『我爱机器学习』深入理解SVM(一) 原始问题和对偶问题

2018年3月19日 (https://www.hrwhisper.me/machine-learning-support-vector-machine-1/) hrwhisper (https://www.hrwhisper.me/author/hrsay/) 2 Comments (https://www.hrwhisper.me/machine-learning-support-vector-machine-1/#comments) 635 views

本文尽可能通俗、详细的介绍支持向量机SVM内容。

包括

- 线性可分SVM
- SVM对偶问题

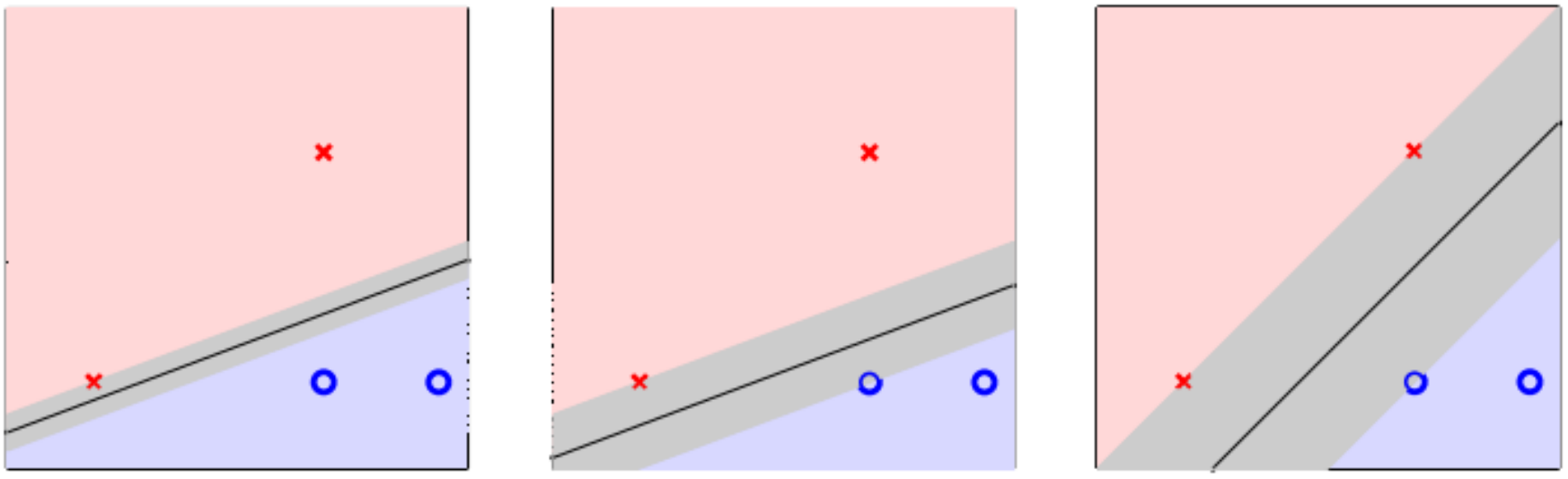
线性可分SVM

回顾一下感知机，

> 数据：$\{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \cdots, (\mathbf{x_n}, y_n)\},\ y \in \{-1, +1\}$
>
> 模型为：
>
> $$f(x) = \text{sign}(\mathbf{w^T x} + b = 0)$$
>
> 其损失函数为所有误分类点到超平面S的总函数距离：
>
> $$-\sum_{x_i \in M} y_i(\mathbf{w^T x_i} + b)$$
>
> 感知机选取不同的初值或选取不同的分类点，解可能不同。

PS: 支持向量机的输入和感知机是一样的。

那么感知机的许多解中，哪个比较好？比如下面的这三个：



直观上看，我们希望得到的是第三个结果，因为分类的边界离样本点远，这样，容错性（鲁棒性robust）会比较好。

仍然设我们的平面为$\mathbf{w}^{\mathbf{T}}\mathbf{x}+b$，我们有如下优化问题：

$$\begin{aligned} \max_{\mathbf{w},\mathbf{b}} \quad & \text{margin}(\mathbf{w}, b) \\ \text{s. t.} \quad & y_i(\mathbf{w}^{\mathsf{T}}\mathbf{x}_i + b) > 0, \qquad i = 1, 2..n \\ & \text{margin}(\mathbf{w}, b) = \min_{i=1,\ldots,m} \text{distance}(\mathbf{x}_i, \mathbf{w}, b) \end{aligned} \tag{1-1}$$

上面的第一个约束条件我们已经在感知机中见过，表达的意思就是全部样本都分类正确（这里我们先将设数据集线性可分），其中$y_i(\mathbf{w}^{\mathbf{T}}\mathbf{x_i}+b)$也叫做**函数距离**。第二个优化条件则是定义边界为样本到超平面最短的距离，这个距离稍后详解，我们的目标函数则是最大化它。

在感知机中，我们证明过点x到平面的距离为：

$$\text{distance}(\mathbf{x}_i, \mathbf{w}, b) = \frac{1}{\|\mathbf{w}\|}|\mathbf{w}^{\mathbf{T}}\mathbf{x_i}+b|$$

在数据集**线性可分**的情况下，分离超平面对数据集中每个点都有$y_i(w^T x_i + b) > 0$。因此，距离的计算公式又可以写为:

$$\text{distance}(\mathbf{x}_i, \mathbf{w}, b) = \frac{1}{\|\mathbf{w}\|}y_i(\mathbf{w}^{\mathbf{T}}\mathbf{x_i}+b) \tag{1-2}$$

1-2这个距离称为**几何距离**。函数距离和几何距离有啥关系呢？为什么要用几何距离呢？

几何距离就相当于函数距离做了规范化。函数距离只能表示分类的正确性，但是不同超平面的系数是可以放缩的，比如$\mathbf{w^T x_i} + b = 0$和 $3\mathbf{w^T x_i} + 3b = 0$ 所表示的超平面是一样的，但是函数距离后者是前者的3倍！因此，确定超平面的时候，应该使用几何距离。

于是，我们假定让点到超平面的函数最小几何距离为γ，式1-1可以写为：

$$\begin{aligned} \max_{\mathbf{w,b}} \quad & \gamma \\ \text{s. t.} \quad & \frac{y_i(\mathbf{w^T x_i} + b)}{\|\mathbf{w}\|} \geq \gamma, \qquad i = 1, 2, \cdots, n \end{aligned} \tag{1-3}$$

PS: 原来的$y_i(\mathbf{w^T x_i} + b) > 0$的条件限制不如新的条件，因此略去了。

考虑函数间隔和几何间隔的关系，可以将1-3写为：

$$\begin{aligned} \max_{\mathbf{w,b}} \quad & \frac{\gamma}{\|\mathbf{w}\|} \\ \text{s. t.} \quad & y_i(\mathbf{w^T x_i} + b) \geq \gamma, \qquad i = 1, 2, \cdots, n \end{aligned} \tag{1-3}$$

原来γ是点到超平面的最小几何距离，现在仍然是γ则是最小的函数距离，但是优化的目标仍然是几何距离。

现在，可以进一步化简，γ的取值其实对约束条件没有影响，也对优化目标没有影响。因为假设$\mathbf{w}$和b按比例变为$\lambda\mathbf{w}$和λb，则函数间隔也变成了$\lambda\gamma$（PS：同时可以看出几何间隔还是一样的）。因此，不妨令$\gamma = 1$，于是可以重写问题如下：

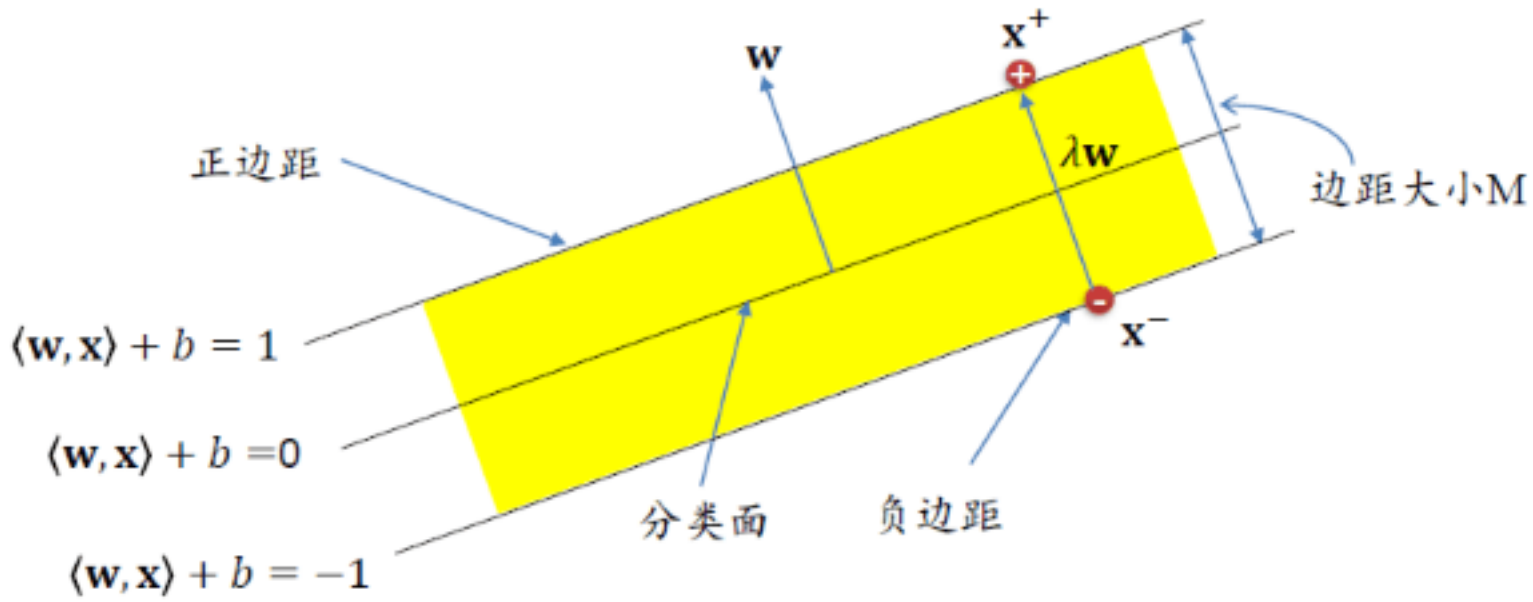
$$\begin{aligned} \max_{\mathbf{w,b}} \quad & \frac{1}{\|\mathbf{w}\|} \\ \text{s. t.} \quad & y_i(\mathbf{w^T x_i} + b) \geq 1, \qquad i = 1, 2, \cdots, n \end{aligned}$$

注意最大化 $\frac{1}{||w||}$ 和最小化$\frac{1}{2}||\mathbf{w}||^2 = \frac{1}{2}\mathbf{w^T w}$是等价的，因此有：

$$\begin{aligned} \min_{\mathbf{w,b}} \quad & \frac{1}{2}\mathbf{w^T w} \\ \text{s. t.} \quad & y_i(\mathbf{w^T x_i} + b) \geq 1, \qquad i = 1, 2, \cdots, n \end{aligned} \tag{1-4}$$

式1-4所表示的优化问题就称为**SVM的原始问题**。

在很多介绍SVM的资料中（包括西瓜书），对于SVM的原始问题都十分简单。它们都很直接的说：限制条件为$y_i(\mathbf{w}^T\mathbf{x_i}+b)\geq 1$，然后直接做类似如下推导：



$$\mathbf{w}^T\mathbf{x}^+ + b = +1$$

$$\mathbf{w}^T\mathbf{x}^- + b = -1$$

联立两式：$\mathbf{w}^T(\mathbf{x}^+ - \mathbf{x}^-) = 2$

因为：$\mathbf{x}^+ = \mathbf{x}^- + \lambda\mathbf{w}$ $\Rightarrow \lambda = \dfrac{2}{\mathbf{w}^T\mathbf{w}}$

间距和为：$|\mathbf{x}^+ - \mathbf{x}^-| = |\lambda\mathbf{w}| = \dfrac{2}{\mathbf{w}^T\mathbf{w}}|\mathbf{w}| = \dfrac{2}{|\mathbf{w}|}$

然后最大化$\dfrac{2}{\|w\|}$和最小化$\dfrac{1}{2}\mathbf{w}^T\mathbf{w}$等价

但其实限制条件是从上面这样缩放而来的，因此本文关于原问题的基本是按照林轩田老师和李航老师的思路写的。

这个图很直观的反应了**最大间隔分离超平面完全由离该超平面最近的点定义**，其它点如果从数据集中删除掉，对最优超平面的选择没有任何影响。这些决定了最大间隔分离超平面的点通常被称为**支持向量**，因此SVM的意思就是使用支持向量学习出最优的超平面。

二次规划求解

式1-4所表示的SVM原始问题是一个凸的二次规划问题，可以直接用现成的优化计算包求解。

在此之前，需要表示为标准形式。

二次规划的标准形式为：

$$\begin{aligned} \min_{\mathbf{u}} \quad & \frac{1}{2}\mathbf{u}^{\mathsf{T}}\mathrm{Q}\mathbf{u} + \mathbf{p}^{\mathsf{T}}\mathbf{u} \\ \text{s. t.} \quad & \mathbf{a}_i^{\mathsf{T}}\mathbf{u} \geq c_i, \text{for } i = 1, 2, \ldots, M \end{aligned}$$

根据和1-4的对应关系，设维度为d得：

$$\begin{aligned} \mathbf{u} &= \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix} \\ \mathrm{Q} &= \begin{bmatrix} 0 & \mathbf{0}_d^{\mathsf{T}} \\ \mathbf{0}_d & \mathrm{I}_d \end{bmatrix} \\ \mathbf{p} &= \mathbf{0}_{d+1} \\ \mathbf{a}_n^{\mathsf{T}} &= y_n \begin{bmatrix} 1 & \mathbf{x}_n^{\mathsf{T}} \end{bmatrix} \\ c_n &= 1 \\ M &= n \end{aligned}$$

SVM对偶问题

虽然可以求解1-4的SVM原始问题，但是我们可以利用拉格朗日乘子法得到对偶问题。这样的好处是：

1. 对偶问题往往更容易求解
2. 自然引入核函数，进而推广到非线性分类问题

好处2在下一小结再讲。现在讲讲好处1。

在线性回归一章中，讲到可以进行空间的变换。

设$\mathbf{z} = \Phi(\mathbf{x})$, 则SVM 原始问题写为：

$$\begin{aligned} \min_{\mathbf{w},\mathbf{b}} \quad & \frac{1}{2}\mathbf{w}^{\mathbf{T}}\mathbf{w} \\ \text{s. t.} \quad & y_i(\mathbf{w}^{\mathbf{T}}\boldsymbol{\Phi}(\mathbf{x}) + b) \geq 1, \qquad i = 1, 2, \cdots, n \end{aligned}$$

x经过非线性特征转换以后，映射到新空间里得到的z通常维度比较高。

记映射前维度为d, 映射后向量的维度为$\tilde{d}$，则求解二次规划时需要面对$\tilde{d} + 1$个变量(**w**和b)和m个限制条件。

因此，需要探索一种方法使得SVM不去依赖$\tilde{d}$。

拉格朗日乘子法

首先讲解拉格朗日乘子法，给定优化问题：

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{s. t.} \quad & g_i(\mathbf{x}) \leq 0, \ \text{for}\, i = 1, 2, \ldots, m \\ & h_j(\mathbf{x}) = 0, \ \text{for}\, j = 1, 2, \ldots, n \end{aligned}$$

可以写出拉格朗日函数：

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\mathbf{x}) + \sum_{i=1}^{m} \alpha_i g_i(\mathbf{x}) + \sum_{j=1}^{n} \beta_j h_j(\mathbf{x}), \alpha_i \geq 0$$

则原问题等价于：

$$\begin{aligned} \min_{\mathbf{x}} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad & \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ \text{s. t.} \quad & \alpha_i \geq 0 \end{aligned} \tag{2-1}$$

为什么呢？做个简单的推导：

$$\min_{\mathbf{x}} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

$$= \min_{\mathbf{x}} \left(f(x) + \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left(f(x) \sum_{i=1}^{m} \alpha_i g_i(\mathbf{x}) + \sum_{j=1}^{n} \beta_j h_j(\mathbf{x}) \right) \right)$$

$$= \min_{\mathbf{x}} \left(f(x) + \begin{cases} 0 & \text{u满足约束条件} \\ \infty & \text{otherwise} \end{cases} \right)$$

- 当g不满足约束条件的时候（$g_i(x) > 0$），我们内层优化取max，因此
  $\alpha_i = \infty \Rightarrow \alpha_i g_i(\mathbf{x}) = \infty$
- 当g满足约束条件的时候（$g_i(x) \leq 0$），我们内层优化取max，因此
  $\alpha_i = 0 \Rightarrow \alpha_i g_i(\mathbf{x}) = 0$
- 当h不满足约束条件的时候（$h_j(x) \neq 0$），同理可以取
  $\beta_j = \text{sign}(\text{h}_\text{j}(\mathbf{x}))\infty \Rightarrow \beta_\text{j} \text{h}_\text{j}(\mathbf{x}) = \infty$
- 当h满足约束条件的时候（$h_j(x) = 0$），同理可以取

$$\beta_j h_j(\mathbf{x}) = 0$$

为了使2-1达到优值，需要满足如下条件（**KKT条件**）：

- 主问题可行： $g_i(\mathbf{x}) \leq 0, h_i(\mathbf{x}) = 0$
- 对偶问题可行： $\alpha_i \geq 0$
- 互补松弛: $\alpha_i g_i(\mathbf{x}) = 0$

主问题可行是上面推导的结果，对偶问题可行为2-1的约束项。

互补松弛是主问题和对偶问题都可行的条件下的最大值。

拉格朗日对偶问题

定义2-1的对偶问题为：

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \min_{\mathbf{x}} \quad \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \qquad (2\text{-}2)$$

$$\text{s. t.} \quad \alpha_i \geq 0$$

对偶问题是原问题的下界，即：

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \leq \min_{\mathbf{x}} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

因为总是有 $\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \leq \min_{\mathbf{x}} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ 成立，对任意可能的值求极值，肯定不如直接求到的极值大。有句俗语叫“瘦死的骆驼比马大”说的就是这个道理。

2-2只是交换了max-min的顺序，得到的就是**拉格朗日对偶问题**。

SVM 对偶问题推导

首先变换SVM原问题为：

$$\begin{aligned} \min_{\mathbf{w,b}} \quad & \frac{1}{2}\mathbf{w}^{\mathbf{T}}\mathbf{w} \\ \text{s. t.} \quad & 1-y_i(\mathbf{w}^{\mathbf{T}}\mathbf{x_i}+b) \leq 0, \qquad i=1,2,\cdots,n \end{aligned}$$

可以写出拉格朗日函数如下：

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^{\mathbf{T}}\mathbf{w} + \sum_{i=1}^{n} \alpha_i \left(1-y_i(\mathbf{w}^{\mathbf{T}}\mathbf{x_i}+b)\right), \alpha_i \geq 0 \tag{2-3}$$

根据拉格朗日对偶性，得到对偶问题为：

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \min_{\mathbf{w},b} \quad & \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) \\ \text{s. t.} \quad & \alpha_i \geq 0, i=1,2,\cdots,n \end{aligned} \tag{2-4}$$

求解$\min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})$，拉格朗日函数2-3分别对**w**和b求偏导，得：

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x_i}$$

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i$$

令偏导为0，得:

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x_i} \tag{2-5}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{2-6}$$

带入拉格朗日函数2-3得：

$$\min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x_i^T x_j}$$

再对$\min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})$求极大值得到对偶问题：

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x_i^T x_j} \\ \text{s. t.} \quad & \alpha_i \geq 0, i = 1, 2, \cdots, n \\ & \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned} \tag{2-7}$$

注意我们把$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x_i}$去掉了，因为这里求的是$\alpha$使得问题最大，和$\mathbf{w}$无关。

2-7就是我们的**SVM对偶问题**。需要满足的KKT条件为：

- 主问题可行： $1 - y_i(\mathbf{w^T x_i} + b) \leq 0$
- 对偶问题可行：$\alpha_i \geq 0$
- 互补松弛: $\alpha_i \left(1 - y_i(\mathbf{w^T x_i} + b)\right) = 0$
- 2-5和2-6的条件: $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x_i}$; $\sum_{i=1}^{n} \alpha_i y_i = 0$

PS: 周志华老师的KKT条件中没有2-5和2-6两式，而李航老师和林轩田老师的都有。 笔者认为也应该加上:)

这里有个很有意思的结论： 由于我们总有$\alpha_i \geq 0$，当$\alpha_i > 0$时，$y_i(\mathbf{w^T x_i} + b) = 1$,此时所对应的样本点位于最大间隔边界上，是一个**支持向量**。这显示出一个重要的性质：训练完之后，大部分的训练样本都不需要保留，**最终模型只与支持向量有关**（同时也暗示复杂度主要与支持向量的数目有关）。

求解对偶问题

2-7所表示的SVM对偶问题仍可以用二次规划算法来求解，但是问题规模正比于训练样本数，在实际任务中开销很大。往往采用钆更高效的算法如**SMO**或者 **Pegasos** ， 这将在最后介绍。

假设求解出了α, 那么根据KKT条件, 可得：

- $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x_i}$,
- 选取一个$\alpha_i > 0$ 可得$y_i(\mathbf{w^T x_i} + b) = 1 \Rightarrow \mathbf{w^T x_i} + b = y_i \Rightarrow b = y_i - \mathbf{w^T x_i}$

观察对偶支持向量机求解可以发现对偶问题的好处为：

1. 只需要优化α而不是b和**w**，降低了算法的时间复杂度
2. 通过查看$a_i > 0$ 便可找出支持向量

求得**w**和b即可构造超平面和对应的决策函数，和感知机一样：

$$g_{\text{svm}}(\mathbf{x}) = \text{sign}(\mathbf{w^T x} + \text{b})$$

小结

若进行空间变换$\mathbf{z} = \Phi(\mathbf{x})$则：

$$g_{\text{svm}}(\mathbf{z}) = \text{sign}(\mathbf{w^T z} + \text{b})$$

原始的SVM和对偶SVM对比如下表：

	变量个数	适用情况	物理意义
原始SVM	$\tilde{d}+1$	数据维度较小	对(**w**, b)进行缩放，得到一个合适的值
对偶SVM	n	数据量比较小	找出支持向量和对应的α

回想一下SVM对偶问题的初衷是避免对对数据维度$\tilde{d}$的依赖。虽然我们的对偶问题看上去最后求解时只依赖了数据量n，但是看看目标函数将会有一个$\mathbf{z_i^T z_j}$的内积（将x都映射到z），直接计算的话复杂度仍是$O(\tilde{d})$，如何避免内积计算，请看下一篇 SVM的核函数讲解。

参考资料

- 机器学习技法 – 林轩田
- 《统计学习方法》 – 李航
- 《机器学习》 – 周志华
- 《从零构建支持向量机(SVM)》 – 张皓

机器学习 (https://www.hrwhisper.me/category/study/machine-learning/) Machine Learning (https://www.hrwhisper.me/tag/machine-learning/), Machine Learning model (https://www.hrwhisper.me/tag/machine-learning-model/). permalink (https://www.hrwhisper.me/machine-learning-support-vector-machine-1/).

『我爱机器学习』正则化 (https://www.hrwhisper.me/machine-learning-regularization/)

『我爱机器学习』深入理解SVM(二) – 核函数和软边距 (https://www.hrwhisper.me/machine-learning-support-vector-machine-2-kernel-function-and-soft-margin-svm/)

---

2 thoughts on “『我爱机器学习』深入理解SVM(一) 原始问题和对偶问题”

昊霖 says:
2018年6月7日 at am11:51 (https://www.hrwhisper.me/machine-learning-support-vector-machine-1/#comment-10675)

您好，显示异常了..latex代码都显示出来了呢..

Reply

hrwhisper (https://www.hrwhisper.me) says:
2018年6月7日 at pm3:56 (https://www.hrwhisper.me/machine-learning-support-vector-machine-1/#comment-10682)

这是因为你的js没有加载出来。。。

Reply

Leave a Reply

Your email address will not be published. Required fields are marked *

Comment

Name *

Email *

Website

Save my name, email, and website in this browser for the next time I comment.

Post Comment

(http://weibo.com/murmured)

(http://www.linkedin.com/in/huangrong-yang-548632119/)

(https://instagram.com/hr_say/)

(https://github.com/hrwhisper)

(https://www.hrwhisper.me/feed)

Csdn (http://blog.csdn.net/murmured)　　博客园 (http://www.cnblogs.com/murmured/)

Lofter (http://hrsay.lofter.com/)　　知 乎 (http://www.zhihu.com/people/hrwhisper)

豆瓣 (http://www.douban.com/people/hrwhisper/)

努力的人本身就有奇迹 | 快乐是我们共同的信仰

by hrwhipser.me