第二讲 支持向量机

牟克典

2021年3月19日

二分类任务与分离超平面

- 在二分类任务中，输入(特征)空间中最简单的分类边界就是超平面.
- 在能把训练数据中两类样本分开的分离超平面存在的情形下，按照什么样的准则找一个最优的分离超平面？
- 如果除了极个别样本点外大部分样本都是线性可分的，按照什么样的准则来构造一个分离超平面？
- 能否进一步拓展到数据非线性可分的情形？

概要

1 线性可分支持向量机
- 最大间隔分离超平面
- 对偶算法

内积运算

对任意两个向量$x, x' \in \mathbf{R}^n$，我们用$x \cdot x'$表示二者的内积，即

$$x \cdot x' = \sum_{k=1}^{n} x^{(k)} x'^{(k)},$$

这里$x = (x^{(1)}, x^{(2)}, \cdots, x^{(n)})^T$且$x' = (x'^{(1)}, x'^{(2)}, \cdots, x'^{(n)})^T$.

线性模型

- $y = w \cdot x + b$, 其中 $w = (w_1, w_2, \ldots, w_n)^T \in \mathbf{R}^n$和$b \in \mathbf{R}$为参数.
- $y = w^+ \cdot x^+$, 其中
  $w^+ = (w_1, w_2, \ldots, w_n, b)^T$, $x^+ = (x^{(1)}, x^{(2)}, \cdots, x^{(n)}, 1)^T$

训练样本集

- $D = \{(x_i, y_i)\}_{i=1}^{N}$，其中$x_i \in \mathcal{X} = \mathbf{R}^n$，$y_i \in \mathcal{Y} = \{+1, -1\}$.
- 如果$y_i = +1$，则称x_i为正例，否则称x_i为负例，$1 \leq i \leq N$.

D线性可分

如果对D来说，存在特征空间$\mathcal{X} = \mathbf{R}^n$中的超平面 H:

$$w \cdot x + b = 0$$

能够将训练样本正确分类，即对任一$(x_i, y_i) \in D$,

- 若 $y_i = +1$，则$w \cdot x_i + b > 0$;
- 若 $y_i = -1$，则$w \cdot x_i + b < 0$。

则我们称D是线性可分的。

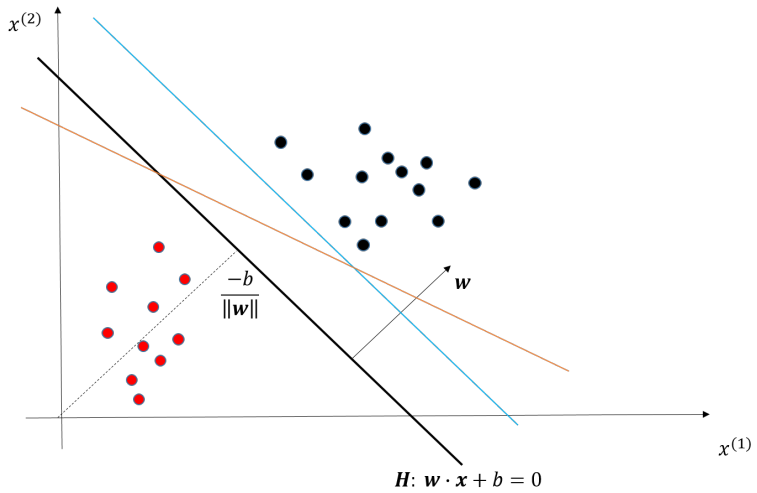

Figure: 线性可分与分离超平面

- 若超平面 H完全正确地将D中的正负样本划分开，我们称其为分离超平面.
- 对于超平面H: $w \cdot x + b = 0$ 来说，w为其法向量，b 为偏置，原点到该超平面的距离为$\frac{|b|}{\|w\|}$.
- 我们用(w, b)表示超平面H.
- 注意超平面H并非惟一的分离超平面.

如何选择一个“最优”分离超平面呢？

- 若超平面(w, b)将训练样本正确分类，则对任一$(x_i, y_i) \in D$来说，都有

$$y_i(w \cdot x_i + b) > 0.$$

- 令$\hat{\gamma} = \min\limits_{1 \le i \le N} y_i(w \cdot x_i + b)$, 则

$$y_i(w \cdot x_i + b) \ge \hat{\gamma} > 0.$$

- 任给的$k \neq 0$来说，(w, b)和(kw, kb)是同一超平面，因此可选择合适的(w, b)使得对任一$(x_i, y_i) \in D$来说，都有

$$y_i(w \cdot x_i + b) \ge 1.$$

- 特别地，对满足
$$y_i(w \cdot x_i + b) = 1$$
的样本点来说，
  - 若$y_i = +1$，则x_i落在超平面 H_1: $w \cdot x + b = 1$上；
  - 若 $y_i = -1$，则x_i落在超平面 H_2: $w \cdot x + b = -1$上.
- 它们是 距离分离超平面(w, b)最近的样本点，它们到(w, b)的几何距离是$\frac{1}{\|w\|}$.
- 超平面H_1和H_2均与分离超平面H平行，且等距离分处H 的两侧.
- 我们把H_1和H_2之间的距离$\frac{2}{\|w\|}$称为间隔.

最大间隔分离超平面

- 样本点到分离超平面的距离刻画了对该样本点分类预测的确信程度.
- 样本点到分离超平面的最短距离 $\frac{1}{\|w\|}$ 刻画了对训练样本点分类预测的最小确信度.
- 最大化对训练样本点分类预测的最小确信度就是最大化间隔.

最优分离超平面不仅要分类正确，而且要使得间隔最大化，即

$$\max_{w,b} \frac{1}{\|w\|}$$
$$\text{s.t.} \quad y_i(w \cdot x_i + b) \geq 1, \quad i = 1, 2, \cdots, N.$$

最大间隔分离超平面

$$\begin{aligned} &\max_{w,b} \frac{1}{\|w\|} \\ &\text{s.t.} \quad y_i(w \cdot x_i + b) \geq 1, \quad i = 1, 2, \cdots, N. \end{aligned} \tag{1}$$

最大化$\frac{1}{\|w\|}$和最小化$\frac{1}{2} \| w \|^2$是等价的，因此式(1)可以重写成如下的凸二次规划问题：

$$\begin{aligned} &\min_{w,b} \frac{1}{2} \| w \|^2 \\ &\text{s.t.} \quad y_i(w \cdot x_i + b) \geq 1, \quad i = 1, 2, \cdots, N. \end{aligned}$$

- 若D是线性可分的，则上述凸二次规划问题的解存在且唯一.

最大间隔分离超平面

我们称

- 这个凸二次规划问题

$$\min_{w,b} \frac{1}{2} \| w \|^2$$
$$\text{s.t. } y_i(w \cdot x_i + b) \geq 1, \quad i = 1, 2, \cdots, N.$$

的解对应的分离超平面(w, b)为最大间隔分离超平面;

- 对应的分类决策函数

$$f(x) = \text{sign}(w \cdot x + b)$$

为线性可分支持向量机, 这里

$$\text{sign}(x) = \begin{cases} +1, & x > 0 \\ -1, & x \leq 0 \end{cases}$$

是符号函数。

对偶算法

采用拉格朗日乘子法来求解凸二次规划问题:

$$\begin{aligned} &\min_{w,b} \frac{1}{2} \parallel w \parallel^2 \\ &\text{s.t.} \quad y_i(w \cdot x_i + b) \geq 1, \quad i = 1, 2, \cdots, N. \end{aligned} \tag{2}$$

我们对上式中的每条约束引进一个拉格朗日乘子$\alpha_i \geq 0$，构造拉格朗日函数如下:

$$L(w, b, \alpha) = \frac{1}{2} \parallel w \parallel^2 + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i y_i (w \cdot x_i + b), \tag{3}$$

其中$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_N)^T$为拉格朗日乘子向量。

$$L(w,b,\alpha)=\frac{1}{2}\parallel w\parallel^2+\sum_{i=1}^{N}\alpha_i-\sum_{i=1}^{N}\alpha_i y_i(w\cdot x_i+b),$$

分别令$L(w,b,\alpha)$对w和b的偏导为0可得到

$$w=\sum_{i=1}^{N}\alpha_i y_i x_i, \tag{4}$$

$$\sum_{i=1}^{N}\alpha_i y_i=0. \tag{5}$$

将式(4)代入(3)，得到

$$L(w,b,\alpha) = \frac{1}{2}\| \sum_{i=1}^{N} \alpha_i y_i x_i \|^2 - \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i \cdot x_j - \sum_{i=1}^{N} \alpha_i y_i b + \sum_{i=1}^{N} \alpha_i \quad (6)$$

利用(5)的约束，得到

$$L(w,b,\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i \cdot x_j. \quad (7)$$

式 (2)

$$\begin{aligned} &\min_{w,b} \frac{1}{2} \| w \|^2 \\ &\text{s.}t. \ \ y_i(w \cdot x_i + b) \geq 1, \quad i = 1, 2, \cdots, N. \end{aligned}$$

的对偶问题为

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \\ &\text{s.}t. \ \ \sum_{i=1}^{N} \alpha_i y_i = 0, \\ &\quad \alpha_i \geq 0, \ \ i = 1, 2, \cdots, N. \end{aligned} \tag{8}$$

设$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_N)^T$ 为式(8)的解，(w, b)为式 (2)的解，则由KKT条件可得

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i, \tag{9}$$

$$\alpha_i(y_i(w \cdot x_i + b) - 1) = 0, i = 1, 2, \cdots, N \tag{10}$$

$$y_i(w \cdot x_i + b) - 1 \geq 0, i = 1, 2, \cdots, N. \tag{11}$$

$$\alpha_i \geq 0, i = 1, 2, \cdots, N. \tag{12}$$

由式(4)可知α_i不能全为0，不放设$\alpha_j > 0$，则

$$y_j(w \cdot x_j + b) = 1. \tag{13}$$

将式(13)两边乘以y_j，可得到

$$b = y_j - w \cdot x_j = y_j - \sum_{i=1}^{N} \alpha_i y_i x_i \cdot x_j. \tag{14}$$

在得到对偶问题的解之后，我们可以由(9)和(14)得到相应的分类决策函数

$$f(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i x_i \cdot x + b \right). \tag{15}$$

对式(14)两边乘以$\alpha_j y_j$ 再求和得到

$$\begin{aligned}\sum_{j=1}^{N}\alpha_j y_j b &= \sum_{j=1}^{N}\alpha_j y_j^2 - \sum_{j=1}^{N}\sum_{i=1}^{N}\alpha_i\alpha_j y_i y_j x_i \cdot x_j. \\ &= \sum_{j=1}^{N}\alpha_j - \| w \|^2 .\end{aligned}$$

利用约束(5)，可以得到

$$\sum_{j=1}^{N}\alpha_j = \| w \|^2 . \tag{16}$$

由此得到间隔的如下表示：

$$\frac{2}{\| w \|} = \frac{2}{\sqrt{\sum_{j=1}^{N}\alpha_j}}. \tag{17}$$

- 由

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i,$$

可知，只有$a_i > 0$的样例(x_i, y_i) 才对模型的构建起作用.

- 由（13）可知，$a_i > 0$的的正例点和负例点正好分别落在间隔边界

$$H_1 : w \cdot x + b = 1$$

和

$$H_2 : w \cdot x + b = -1$$

上，我们称这样的实例点x_i为支持向量。

线性可分训练样本所学习到的最大间隔分离超平面、间隔、间隔边界以及支持向量之间的关系：

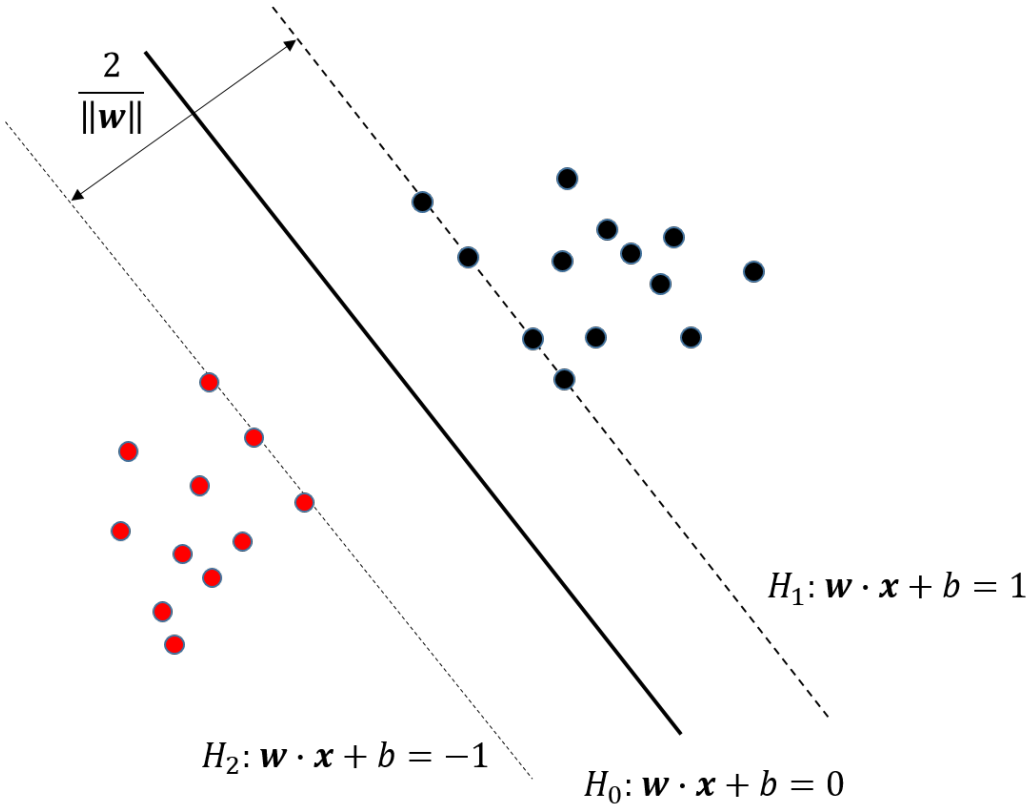

Figure: 支持向量与间隔

支持向量机模型的留一误差

- 定义支持向量机学习算法在D上的留一误差为

$$\hat{R}_{loo} = \frac{1}{N}\sum_{i=1}^{N} I(f_{D-i}(x_i) \neq y_i), \tag{18}$$

其中f_{D-i}为从训练集$D - \{x_i\}$中学习出的支持向量机模型.

- 则

$$\hat{R}_{loo} \leq \frac{N_{SV}(f)}{N} \tag{19}$$

其中$N_{SV}(f)$表示由D学习出的支持向量机f的支持向量个数.

概要

1. 线性可分支持向量机
   - 最大间隔分离超平面
   - 对偶算法
2. 线性支持向量机
   - 软间隔最大化
   - 合页损失函数
   - 对偶算法

- 要求$D=\{(x_i,y_i)\}_{i=1}^N$是线性可分有些严苛.
- 在很多应用场景中，D其实是线性不可分的，即
  - 对任何一个超平面(w,b)来说，一定存在被(w,b)错分的实例点x_i使得

$$y_i(w\cdot x_i+b)<0. \tag{20}$$

  - 除误分类点外，落在以超平面(w,b)为中心，以H_1和H_2为边界的带形区域中的其它样本点尽管被(w,b)正确分类，但

$$0<y_i(w\cdot x_i+b)<1.$$

  - 剩下的样本点就能被超平面(w,b)象线性可分情形一样完全分离.

- 策略: 通过容忍这些点对约束 $y_i(w \cdot x_i + b) \geq 1$ 的违反，来保证超平面(w, b) 能把大多数样本点象线性可分情形一样完全分离.
- 目标不再是只考虑间隔最大化，也要考虑选择的超平面使得不满足约束

$$y_i(w \cdot x_i + b) \geq 1$$

的样本点尽可能少.
- 对每个实例点x_i引入一个松弛变量$\xi_i \geq 0$，以使得x_i满足放宽以后的约束

$$y_i(w \cdot x_i + b) + \xi_i \geq 1. \tag{21}$$

- 对于能被H： $w \cdot x + b = 0$正确分类而且满足约束

$$y_i(w \cdot x_i + b) \geq 1.$$

的实例点x_i来说，松弛变量$\xi_i = 0$即可满足上面放宽的约束.

- 对于能被H正确分类, 但

$$0 < y_i(w \cdot x_i + b) < 1,$$

的实例点x_i来说，$0 < \xi_i < 1$即可满足上面放宽的约束.

- 满足$y_i(w \cdot x_i + b) = 0$的实例点x_i正好落在超平面H上，对应的松弛变量$\xi_i = 1$即可满足放宽的约束.
- 对被超平面H错误分类的实例点x_i来说， 要满足放宽的约束， 对应的松弛变量$\xi_i > 1$才可以.

- 把$\xi_i \neq 0$所对应的实例点x_i视为特异点(Outliers).
- 在忽略掉特异点的情况下，其他训练样本数据的正负实例点的边界之间的距离正好也是$\frac{2}{\|w\|}$.
- 为了和线性可分的情况相区分，我们称这样的间隔为软间隔，线性可分情况下的间隔为硬间隔.
- 既要考虑软间隔最大化，也要考虑特异点的个数尽可能少，但这两个目标通常是相互冲突的.

比较现实的策略是对二者进行权衡，构造如下最优化问题：

$$\min_{w,b} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} I(\xi_i \neq 0) \tag{22}$$

进一步,我们考虑用对这些特异点的总的松弛程度来替代特异点的个数，通过如下最优化问题来刻画线性不可分情形的最优分离超平面:

$$\begin{aligned} \min_{w,b,\xi} \ & \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} \xi_i \\ \text{s.t.} \ & y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad i = 1, 2, \cdots, N. \\ & \xi_i \geq 0, \quad i = 1, 2, \cdots, N. \end{aligned} \tag{23}$$

这里$\xi = (\xi_1, \xi_2, \cdots, \xi_N)^T$是松弛变量向量，$C > 0$是惩罚参数，用于权衡对特异点的总的松弛程度和软间隔.

- 当C值比较大时，对特异点的松弛幅度惩罚比较大.
- 具体C值通常通过k-折交叉验证的方法来确定.

- 式（23）仍然是一个凸二次规划问题.
- 对于给定D和相应的惩罚系数C，求解软间隔最大化问题式（23）的解(w, b)，得到分离超平面

$$w \cdot x + b = 0$$

及其相应的分类决策函数

$$f(x) = \text{sign}(w \cdot x + b).$$

我们称这样的分类模型为训练样本线性不可分时的**线性支持向量机**.

也可以采用$\sum_{i=1}^{N}\xi_i^p$ (其中$p \geq 1$) 来作为总的松弛幅度的度量，即

$$\begin{aligned}
&\min_{w,b,\xi} \ \frac{1}{2} \parallel w \parallel^2 + C\sum_{i=1}^{N}\xi_i^p \\
&\text{s.t.} \ \ y_i(w \cdot x_i + b) \ \geq \ 1 - \xi_i, \qquad i = 1, 2, \cdots, N. \\
&\qquad \xi_i \geq 0, \qquad\qquad\qquad\qquad i = 1, 2, \cdots, N.
\end{aligned}$$

回顾引进松弛变量的初衷:

- $\xi_i = 0$ 对应的样本点x_i都满足约束 $y_i(w \cdot x_i + b) \geq 1$.
- 而对特异点x_i 来说，$\xi_i \geq 1 - y_i(w \cdot x_i + b)$.
- 考虑到目标函数里面松弛的总幅度要尽可能小，对特异点x_i我们这里不妨取

$$\xi_i = 1 - y_i(w \cdot x_i + b).$$

我们引进合页损失函数

$$h(z) = \max(0, 1 - z),$$

则ξ_i可以用合页损失函数表示为:

$$\xi_i = h\left(y_i(w \cdot x_i + b)\right).$$

显然，这样的ξ_i都满足放宽的约束.

$$\begin{aligned}
\min_{w,b,\xi} \ & \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} \xi_i \\
\text{s.t.} \ & y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad i = 1, 2, \cdots, N. \\
& \xi_i \geq 0, \quad i = 1, 2, \cdots, N.
\end{aligned}$$

我们得到与上式(式(23))等价的最优化问题:

$$\min_{w,b} \ \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} h(y_i(w \cdot x_i + b)) .$$

与(23)等价的最优化问题:

$$\min_{w,b} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} h(y_i(w \cdot x_i + b)) .$$

进一步等价于

$$\min_{w,b} \sum_{i=1}^{N} h(y_i(w \cdot x_i + b)) + \frac{1}{2C} \| w \|^2 .$$

这其实对应于损失函数为合页损失时的结构风险最小化策略.

与

$$\min_{w,b} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} I(\xi_i \neq 0)$$

相比，

$$\min_{w,b} \sum_{i=1}^{N} h(y_i(w \cdot x_i + b)) + \frac{1}{2C} \| w \|^2 .$$

相当于我们采用0-1损失函数的一致替代损失函数合页损失函数来替代0-1损失函数.

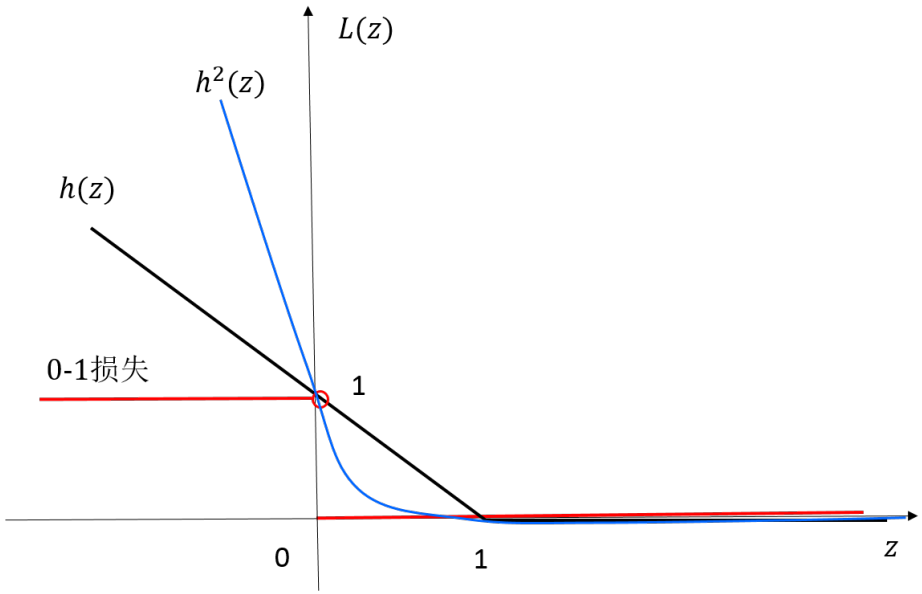

Figure: 合页损失与二次合页损失函数

我们也可以采用0-1损失函数的其他一致替代函数，比如可以采用二次合页损失函数如下：

$$\min_{w,b} \sum_{i=1}^{N} h^2\left(y_i(w \cdot x_i + b)\right) + \frac{1}{2C} \| w \|^2 .$$

$$\begin{aligned}
&\min_{w,b,\xi} \ \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{N} \xi_i \\
&\text{s.t.} \ \ y_i(w \cdot x_i + b) \ \geq \ 1 - \xi_i, \qquad i = 1, 2, \cdots, N. \\
&\qquad \xi_i \geq 0, \qquad\qquad\qquad\qquad i = 1, 2, \cdots, N.
\end{aligned}$$

- 对每条约束$y_i(w \cdot x_i + b) \ \geq \ 1 - \xi_i$引进一个拉格朗日乘子$\alpha_i \geq 0$,
- 对每条约束$\xi_i \geq 0$引进一个拉格朗日乘子$\beta_i \geq 0$

构造拉格朗日函数如下：

$$\begin{aligned}
&L(w, b, \xi, \alpha, \beta) \\
&= \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i \left[y_i(w \cdot x_i + b) - 1 + \xi_i \right] - \sum_{i=1}^{N} \beta_i \xi_i,
\end{aligned}$$

其中$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_N)^T$和$\beta = (\beta_1, \beta_2, \cdots, \beta_N)^T$为拉格朗日乘子向量.

令$L(w, b, \xi, \alpha, \beta)$对$w$、$b$和每个$\xi_i$的偏导为0可得到

$$\nabla_w L = w - \sum_{i=1}^{N} \alpha_i y_i x_i = 0, \implies w = \sum_{i=1}^{N} \alpha_i y_i x_i, \tag{24}$$

$$\nabla_b L = -\sum_{i=1}^{N} \alpha_i y_i = 0, \implies \sum_{i=1}^{N} \alpha_i y_i = 0. \tag{25}$$

$$\nabla_{\xi_i} L = C - \alpha_i - \beta_i = 0, \implies \alpha_i + \beta_i = C. \tag{26}$$

将式(24)代入拉格朗日函数，并利用(25)和(26)的约束，得到

$$L(w,b,\xi,\alpha,\beta)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i\cdot x_j.$$

那么式 (23)的对偶问题为

$$\begin{aligned}&\max_{\alpha}\ \sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i\cdot x_j\\&\text{s.}t.\ \sum_{i=1}^{N}\alpha_i y_i=0,\\&\quad \alpha_i\geq 0\ \wedge\ \beta_i\geq 0\ \wedge\ \alpha_i+\beta_i=C,\quad i=1,2,\cdots,N.\end{aligned}\tag{27}$$

进一步利用约束$\alpha_i + \beta_i = C$消去β_i，得到如下对偶问题：

$$\begin{aligned} \max_{\alpha} & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \\ s.t. & \sum_{i=1}^{N} \alpha_i y_i = 0, \\ & 0 \leq \alpha_i \leq C, \quad i = 1, 2, \cdots, N. \end{aligned} \tag{28}$$

和D线性可分的情形相比，

- 对偶问题的目标函数完全一样.
- 约束部分只是增加了约束$\alpha_i \leq C$.

设$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_N)^T$ 为对偶问题的解，

- 由式(24)得到原始问题的解w仍然可以表示为训练样本点x_i的线性组合，即

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i.$$

- 仍然称$\alpha_i > 0$的样本点x_i为支持向量.
- 但对线性不可分的情形来说，并非所有的支持向量都落在间隔边界

$$w \cdot x + b = \pm 1$$

上.

考虑KKT条件：

$$\begin{aligned}
\alpha_i(y_i(w\cdot x_i+b)-1+\xi_i)=0, i=1,2,\cdots,N\\
y_i(w\cdot x_i+b)-1+\xi_i\geq 0, i=1,2,\cdots,N.\\
\beta_i\xi_i=0, i=1,2,\cdots,N.\\
\alpha_i\geq 0, i=1,2,\cdots,N.\\
\beta_i\geq 0, i=1,2,\cdots,N.
\end{aligned}$$

则$\alpha_i>0$的样本点x_i来说，

$$y_i(w\cdot x_i+b)=1-\xi_i.$$

进一步，如果$0<\alpha_i<C$，则$\beta_i>0$，由KKT条件可知$\xi_i=0$。因此，$y_i(w\cdot x_i+b)=1$. 这意味着对$0<\alpha_i<C$的样本点x_i仍然落在间隔边界H_1或H_2上。

进一步，将式$y_i(w \cdot x_i + b) = 1$.两边乘以y_i，可得到

$$b = y_i - w \cdot x_i = y_i - \sum_{j=1}^{N} \alpha_j y_j x_j \cdot x_i.$$

因此，在得到对偶问题的解之后，我们可以得到最优分离超平面

$$\sum_{i=1}^{N} \alpha_i y_i x_i \cdot x + b = 0$$

和相应的分类决策函数

$$f(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i x_i \cdot x + b\right)$$

- 最优分离超平面的法向量惟一，但偏置不一定惟一.
- 通常在具体求解算法中用若干个如上求得的偏置的均值作为最优超平面的偏置的估计值.

- $\alpha_i = C$来说，$\xi_i > 0$的支持向量都是特异点:
  - x_i到所属类别的边界超平面的距离为$\frac{\xi_i}{\|w\|}$.
  - 如果$0 < \xi_i < 1$，则x_i落在边界和分离超平面之间，仍然被正确分类.
  - 如果$\xi_i = 1$，则x_i正好落在分离超平面上.
  - 如果$\xi_i > 1$，则x_i被分离超平面错误分类.

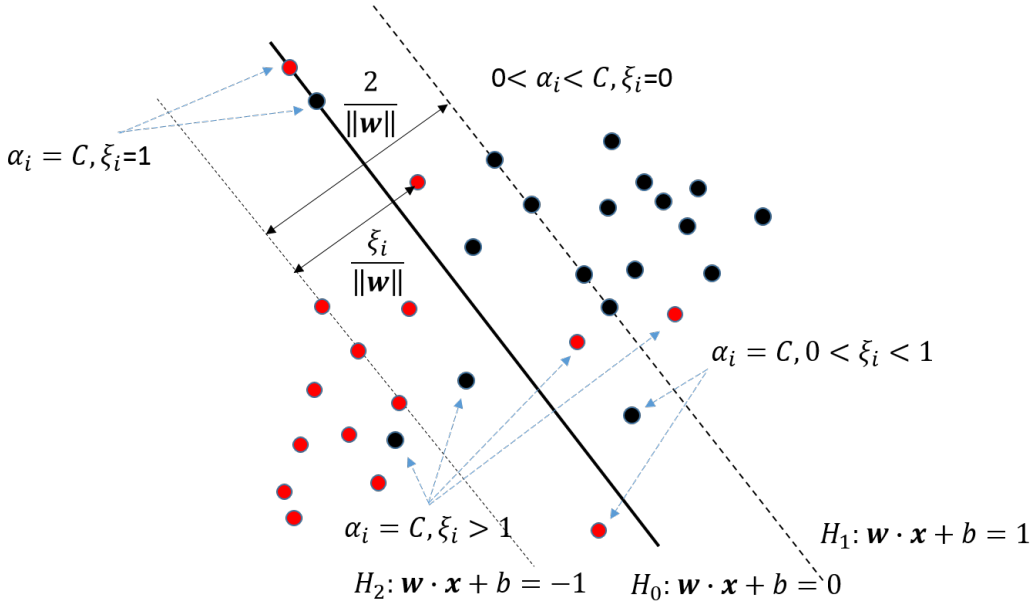

Figure: 软间隔与支持向量

概要

1. 线性可分支持向量机
   - 最大间隔分离超平面
   - 对偶算法
2. 线性支持向量机
   - 软间隔最大化
   - 合页损失函数
   - 对偶算法
3. SMO算法

回顾对偶问题

$$\begin{aligned}
&\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \\
&\text{s.t.} \sum_{i=1}^{N} \alpha_i y_i = 0, \\
&\quad 0 \leq \alpha_i \leq C, \quad i = 1, 2, \cdots, N.
\end{aligned}$$

- 参数的个数为训练样本容量N.
- 当N比较大时，如何高效地求解对偶问题?

由Platt提出的序列最小最优化算法(Sequential Minimal Optimization，简称SMO)就是针对这个凸二次规划问题的一种快速实现算法.

SMO算法的主要思想:

- 将参数（拉格朗日乘子）个数比较多的凸二次规划问题约减为一系列仅涉及两个拉格朗日算子的小规模的二次规划问题;
- 只涉及两个拉格朗日乘子的二次规划问题可以通过解析方法求解，从而加速支持向量机的训练过程.

SMO算法的核心问题:

- 如何确定要更新的两个拉格朗日乘子;
- 在要更新的拉格朗日乘子确定之后如何求解只涉及这两个变量的二次规划问题.

我们先介绍求解只涉及两个拉格朗日乘子的二次规划问题的解析方法。

设当前选择的需要更新的拉格朗日乘子为α_1和α_2，其他拉格朗日乘子α_i $(3 \leq i \leq N)$在本轮参数更新中保持不变. 由约束$\sum_{i=1}^{N} \alpha_i y_i = 0$ 可知

$$\alpha_1 y_1 + \alpha_2 y_2 + \sum_{i=3}^{N} \alpha_i y_i = 0.$$

上式两边乘以y_1得到

$$\alpha_1 + \alpha_2 y_1 y_2 + \sum_{i=3}^{N} \alpha_i y_1 y_i = 0.$$

令$\gamma = -\sum_{i=3}^{N} \alpha_i y_1 y_i$，$s = y_1 y_2 \in \{-1, +1\}$, 则

$$\alpha_1 + s\alpha_2 = \gamma.$$

令$K_{ij} = x_i \cdot x_j$ 且 $v_i = \sum_{j=3}^{N} \alpha_j y_j K_{ij}$ $(i = 1, 2)$，则将对偶问题转化为下面关于拉格朗日乘子α_1 和α_1的优化问题：

$$\begin{aligned} \max_{\alpha_1, \alpha_2} \quad & W_1(\alpha_1, \alpha_2), \\ s.t. \quad & 0 \le \alpha_1, \alpha_2 \le C, \\ & \alpha_1 + s\alpha_2 = \gamma. \end{aligned}$$

其中

$$\begin{aligned} W_1(\alpha_1, \alpha_2) = \; & \alpha_1 + \alpha_2 - \frac{1}{2} K_{11} \alpha_1^2 - \frac{1}{2} K_{22} \alpha_2^2 \\ & - s K_{12} \alpha_1 \alpha_2 - y_1 \alpha_1 v_1 - y_2 \alpha_2 v_2. \end{aligned} \tag{29}$$

由约束$\alpha_1 + s\alpha_2 = \gamma$可得

$$\alpha_1 = \gamma - s\alpha_2.$$

将其代入$W_1(\alpha_1, \alpha_2)$，并令

$$W_2(\alpha_2) = W_1(\gamma - s\alpha_2, \alpha_2). \tag{30}$$

令$W_2(\alpha_2)$对α_2的导数为0，可以得到

$$\alpha_2 = \frac{s(K_{11} - K_{12})\gamma + y_2(v_1 - v_2) - s + 1}{\eta}, \tag{31}$$

其中

$$\eta = K_{11} + K_{22} - 2K_{12}. \tag{32}$$

进一步，令α_i^*为当前（即开始更新α_1和α_2之前）的α_i的值，定义

$$f(x) = \sum_{i=1}^{N} \alpha_i^* y_i x_i \cdot x + b^*, \tag{33}$$

则

$$v_1 - v_2 = f(x_1) - f(x_2) + \alpha_2^* y_2 \eta - s y_2 \gamma (K_{11} - K_{12}). \tag{34}$$

由此我们可以得到未经剪辑时的α_2的最优解：

$$\alpha_2 = \alpha_2^* + y_2 \frac{(y_2 - f(x_2)) - (y_1 - f(x_1))}{\eta}, \tag{35}$$

注意到约束$0 \le \alpha_1, \alpha_2 \le C$，我们需要对$\alpha_2$进行剪辑。

由于$\alpha_1 + s\alpha_2 = \gamma$,

- 若$s = -1$, 则定义α_2 的下界为$L = \max\{0, -\gamma\}$, 上界为$H = \min\{C, C - \gamma\}$.
- 若$s = +1$, 则定义α_2的下界为$L = \max\{0, \gamma - C\}$, 上界为$H = \min\{C, \gamma\}$.

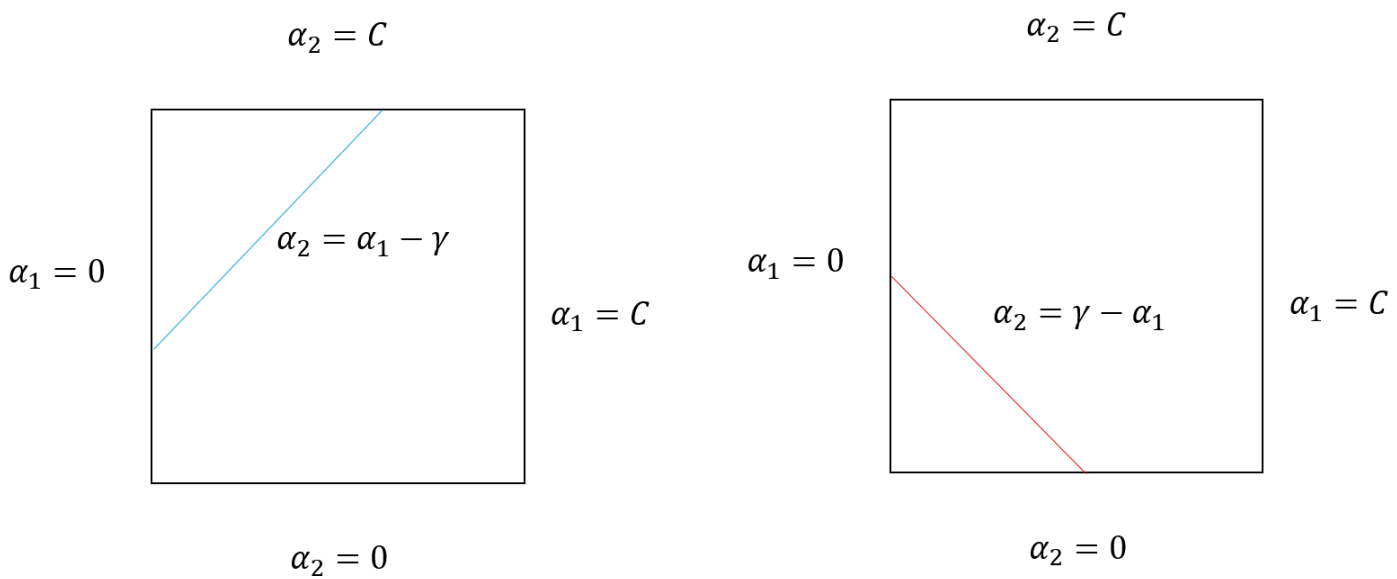

Figure: α_2的上下界

则经剪辑以后的α_2的解为：

$$\alpha_2^{clip} = \begin{cases} \alpha_2 & \text{if } L < \alpha_2 < H \\ L & \text{if } \alpha_2 \leq L \\ H & \text{if } \alpha_2 \geq H \end{cases} \tag{36}$$

相应地，α_1的最优解为

$$\alpha_1 = \alpha_1^* + s(\alpha_2^* - \alpha_2^{clip}). \tag{37}$$

在每次完成两个变量的更新后，都需要重新计算b的值。如果$0 < \alpha_i < C, i = 1$或2，则可通过求解

$$\sum_{j=1}^{N} \alpha_j y_j K_{ij} + b = y_i \tag{38}$$

来计算b。

SMO算法如何选择需要更新的拉格朗日乘子α_1和α_2呢？

- 选取当前违背KKT条件程度最严重的变量作为一个变量，
- 第二个变量应选择一个使得目标函数值增长最快的变量.

关于SMO算法的具体框架可以进一步参阅[3]或者[1]。

概要

1. 线性可分支持向量机
   - 最大间隔分离超平面
   - 对偶算法
2. 线性支持向量机
   - 软间隔最大化
   - 合页损失函数
   - 对偶算法
3. SMO算法
4. 核方法与非线性支持向量机

在很多实际学习任务中，正负类之间的划分边界可能是非线性的，比如著名的异或问题:

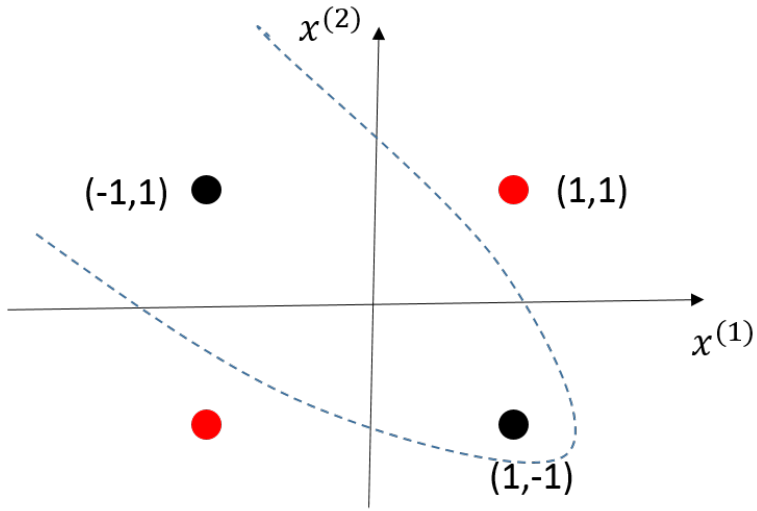

Figure: 异或问题

比较自然的想法是将样本从原始特征空间$\mathcal{X}$映射到一个更高维的特征空间（Hilbert空间）$\mathcal{Z}$，使得样本在新的特征空间是线性不可分（可分）的。

假设能够直接构建这样的映射$\phi : \mathcal{X} \mapsto \mathcal{Z}$, 则可以直接在新的更高维特征空间$\mathcal{Z}$中构建相应的线性支持向量机模型如下:

$$\begin{aligned} \min_{w,b,\xi} \ & \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{N} \xi_i \\ \text{s.t.} \ & y_i(w \cdot \phi(x_i) + b) \geq 1 - \xi_i, \quad i = 1, 2, \cdots, N. \\ & \xi_i \geq 0, \quad i = 1, 2, \cdots, N. \end{aligned}$$

其对偶问题是

$$\begin{aligned} \max_{\alpha} & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \phi(x_i) \cdot \phi(x_j) \\ \text{s.t.} & \sum_{i=1}^{N} \alpha_i y_i = 0, \\ & 0 \leq \alpha_i \leq C, \quad i = 1, 2, \cdots, N. \end{aligned}$$

这里$\phi(x_i) \cdot \phi(x_j)$是$\mathcal{Z}$空间中$\phi(x_i)$和$\phi(x_j)$的内积。

如果α是对偶问题的解，则我们可以得到相应的分类决策函数为

$$f(x) = \text{sign}(\sum_{i=1}^{N} \alpha_i y_i \phi(x_i) \cdot \phi(x) + b)$$

问题:

在实际学习任务中ϕ也很难构建出来，直接计算$\phi(x_i) \cdot \phi(x_j)$比较困难.

怎么办?

我们引进核技巧来解决这个问题!

核方法

- 不直接定义$\mathcal{Z}$和相应的映射ϕ，而是通过引进一个使得对任意的$x_i, x_j \in \mathcal{X}$，

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$$

成立的$\mathcal{X}$上的所谓核函数$K(x_i, x_j)$来隐式地表示它们.

- 相应的对偶问题可以表示成

$$\begin{aligned}
&\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\
&\text{s.}t. \sum_{i=1}^{N} \alpha_i y_i = 0, \\
&\quad 0 \leq \alpha_i \leq C, \quad i = 1, 2, \cdots, N.
\end{aligned}$$

核方法

- 求得对偶问题的解α，则得到相应的分类决策函数为

$$f(x) = \text{sign}(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b),$$

这里 $b = y_i - \sum_{j=1}^{N} \alpha_j y_j K(x_j, x_i)$，其中$x_i$对应的$\alpha_i$满足$0 < \alpha_i < C$。

- 我们称上述分类决策函数为非线性支持向量机.
- 显然线性支持向量机是非线性支持向量机中核函数 $K(x_i, x_j) = x_i \cdot x_j$ 的特殊情形.

核函数

正定对称核函数

$\mathcal{X}$上的函数$K:\mathcal{X}\times\mathcal{X}\to\mathbf{R}$被称为正定对称核函数如果对任意的$\{x_1,x_2,\cdots,x_m\}\subseteq\mathcal{X}$，核矩阵或Gram矩阵$[K(x_i,x_j)]_{m\times m}$是对称半正定矩阵.

- 常用的正定对称核函数有：
  - 多项式核函数：$K(x,x')=(x\cdot x'+c)^d$,其中$d\in\mathbf{N}$为多项式的次数，$c>0$为常数。
  - 高斯核函数：$K(x,x')=\exp\left(-\frac{\|x'-x\|^2}{2\sigma^2}\right)$,其中$\sigma>0$。
- 核函数选择也是支持向量机模型学习的重要一环：
  - 核函数 仅是隐式地定义了高维特征空间；
  - 同一核函数而言，可能对应到若干从原始特征空间到不同高维特征空间的映射.

小结

- 线性可分支持向量机
  - 最大间隔分离超平面
  - 对偶算法
- 线性支持向量机
  - 松弛变量与软间隔最大化
  - 合页损失函数
  - 对偶算法
- SMO算法
- 非线性支持向量机与核方法

参考文献

[1] 李航，统计学习方法，清华大学出版社，2012

[2] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, Foundations of Machine Learning, MIT Press, 2012

[3] John C.Platt, Fast training of support vector machines using sequential minimal optimization. In Advances in Kernel Methods, Pages 185-208, MIT Press, 1999

[4] 周志华，机器学习，清华大学出版社，2016

[5] Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning, Springer.2001