq 0\end{gathered}$$

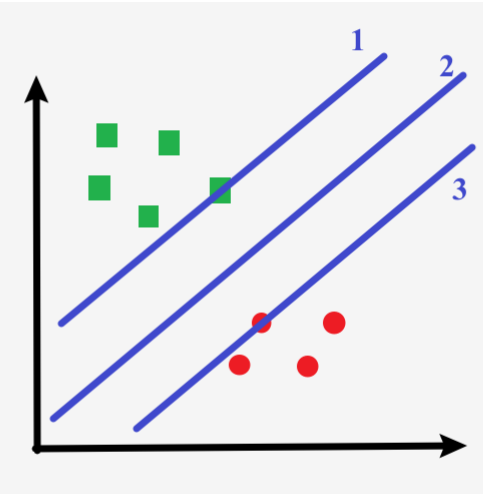
- 等价于，

$$\begin{gathered}\min_{\lambda}\ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j x_i^T x_j+\sum_{i=1}^{N}\lambda_i\\ s.t.\quad \lambda_i\geq 0\end{gathered}$$

因为原问题与对偶问题具有强对偶关系⇔满足KKT条件:

$$KKT\begin{cases}\text{可行条件}:\begin{cases}\lambda_i\geq 0\\1-y_i(w^Tx+b)\leq 0\end{cases}\\\text{互补松弛条件}:\lambda_i(1-y_i(w^Tx+b))=0\\\text{梯度为0}:\frac{\partial L}{\partial w}=0,\ \frac{\partial L}{\partial b}=0\end{cases}$$

现在，我们再来看图示。中间蓝线便是我们要找出的最优决策边界$w^Tx+b=0$（线2），根据前文所述，我们将γ限制在1，故而，经过距离最优决策边界最近的两个样本点的直线可以分别表示为$w^Tx+b=1$（线1）和$w^Tx+b=-1$（线3）。



以$w^Tx+b=1$为例（绿色区域），由可行条件$1-y_i(w^Tx+b)\leq 0$可知，此时为取等条件，即$1-y_i(w^Tx+b)=0$。由互补松弛条件可知，λ可取任意值，满足$\lambda_i\geq 0$。下面我们看到线1上方的区域，这里仍然存在四个绿色样本点，对于它们来说，$w^Tx+b>1$，所以$1-y_i(w^Tx+b)<0$，故而由互补松弛条件可知，λ_i一定等于零。所以，对于λ_i的值有贡献的样本点仅存在于$w^Tx+b=1$和$w^Tx+b=-1$两条线上，我们称这些点为支持向量。

下面，我们做最后的结果推导。根据KKT条件以及前文推导，我们已知：

$$w^*=\sum_{i=1}^{N}\lambda_i y_i x_i$$

从1和3两条线入手，进行b^*的推导：

$$\begin{gathered}\text{我们假设}\exists(x_k,y_k),\ s.t.\ 1-y_k(w^Tx_k+b)=0\Rightarrow y_k(w^Tx_k+b)=1\\ \text{因为}y_k=\pm1,\text{等式两边同乘}y_k\text{仍然成立, 则：}\\ y_k^2(w^Tx_k+b)=y_k\\ \Rightarrow b^*=y_k-w^Tx_k=y_k-\sum_{i=1}^{N}\lambda_i y_i x_i^T x_k\end{gathered}$$

至此，我们便完成了Hard-margin SVM的全部推导，该判别模型最终可表示为一下分类决策函数：

$$\begin{gathered}f(x)=sign(w^{*T}x+b^*)\\ w^*=\sum_{i=1}^{N}\lambda_i y_i x_i\\ b^*=y_k-w^Tx_k=y_k-\sum_{i=1}^{N}\lambda_i y_i x_i^T x_k\end{gathered}$$

Soft-Margin SVM

在引入部分我们便给Hard-margin SVM添加了一个限制条件，即我们的二分类问题一定是线性可分的。那么如果我们拿到的二分类问题是线性不可分问题呢？这里，我们就引入了Soft-margin SVM，它相当于在线性可分问题的基础上允许一些错误样本点（整体分类问题是线性不可分的）。

回顾我们在推导Hard-margin SVM时的原问题，即：

$$\begin{gathered}\min_{w,b}\ \frac{1}{2}w^Tw\\ s.t.\quad 1-y_i(w^Tx_i+b)\leq 0,\ \ for\ \ \forall i=1,\dots,N\end{gathered}$$

Soft-margin SVM的基本思想是引入一个loss function用来表示错误分类损失，并将其一起最小化，获得尽可能优的决策边界，我们可以写作：

$$\begin{gathered}\min_{w,b}\ \frac{1}{2}w^Tw+loss\ function\\ s.t.\quad 1-y_i(w^Tx_i+b)\leq 0,\ \ for\ \ \forall i=1,\dots,N\end{gathered}$$

下面我们来探讨合适的loss function。

首先，最为简单的loss function便是错分的样本点数量，数学公式可以写作：

$$loss=\sum_{i=1}^{N}I\{y_i(w^Tx_i+b)<1\}$$

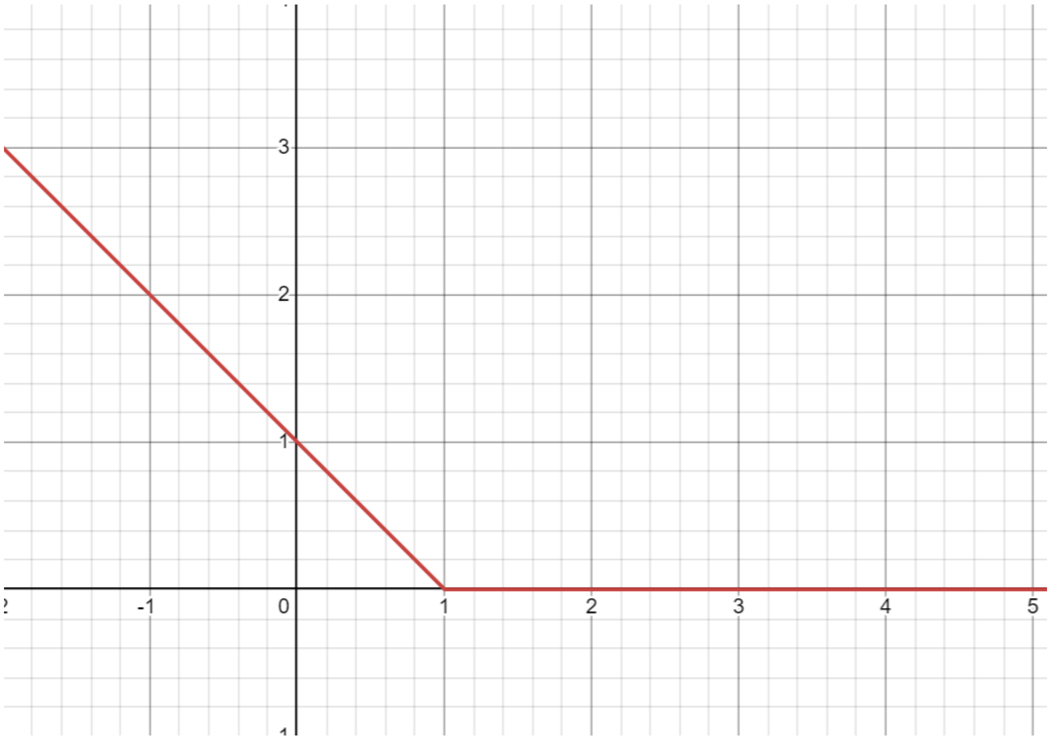
但此时的问题是这个函数并不连续，这样的优化问题很难求解，我们选择考虑别的loss function。

既然数量不行，距离行不行呢？这里，我们便引入了Hinge Loss。我们做如下考虑，假设我们已分类了样本点(x_i, y_i)，如果该样本点满足约束条件（函数间隔（确信度）），即$y_i(w^Tx_i+b) \geq 1$，令$loss=0$；相反，如果该样本点不满足约束条件，即$y_i(w^Tx_i+b) < 1$，令$loss = 1-y_i(w^Tx_i+b)$。

合并成一个连续的loss function，

$$loss = max\{0, 1-y_i(w^Tx_i+b)\}$$

设$z=y_i(w^Tx_i+b)$，$loss = max\{0, 1-z\}$，画出图像如下（合页形状）：



我们发现，这个损失函数是连续可微的，可以使用，于是写出Soft-margin SVM的优化问题如下：

$$\min_{w,b} \frac{1}{2}w^Tw + C\sum_{i=1}^{N}\max\{0, 1-y_i(w^Tx_i+b)\}$$
$$s.t. \quad y_i(w^Tx_i+b) \geq 1, \ \ for \ \ \forall i = 1, \ldots, N$$

这里，$C>0$是一个超参数，我们称为惩罚参数，由实际应用问题决定。C越大，对错误分离的惩罚越大。

为进一步消除max函数，我们引入一个松弛变量$\xi_i = 1-y_i(w^Tx_i+b)$，$\xi_i \geq 0$。上式可简化为，

$$\min_{w,b} \frac{1}{2}w^Tw + C\sum_{i=1}^{N}\xi_i$$
$$s.t. \quad y_i(w^Tx_i+b) \geq 1, \ \ for \ \ \forall i = 1, \ldots, N$$
$$\xi_i \geq 0$$

接下来，使用约束优化问题的方法同Hard-margin SVM一样正常推导即可。

Kernel Method

前文中，我们讨论了对于线性可分二分类问题、线性不可分二分类问题的解决方法，现在，如果二分类问题变成了非线性问题，我们该怎么解决呢？对于这种问题，我们便要引入核方法。因篇幅有限，我会在下一篇文章详细讨论核方法的原理、核函数及正定核等内容。