支持向量机讲义

时明

shiming@hbut.edu.cn

2022 年 10 月 06 日

1. 问题描述

对于数据集$T=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_N,y_N)\}$，其中$x_i\in\boldsymbol{R}^p=\mathcal{X}$，$y_i\in\{-1,+1\}$.称该数据集是线形可分的，若存在$\boldsymbol{w}\in\boldsymbol{R}^p,b\in\boldsymbol{R}$，使得下式成立。

$$\begin{cases} w^Tx_i+b\geq 1, & if\ y_i=1 \\ w^Tx_i+b\leq -1, & if\ y_i=-1 \end{cases} \tag{1}$$

即

$$y_i(w^Tx_i+b)\geq 1,\forall i\in\{1,2,\cdots,N\} \tag{2}$$

一个线形可分的数据集的例子如下图所示

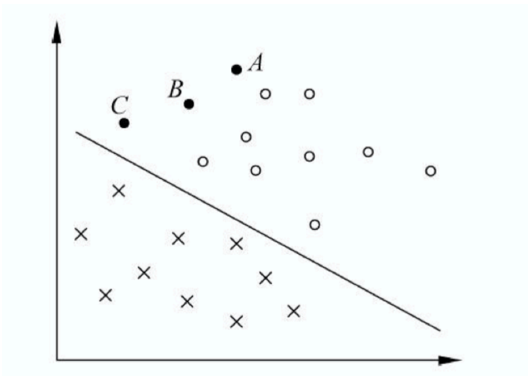

图 1 线形可分数据集实例

此时，假设$(\boldsymbol{w},b)$满足分类不等式组(2)，那么它对应的分类间隔为

$$\mathcal{r}=\left[\min_{i:y_i=1}(w^Tx_i+b)-\max_{j:y_j=-1}\left(w^Tx_j+b\right)\right]\cdot\frac{1}{\|w\|_2} \tag{3}$$

从上式容易看出，分类间隔$\mathcal{r}$的值与b无关。最大化该分类间隔即

$$\max_{\boldsymbol{w},b}\mathcal{r} \tag{4}$$

该目标函数等价于

$$\min_{w,b}\frac{1}{2}\|w\|_2^2 \\ \text{s.t.}\quad y_i(w^Tx_i+b)-1\geq 0,i=1,2,\cdots,N \tag{5}$$

定理 1：如果数据集T是线性可分的，那么关于(w,b)目标函数(5)的最优解(w^*,b^*)是唯一存在的。

证明：首先证明最优解的存在性。由于数据集T是线性可分，那么不等式组(2)的可行解存在。考虑到(5)中约束条件是关于(w,b)的连续函数，因此可行解集

是闭的，并且存在(w^*,b^*)是可行解是最优解。

然后，证明最优解是唯一的。假设存在(w_1^*,b_1^*)和(w_2^*,b_2^*)是最优解，则可得$w_1^*=w_2^*=w^*$。考虑到T的线性可分，则分别存在正样本$y_i=y_j=1$使得

$$\begin{cases} w_1^* x_i + b_1^* = w^* x_i + b_1^* = 1 \\ w_2^* x_j + b_2^* = w^* x_j + b_2^* = 1 \end{cases} \tag{6}$$

考虑到(w_1^*,b_1^*)和(w_2^*,b_2^*)是目标函数的解，那么

$$\begin{cases} w_1^* x_j + b_1^* \geq 1 \\ w_2^* x_i + b_2^* \geq 1 \end{cases} \tag{7}$$

结合不等式(5)和(6)可得$b_1^* \geq b_2^*$和$b_2^* \geq b_1^*$，即$b_1^*=b_2^*$.因此$(w_1^*,b_1^*)=(w_2^*,b_2^*)$，最优解是唯一的。

2. 拉格朗日对偶问题

考虑拉格朗日函数(Lagrange function)$\mathcal{L}(w,b,\alpha)$如下

$$\begin{aligned} \mathcal{L}(w,b,\alpha) &= \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{N} \alpha_i\left(y_i(w^T x_i + b) - 1\right) \\ &= \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{N} \alpha_i y_i (w^T x_i + b) + \sum_{i=1}^{N} \alpha_i \end{aligned} \tag{8}$$

其中$(w,b,\alpha) \in R^p \times R \times [0,+\infty)^N$，则以下等式成立

$$\max_{\alpha_i \geq 0} \mathcal{L}(w,b,\alpha) = \begin{cases} \frac{1}{2}\|w\|_2^2, \quad if \quad y_i(w^T x_i + b) - 1 \geq 0 \\ \infty, \ others \end{cases} \tag{9}$$

此时，原目标函数(5)等价于

$$\min_{w,b} \max_{\alpha_i \geq 0} \mathcal{L}(w,b,\alpha) \tag{10}$$

命题 2：假设二元函数$f(x,y)$的定义域为$\{(x,y)|x \in P, y \in Q\}$,其中数集$P$、$Q \subseteq R$.则以下不等式成立：

$$\min_{x \in P} \max_{y \in Q} f(x,y) \geq \max_{y \in Q} \min_{x \in P} f(x,y) \tag{11}$$

证明：对于$\forall x_0 \in P$和$\forall y_0 \in Q$，不等式$\max_{y \in Q} f(x_0,y) \geq f(x_0,y_0) \geq \min_{x \in P} f(x,y_0)$成立，因此$\min_{x_0 \in \mathrm{P}} \max_{y \in Q} f(x_0,y) \geq \max_{y_0 \in Q} \min_{x \in P} f(x,y_0)$，命题成立。

根据以上命题可以得到，对于定义在$R^p \times R \times [0,+\infty)^N$上的拉格朗日函数$\mathcal{L}(w,b,\alpha)$，以下不等式

$$p^* = \min_{w,b} \max_{\alpha_i \geq 0} \mathcal{L}(w,b,\alpha) \geq \max_{\alpha_i \geq 0} \min_{w,b} \mathcal{L}(w,b,\alpha) = d^* \tag{12}$$

成立。一般称优化问题

$$\max_{\alpha_i \geq 0} \min_{w,b} \mathcal{L}(w,b,\alpha) \tag{13}$$

为拉格朗日原问题(10)的对偶问题。

根据拉格朗日函数对偶定理[1]，如果约束条件(5)是严格可行的，即存在可行解(w', b')，使得

$$y_i(w'^T x_i + b') - 1 > 0, \forall\, i = 1,2,\cdots,N \tag{14}$$

那么拉格朗日原问题(10)与对偶问题(13)的最优值是相同的，

$$p^* = d^* \tag{15}$$

并且(w^*, b^*)和α^*是分别是拉格朗日原问题(10)的解和对偶问题(13)的解当且仅当它们满足 KKT 条件

$$\begin{cases} \nabla_w \mathcal{L}(w^*, b^*, \alpha^*) = 0 \\ \nabla_b \mathcal{L}(w^*, b^*, \alpha^*) = 0 \\ \alpha_i^* \left(y_i\left({w^*}^T x_i + b^*\right) - 1\right) = 0, i = 1,2,\cdots,N \\ y_i\left({w^*}^T x_i + b^*\right) - 1 \geq 0, i = 1,2,\cdots,N \\ \alpha_i^* \geq 0, i = 1,2,\cdots,N \end{cases} \tag{16}$$

若将(w^*, b^*, α^*)代入原方程可得

$$\mathcal{L}(w^*, b^*, \alpha^*) = \frac{1}{2}\|w^*\| = p^* = d^* \tag{17}$$

由于数据集T是线性可分的，因此约束条件(5)是严格可行的。据此，求解拉格朗日原问题(10)最优解的一个思路是：首先求解对偶问题(13)的最优解α^*，然后根据 KKT 条件求解拉格朗日原问题(10)的最优解(w^*, b^*)。根据以上思路，首先求解拉格朗日对偶问题(13)的最优解，具体来说分成 3 步：

(a) 拉格朗日函数$\mathcal{L}(w, b, \alpha)$关于$(w, b)$的极小：$\min\limits_{w,b} \mathcal{L}(w, b, \alpha)$

令梯度

$$\begin{aligned} \nabla_w \mathcal{L}(w, b, \alpha) &= w - \sum_{i=1}^{N} \alpha_i y_i x_i = 0 \\ \nabla_b \mathcal{L}(w, b, \alpha) &= -\sum_{i=1}^{N} \alpha_i y_i = 0 \end{aligned} \tag{18}$$

即得

$$\begin{aligned} w &= \sum_{i=1}^{N} \alpha_i y_i x_i \\ 0 &= \sum_{i=1}^{N} \alpha_i y_i \end{aligned} \tag{19}$$

将上述等式代入拉格朗日函数(8)可得

[1] 《统计学习方法(第 2 版)》，李航：第 447-450 页

$$\mathcal{L}(w,b,\alpha)=\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j\left(x_i^T x_j\right)-\sum_{i=1}^{N}\alpha_i y_i\left[\left(\left(\sum_{j=1}^{N}\alpha_j y_j x_j\right)^T x_i+b\right)\right]$$
$$+\sum_{i=1}^{N}\alpha_i=-\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j\left(x_i^T x_j\right)+\sum_{i=1}^{N}\alpha_i \tag{20}$$

(b) 极小值$\min\limits_{w,b}\mathcal{L}(w,b,\alpha)$关于α的极大

由于$\alpha\in[0,+\infty)^N$，并且根据 KKT 条件

$$\nabla_b\mathcal{L}(w,b,\alpha)=\sum_{i=1}^{N}\alpha_i y_i=0 \tag{21}$$

因此求解以下问题的最优解$\alpha^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_N^*)$

$$\begin{aligned}&\min_{\alpha}-\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j\left(x_i^T x_j\right)+\sum_{i=1}^{N}\alpha_i\\ &\mathrm{s.t.}\ \sum_{i=1}^{N}\alpha_i y_i=0\\ &\alpha_i\geq 0\end{aligned} \tag{22}$$

求解该最优解的一种方法是序列最优化(sequential minimal optimization, SMO)算法。

(c) 拉格朗日原问题的最优解(w^*,b^*)

根据 KKT 条件

$$\nabla_w\mathcal{L}(w^*,b^*,\alpha^*)=w^*-\sum_{i=1}^{N}\alpha_i^* y_i x_i=0 \tag{23}$$

可得

$$w^*=\sum_{i=1}^{N}\alpha_i^* y_i x_i \tag{24}$$

考虑到数据集T是线性可分的，因此

$$w^*=\sum_{i=1}^{N}\alpha_i^* y_i x_i\neq 0 \tag{25}$$

即存在$\alpha_i^*\neq 0$，又根据 KKT 条件

$$\alpha_i^*\left(y_i\left({w^*}^T x_i+b^*\right)-1\right)=0 \tag{26}$$

因此

$$y_i\left({w^*}^T x_i+b^*\right)-1=0 \tag{27}$$

于是

$$b^*=y_i-{w^*}^T x_i \tag{28}$$

3. 松弛变量

当数据集T是线性不可分时，如下图示例，则优化问题(5)不存在可行解，按照上述过程无法求得最优分类平面。此时可以考虑基于以下目标函数寻找最优分类平面

$$\min_{w,b}\frac{1}{2}\|w\|_2^2+C\sum_{i=1}^{N}l_{0/1}(y_i(w^Tx_i+b)-1) \tag{29}$$

其中$C\geq 0$是指定的正常数，$l_{0/1}(\cdot)$是 0/1 损失函数，

$$l_{0/1}(z)=\begin{cases}1, & \text{if } z<0;\\ 0, & \text{otherwise.}\end{cases} \tag{30}$$

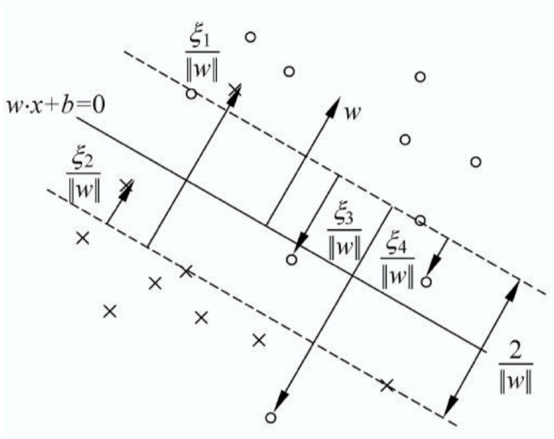

图 2 一个线形不可分数据集的示例。

考虑$l_{0/1}$损失函数是非凸的，因此可以采用凸损失函数代替，例如：

$$\begin{aligned}&\text{合页损失(hinge loss)：}l_{hinge}(z)=\max(0,1-z)\\&\text{指数损失(exponential loss):}\ l_{exp}(z)=\exp(-z)\\&\text{对率损失(logistic loss):}\ l_{log}(z)=\ln(1+\exp(-z))\end{aligned} \tag{31}$$

如果采用合页损失函数l_{hinge}代替$l_{0/1}$损失函数，那么目标函数(29)转化为

$$\min_{w,b}\frac{1}{2}\|w\|_2^2+C\sum_{i=1}^{N}l_{hinge}\big(y_i(w^Tx_i+b)\big) \tag{32}$$

考虑将以上目标函数转化为下面优化问题

$$\begin{aligned}&\min_{w,b}\frac{1}{2}\|w\|_2^2+C\sum_{i=1}^{N}\xi_i\\ \text{s.t.}\ &\xi_i\geq l_{hinge}\big(y_i(w^Tx_i+b)\big)\\ &\quad=\max\big(0,1-y_i(w^Tx_i+b)\big)\end{aligned} \tag{33}$$

该目标函数等价于

$$\min_{w,b}\frac{1}{2}\|w\|_2^2+C\sum_{i=1}^{N}\xi_i \tag{34}$$

$$\text{s.t.}\, \xi_i \geq 1 - y_i(w^T x_i + b)$$
$$\xi_i \geq 0$$

其中$\xi_i, i = 1,2,\cdots,N$被称为松弛变量，该约束优化问题具有和目标函数(5)类似的形式，因而可以采用先转换为对偶函数优化问题(13)，然后利用 KKT 条件(16)求解最优解的思路。

4. 非线性分类问题与核函数

在某些情况下，数据集T并非是线性可分的，需要用某些非线性函数，将样本点非线性投射到更高维空间中，才能够转换为线性可分问题，如下图所示。

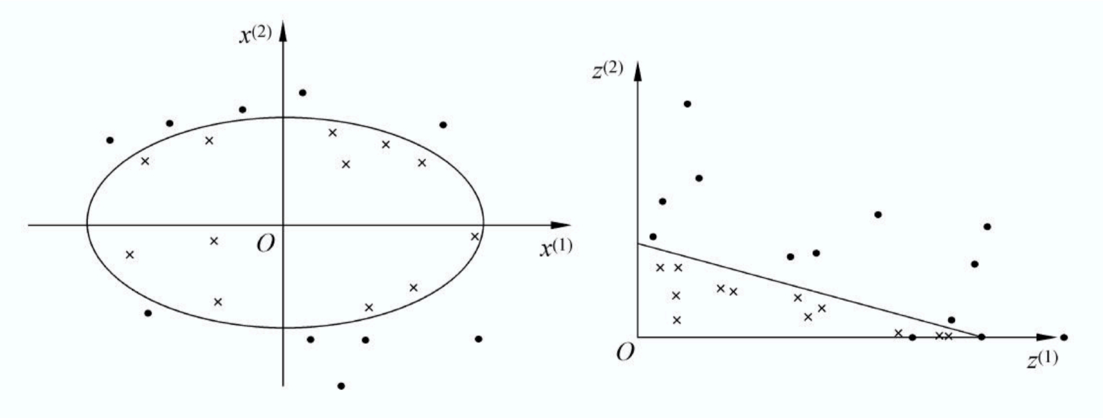

图 3 非线性分类问题

对于某数据空间$\mathcal{X}$，称$\kappa(\cdot,\cdot)$是定义在$\mathcal{X}\times\mathcal{X}$上的对称函数。若$\forall n \in \mathbb{N}^+$，$x_1, x_2, \cdots, x_n \in \mathcal{X}$，Gram 矩阵

$$\boldsymbol{K} = \begin{pmatrix} \kappa(x_1, x_1) & \cdots & \kappa(x_1, x_n) \\ \vdots & \ddots & \vdots \\ \kappa(x_n, x_1) & \cdots & \kappa(x_n, x_n) \end{pmatrix} \tag{34}$$

是半正定(semi-positive definite)的，那么称κ是$\mathcal{X}$上的正定核函数。

定义映射

$$\phi: x \to k(\cdot, x) \tag{35}$$

然后定义集合

$$\mathcal{H}_0 = \text{span}\{k(\cdot, x) | x \in \mathcal{X}\} \tag{36}$$

那么容易验证$\mathcal{H}_0$构成一个线性空间。对于任意$f, g \in \mathcal{H}_0$，则存在α_i, β_j使得

$$\begin{gathered} f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \\ g = \sum_{i=1}^{m} \beta_j k(\cdot, y_j) \end{gathered} \tag{37}$$

则对于$x \in \mathcal{H}$，

$$f(x)=\sum_{i=1}^{n}\alpha_i k(x,x_i)\,. \tag{38}$$

定义以下运算

$$f*g=\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_i\beta_j k(x_i,y_j) \tag{39}$$

容易验证该运算满足以下条件

$$\begin{aligned}
&(a)\ \ (cf)*g=c(f*g),c\in R\\
&(b)\ \ f*g=g*f\\
&(c)\ \ f*(g*h)=(f*g)*h\\
&(d)\ \ f*f\geq 0,\\
&\qquad\qquad f*f=0\Leftrightarrow f=0
\end{aligned} \tag{40}$$

易证(a) − (c),对于(d)中$f*f\geq 0$，可由 Gram 矩阵(34)的正定性得到。同时易证$f=0\Rightarrow\ f*f$。下证若$f*f=0$，那么$f=0$，即$\forall x\in\mathcal{X},f(x)=0$. 考虑到对于$\forall\lambda\in\mathbb{R}$,

$$\begin{aligned}
\big(f+\lambda k(\cdot,x)\big)*\big(f+\lambda k(\cdot,x)\big)&=(\boldsymbol{\alpha}^T,\lambda)\begin{pmatrix}\boldsymbol{K} & \boldsymbol{k}\\ \boldsymbol{k}^T & k(x,x)\end{pmatrix}\begin{pmatrix}\boldsymbol{\alpha}\\ \lambda\end{pmatrix}\\
&=\boldsymbol{\alpha}^T\boldsymbol{k}\boldsymbol{\alpha}+\lambda f(x)+\lambda^2 k(x,x)\\
&\geq 0
\end{aligned} \tag{41}$$

其中$\boldsymbol{K}$为 Gram 矩阵(34)，并且

$$\begin{aligned}
&\boldsymbol{k}=\big(k(x_1,x),k(x_2,x),\cdots,k(x_n,x)\big)^T\\
&\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_n)^T
\end{aligned} \tag{42}$$

公式(38)中不等式左侧是关于λ的二次函数。考虑到$f*f=0$，并且$k(x,x)\geq 0$，那么

$$\lambda f(x)+\lambda^2 k(x,x)\geq 0,\forall\lambda\in\mathbb{R} \tag{43}$$

因此$f(x)=0$.对于$\forall x\in\mathcal{X},f(x)=0$,所以

$$f=0 \tag{44}$$

由于$(\cdot*\cdot)$运算满足以上性质，因此它是线性空间$\mathcal{H}_0$上的内积。根据泛函分析中的完备化定理，可将$\mathcal{H}_0$完备化为希尔伯特空间$\mathcal{H}$。

定义映射$\phi\colon R^p\to\mathcal{H}$，满足

$$\begin{aligned}
&\phi(x)=k(\cdot,x)\\
&\phi(x)^T\phi(y)=k(x,y)=k(y,x)
\end{aligned} \tag{45}$$

映射ϕ可以将R^p空间中的样本集T，映射成高维空间$\mathcal{H}$中的样本集$\phi(T)$，并且$\mathcal{H}$中不同样本点之间的内积可以方便地由核函数(45)计算。希尔伯特空间$\mathcal{H}$中数据

集依然可以采用转换为对偶问题的思路，计算得到最优分类超平面。

5. 小结

以上就是 SVM 算法的基本思路，主要内容包括：最优分类平面的存在性与唯一性、拉格朗日对偶、松弛变量、核函数方法等。受限于作者水平，未进行深入讨论的内容包括：有限维R^p空间和无穷维希尔伯特空间上拉格朗日对偶定理的证明，它们是将拉格朗日原问题转换为对偶问题的关键，感兴趣的读者可以查阅相关凸优化文献了解。