西瓜书阅读笔记之SVM

一、SVM的基本型

首先了解一下SVM是干什么的，SVM用来分类样本的。SVM的目标是寻找到一个最佳的超平面使得（超平面可能有很多，最佳超平面和支持向量之间的间隔最可能大）。划分超平面可以通过线性方程来描述：

$$w^T \boldsymbol{x} + b = 0$$

$ w=(w_1;w_2;...;w_d) $为法向量，决定了超平面的方向，$b$为位移项，据定了超平面与原点之间的距离，$w,b$是确定超平面的两个唯一的重要因素，记为$(w,b)$，样本空间中任意点$x$到超平面$(w,b)$的距离可写为：

$$r = \frac{|w^T \boldsymbol{x} + b|}{||w||}$$

假设超平面(w,b)能够训练样本正确分类，即对于$(x_i, y_i) \in D$，若$y_i = +1$，则有$w^Tx_i+b>0$；若$y_i=-1$，则有$w^Tx_i+b<0$，根据切比雪夫不等于，我们一定可以找到一个ϵ 满足$w^Tx_i+b≤\epsilon<0$,$\epsilon < 0$

$$w^T \boldsymbol{x}_i + b \leq \epsilon < 0$$

$$w^T \boldsymbol{x}_i + b \leq -1,\ y_i = -1$$

$$\text{两边同时除以 } -\epsilon, \frac{-w^T x_i}{-\epsilon} - \frac{b}{\epsilon} \leq -1$$

$$\text{同理，} w^T \boldsymbol{x}_i + b \geq +1,\ y_i = +1$$

$$\begin{cases} w^T \boldsymbol{x}_i + b \leq -1, & y_i = -1 \\ w^T \boldsymbol{x}_i + b \geq +1, & y_i = +1 \end{cases} \tag{6.3}$$

距离超平面最近的这几个样本点使得（1）（2）成立，它们被称为“支持向量”(support vector)，图6.2中表示为类似于正负电荷的圆圈，两个异类支持向量到超平面的距离之和为γ，γ也被称为“间隔”(margin)

$$\gamma = \frac{2}{||w||} = \frac{1}{||w||} + \frac{1}{||w||}$$

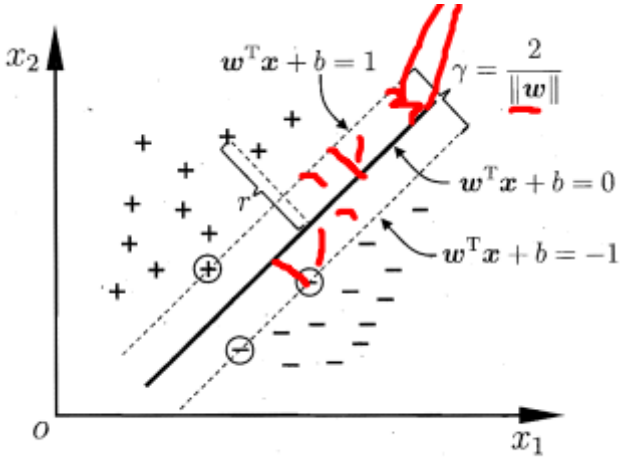

图 6.2 支持向量与间隔

欲找到具有"最大间隔（maximum margin）"的划分超平面，也就是要找到能满足约束条件（1）（2）的参数w和b，使得γ最大，即

$$\max_{w,b} \frac{2}{\|w\|} \qquad (6.5)$$

$$s.t. \quad y_i(w^T \boldsymbol{x}_i + b) \geq 1, i = 1, 2, \ldots, m$$

显然，为了最大化间隔，仅需**最大化**$\frac{1}{||w||^2}=(||w||^2)^-1$，这等价于$||w||^2$最小化，于是上式可以被重写为：

$$\min_{w,b} \frac{1}{2}\|w\|^2 = f(w)$$
$$s.t. \quad y_i(w^T \boldsymbol{x}_i + b) \geq 1, i = 1, 2, \ldots, m \qquad (6.6)$$

这就是支持向量机（Support Vector Machine，简称SVM）的基本型。

二、对偶问题

目标和约束条件都有了，现在就是要解式（6.6）来得到最大间隔划分超平面所对应的模型

当我们要求一个函数 min f(x) 的时候，如果 f(x) 可导，我们通过是通过求 f(x) 的导数来得。

但是如果函数 f(x) 带约束条件，如（6.6），那么问题就开始变复杂了。**凸优化的目标就是解决带约束条件函数的极值问题[https://zhuanlan.zhihu.com/p/89292221?from_voters_page=true]。**

凸优化解决的通用模型是：

$$\begin{cases} \min & f(x) \\ \text{s.t.} & g_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, n \end{cases}$$

不是所有的极值问题都可以适用的**凸优化理论，它必须满足以下条件**：

1、目标函数 f(x) 为凸函数 (二阶可导，且二阶导数 > 0)

2、不等式约束函数 g(x) 为凸函数

3、等式约束函数 h(x) 为仿射函数 (仿射函数=导数最高阶数为1)

只有同时满足以上3个条件，才属于凸优化的范畴。

凸函数：定义域为凸集，凸集几何意义表示为：如果集合中任意2个元素连线上的点也在集合C中，则C为凸集

在(6.6)中，w和b是模型参数，目标函数$\min_{w, b} \frac{1}{2}{\|\mathcal{w}\|}^2$是二次函数，凸函数。约束条件$s.t. y_i(w^Tx_i+b)≥1$是仿射函数，所以SVM本身十一个凸二次规划（convex quadratic programming）问题

$$g(w, b) = 1 - y_i \left(w^T \boldsymbol{x}_i + b\right)$$

对于凸优化的通用模型，由于其带有约束条件，很难处理，因此我们会考虑怎么用一个式子来表述那个通用模型呢？拉格朗日乘子法就是一个很好的方法，使用拉格朗日乘子法可以得到其"对偶问题"(dual problem)，具体来说，对式(6.6)的每个约束条件添加拉格朗日乘子$\alpha_i≥0$（如果α_i与$g(w,b)$同号，最大值就是∞，没有意义），则该问题的拉格朗日函数可以些写为：

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \alpha_i \left(1 - y_i \left(w^T \boldsymbol{x}_i + b\right)\right) \qquad (6.8)$$

上述问题的等价问题为：

$$\min-\max L(w,b,\alpha)$$

其对偶问题为：

$$\max-\min L(w,b,\alpha)$$

先求**min(L)**，$\alpha=\alpha_1;\alpha_2;...;\alpha_m$。令$L(w, b, \alpha)$对$w$和$b$的偏导为零可得

$$\boldsymbol{w}=\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i \tag{6.9}$$

$$0=\sum_{i=1}^{m}\alpha_i y_i \tag{6.10}$$

将(6.9)代入(6.8):

$$\begin{aligned}
&=\frac{1}{2}\sum_{i=1}^{m}(\alpha_i y_i \boldsymbol{x}_i)^2+\sum_{i=1}^{m}\alpha_i\left(1-y_j\left(\sum_{j=1}^{m}\alpha_i y_i \boldsymbol{x}_i^T\boldsymbol{x}_j+b\right)\right)\\
&=\frac{1}{2}\sum_{i=1}^{m}(\alpha_i y_i \boldsymbol{x}_i)^2+\sum_{i=1}^{m}\alpha_i-\alpha_i y_j\sum_{j=1}^{m}\alpha_i y_i \boldsymbol{x}_i^T\boldsymbol{x}_j+b\\
&\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j-\alpha_i y_j\sum_{j=1}^{m}b \qquad \text{代入}6.10\\
&\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j
\end{aligned}$$

于是

$$\begin{cases}\max_\alpha \sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j\\ s.t.\quad 0=\sum_{i=1}^{m}\alpha_i y_i,\quad \alpha\geq 0,\ i=1,2,\dots,m\end{cases} \tag{6.11}$$

解出α后，求出w 和b即可得到模型

$$\begin{aligned}
f(\boldsymbol{x})&=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b\\
&=\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}+b.
\end{aligned} \tag{6.12}$$

从对偶问题(6.11)解出的α_i是式(6.8)中的拉格朗日乘子，它恰对应训练样本$(\boldsymbol{x}_i,y_i)$。如果能满足KKT条件，**原始问题=对偶问题。**

模型最终转化成参数为α的函数，求最大值问题，也就是有w,b两个参数变成一个参数

于是，对于任意训练样本$(\boldsymbol{x}_i,y_i)$，总$\alpha_i =0 $有或$y_if(\boldsymbol{x_i}=1)$。若，则该样本将不会在（6.12）的求和中出现，也就不会对$f(\boldsymbol{x})$有任何影响；若α_i>0，则必有$y_if(\boldsymbol{x_i}=1)$，所对应的样本点位于最大间隔边界上，是一个支持向量。这显示出支持向量机的一个重要特性：训练完成后，大部分的训练样本都不需要保留，最终模型仅与支持向量有关。

那么现在的问题重点又转到求解式子（6.11），不难发现这是一个二次规划问题，可使用通过的二次规划算法来求解，该问题的规模正比于训练样本数，这会在实际任务中造成很大的开销，为了避免这个问题，人们通过利用问题本身的特性，提出了很多高效的算法，SMO(Sequential Minimal Optimization)是一个代表，SMO的思想是先固定α_i之外的所有参数，然后求α_i上的极值。

三、核函数

在前面的公式推理前提是假设样本线性可分，即存在一个划分超平面能将训练样本正确分类，然而在现实生活中，原始样本空间也许并不存在，对于这样的问题，可将样本从原始空间映射到一个更高维的特征空间，使得样本在这个特征空间内线性可分。

我们希望样本在特征空间内线性可分，因此特征空间的好坏对支持向量机的性能至关重要。需要注意的是，在不知道特征映射的形式时，我们并不知道什么样的核函数最合适，而核函数也仅是隐式地定义了这个特征空间。于是，“核函数选择”成为支持向量机的最大变数。若核函数选择不佳，很可能导致性能不佳！！

表 6.1 常用核函数

名称	表达式	参数
线性核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j$	
多项式核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j)^d$	$d \geqslant 1$ 为多项式的次数
高斯核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right)$	$\sigma > 0$ 为高斯核的带宽(width)
拉普拉斯核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|}{\sigma}\right)$	$\sigma > 0$
Sigmoid 核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(\beta \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j + \theta)$	tanh 为双曲正切函数, $\beta > 0, \theta < 0$

四、软间隔与正则化

现实生活中往往很难确定合适的核函数是的训练样本在特征空间中线性可分；退一步说，即便恰好找到了某个核函数使训练样本在特征空间中线性可分，也很难推断这个貌似线性可分的结果不是由于过拟合所造成的。

缓解这个问题的一个办法是**允许支持向量机在一些样本上的出错。**为此，要引入“软间隔”（soft margin）的概念，如下图6.4所示：

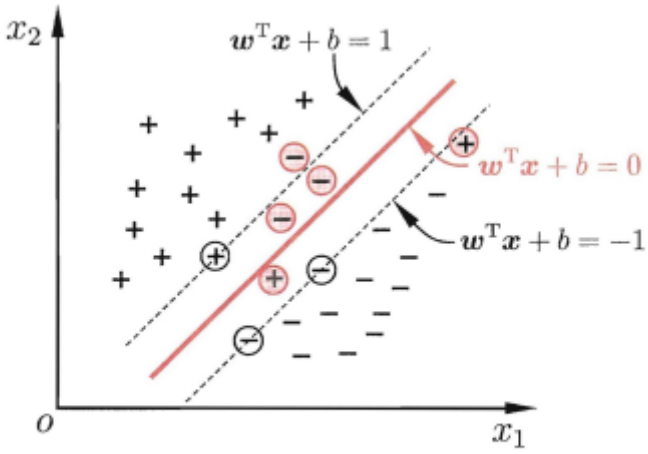

图 6.4 软间隔示意图. 红色圈出了一些不满足约束的样本.

前面介绍的支持向量机形式时保证所有样本正确划分，这称为“硬间隔”（hard margin），而软间隔是允许某些样本不满足条件：

当然，在最大化间隔的同时，不满足约束的样本要尽可能少，于是，优化目标可写为：

其中$C>0$是一个常数，$\ell_{0 / 1}$是“0/1损失函数”

显然，当C无穷大时，式（6.29）迫使所有样本满足约束条件（6.28），于是式（6.29）等价于（6.6）；当C取有限值时，式（6.29）允许一些样本不满足约束。

然而，$\ell_{0 / 1}$非凸、非连续，数学性质不太好，使得式（6.29）不易直接求解，于是人们通常用其他的一些函数来代替$\ell_{0 / 1}$，称为“替代损失”（surrogate loss）。替代损失函数一般具有较好的数学性质，如它们通常是凸函数且是$\ell_{0 / 1}$的上界。图6.5给出了三种常用的替代损失函数：

hinge 损失: $\ell_{\text {hinge }}(z)=\max (0,1-z)$;
指数损失(exponential loss): $\ell_{\exp }(z)=\exp (-z)$;
对率损失(logistic loss $): \ell_{\log }(z)=\log (1+\exp (-z))$.

若采用hinge损失，式（6.29）变成

引入“松弛变量”（slack variables）$\xi_{i} \geqslant 0$，可将式（6.34）重写为：

这就是常用的“软间隔支持向量机”。

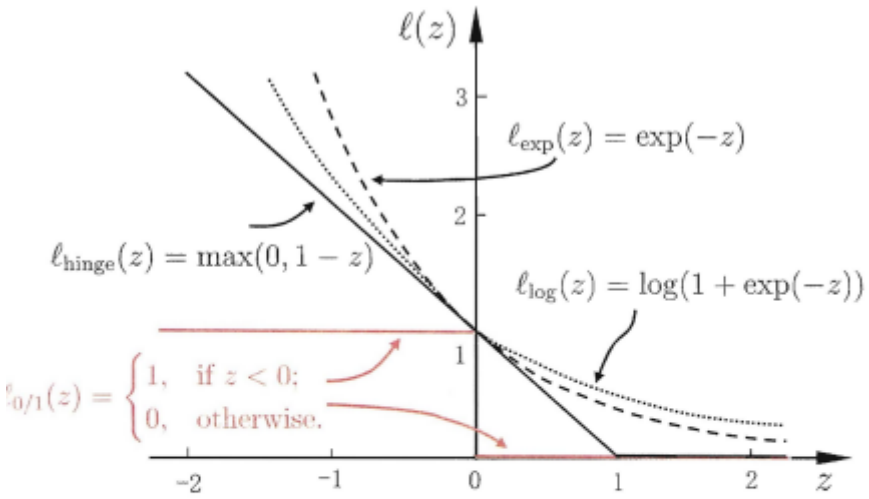

图 6.5 三种常见的替代损失函数: hinge损失、指数损失、对率损失

$$\begin{aligned} \text{s.t.} \quad & y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant 1 - \xi_i \\ & \xi_i \geqslant 0 \ , \ i = 1, 2, \ldots, m. \end{aligned}$$

显然，式（6.35）中每个样本都有一个对应的松弛变量，用以表征该样本不满足约束（6.28）的程度。但是，与式（6.6）相似这仍然是一个二次规划问题。于是，类似（6.8），通过拉格朗日乘子法可得到式（6.35）的拉格朗日函数：

其中，$\alpha_{i} \geqslant 0, \mu_{i} \geqslant 0$是拉格朗日乘子。其后和上述流程一样，对参数求导，求拉格朗日乘子，KKT条件