机器学习导论
习题三

191250026, 丁云翔, 191250026@smail.nju.edu.cn

2021 年 4 月 22 日

学术诚信

本课程非常重视学术诚信规范，助教老师和助教同学将不遗余力地维护作业中的学术诚信规范的建立。希望所有选课学生能够对此予以重视。[1]

(1) 允许同学之间的相互讨论，但是**署你名字的工作必须由你完成**，不允许直接照搬任何已有的材料，必须独立完成作业的书写过程;

(2) 在完成作业过程中，对他人工作（出版物、互联网资料）中文本的直接照搬（包括原文的直接复制粘贴及语句的简单修改等）都将视为剽窃，剽窃者成绩将被取消。**对于完成作业中有关键作用的公开资料，应予以明显引用**；

(3) 如果发现作业之间高度相似将被判定为互相抄袭行为，**抄袭和被抄袭双方的成绩都将被取消**。因此请主动防止自己的作业被他人抄袭。

作业提交注意事项

(1) 请在**第一页填写个人的姓名、学号、邮箱信息**；

(2) 本次作业需提交该pdf文件， pdf文件名格式为**学号_姓名.pdf**， 例如190000001_张三.pdf，**需通过教学立方提交**。

(3) 未按照要求提交作业，或提交作业格式不正确，将会**被扣除部分作业分数**；

(4) 本次作业提交截止时间为**4月25日23:55:00**。

[1]参考尹一通老师高级算法课程中对学术诚信的说明。

1 [30pts] Binary Split of Attributes

本题尝试讨论决策树构建过程中的一种属性划分策略。我们已经知道，决策树学习中的一个关键问题是如何选择最优划分属性。一般来讲，我们可以使用贪心策略，基于某种指标（信息增益、Gini指数等）选择当前看来最好的划分属性，并对其进行划分。然而，在获得了最优属性A后，如何对其进行划分也是一个重要的问题。如果A是一个无序的离散属性，我们可以在当前节点考虑A的所有可能取值，从而对节点进行划分；如果A是一个有序的连续属性，则可以考虑将其离散化，并将该属性划分为多个区间。

(1) **[9pts]** (二分划分) 考虑当前的划分属性A，假设它是一个离散属性且有K个不同的取值，我们可以依据A将当前节点划分成K份。另一种策略是“二分划分”，即将K个不同取值划分为两个不相交的集合，并由此将当前节点划分为两份。在之后的节点中，仍然允许再次选择A作为划分属性。相较于将当前节点直接划分为K份，请定性地说明二分划分策略有何优势；

(2) **[6pts]** (K较大的情况) 二分划分策略在K较大时会遇到困难，因为此时将属性取值集合划分为两个不相交子集的方案数很多。试计算该方案数。注意：划分得到的两个子集是没有顺序的，且“全集+空集”这种划分不计算在内。

(3) **[15pts]** (特殊情况：二分类) 考虑一个二分类问题。如果使用二分划分策略对属性集$A = \{a_1, \ldots, a_k, \ldots, a_K\}$进行划分，且$K$较大，下面这种策略是一个不错的选择：首先，统计在属性A上取值为a_k的样本为**正类**的概率$p_k = \mathbf{Prob}[y = +1|A = a_k]$，并以$p_k$为键值对$K$个属性取值排序。不失一般性，我们假设$p_1 \leq \cdots \leq p_k \cdots \leq p_K$。之后，我们将属性$A$当作是有序属性，寻找一个最优的$\bar{k}$，将属性集划分为子集$\{a_1, \ldots, a_{\bar{k}}\}$和$\{a_{\bar{k}+1}, \ldots, a_K\}$，并由此将当前节点划分为两个子节点。请尝试分析该策略的合理性。

Solution. 此处用于写解答(中英文均可)

(1)

二分划分策略对每种划分属性不要求将所有的取值都划分开，在单个划分属性的属性取值较多但样本数量较少时，可以保证划分以后各分支的样本数不会过少，可以降低过拟合的风险，提高泛化能力；且一次划分产生的分支大大减少，能降低决策树的宽度和复杂度，提高决策树生成的效率。

(2)

当K为奇数时，划分产生的两个不相交子集大小分别为1到$\frac{K-1}{2}$和$K-1$到$\frac{K+1}{2}$，共$\frac{K-1}{2}$对划分子集大小，故二分划分策略的方案数为：

$$\sum_{i=1}^{\frac{K-1}{2}} \binom{K}{i} = 2^{K-1} - 1$$

当K为偶数时，划分产生的两个不相交子集大小分别为1到$\frac{K}{2}$和$K-1$到$\frac{K}{2}$，共$\frac{K}{2}$对划分子集大小，故二分划分策略的方案数为：

$$\sum_{i=1}^{\frac{K}{2}} \binom{K}{i} - \frac{1}{2}\binom{K}{\frac{K}{2}} = 2^{K-1} - 1$$

综上，划分的方案数为$2^{K-1}-1$.

(3)

由于题给情况是一个二分类问题，故最终是需要将正类与负类划分开，所以每一次属性划分都应该能尽量将正类与负类划分开，即其中一个分支中样本为正类的概率尽可能高，另一个分支中样本为负类的概率尽可能高（即样本为正类的概率尽可能低）。而这一思想的形式化体现为基尼指数，该二分类问题的二分划分情况下基尼指数如下：

$$Gini\_index(D,a)=\sum_{v=1}^{2}\frac{|D^v|}{|D|}(1-\sum_{k=1}^{2}p_{v,k}^2)$$

基尼指数越小，代表划分后产生的两个数据集的纯度越高，即在划分后产生的一个分支中样本为正类的概率越高，且另一个分支中样本为负类的概率越高。

以最小化基尼指数为二分划分的划分依据。若采取遍历搜索的策略，其找到的最优解为使基尼指数最小的划分方法，即将所有属性取值分为该属性取值的样本为正类的概率高的一类和该属性取值的样本为负类的概率高（即样本为正类的概率低）的一类，其必满足前者中任一属性取值的样本为正类的概率大于后者中任一属性取值的样本为正类的概率，这样才能使划分后的两分支数据纯度达到最优，从而使基尼指数最小。而第(3)问中的策略将属性取值按该属性取值的样本为正类的概率排序，并选定一个最优的$\bar{k}$，将属性集合分为样本为正类的概率较高和样本为正类的概率较低（即样本为负类的概率较高）的两类，且前者中任一属性取值的样本为正类的概率大于后者中任一属性取值的样本为正类的概率。由于两种策略的优化目标都是使划分后基尼指数最小，且最优划分都满足划分后一个分支中任一属性取值的样本为正类的概率均大于另一个分支中任一属性取值的样本为正类的概率。发现两种策略的优化目标一致，应得到的最优解形式也一致，所以最优解也一致，且该策略的划分效率大幅提高，所以是合理的。

2 [70pts] Review of Support Vector Machines

在本题中，我们复习支持向量机（SVM）的推导过程。考虑N维空间中的二分类问题，即$\mathbb{X}=\mathbb{R}^N,\mathbb{Y}=\{-1,+1\}$。现在，我们从某个数据分布$\mathcal{D}$中采样得到了一个包含$m$个样本的数据集$S=\{(\mathbf{x}_i,y_i)\}_{i=1}^m$。令$h:\mathbb{X}\to\mathbb{Y}$表示某个线性分类器，即$h\in\mathcal{H}=\{\mathbf{x}\to\text{sign}(\mathbf{w}^\top\mathbf{x}+b):\mathbf{w}\in\mathbb{R}^N,b\in\mathbb{R}\}$。始终假设数据是线性可分的。

(1) **[3pts]** 对于一个样本$(\mathbf{x},y)$，请用包含$\mathbf{x},y,\mathbf{w},b$的不等式表达“该样本被分类正确”；

(2) **[3pts]** 我们知道，线性分类器h对应的方程$\mathbf{w}^\top\mathbf{x}+b=0$确定了$N$维空间中的一个超平面。令$\rho_h(\mathbf{x})$表示点$\mathbf{x}$到由$h$确定的超平面的欧式距离，试求$\rho_h(\mathbf{x})$；

(3) **[4pts]** 定义分类器h的间隔$\rho_h=\min_{i\in[m]}\rho_h(\mathbf{x})$。现在，我们希望在$\mathcal{H}$中寻找“能将所有样本分类正确且间隔最大”的分类器。试写出该优化问题。我们将该问题称为问题$\mathcal{P}^1$；

(4) **[5pts]** $\mathcal{P}^1$是一个关于参数$\mathbf{w},b$的优化问题。然而，该问题有无穷多组最优解。请证明该结论；

(5) **[5pts]** 虽然$\mathcal{P}^1$有无穷多组最优解，但这些最优解将给出等价的分类器。所以，我们希望对$\mathcal{P}^1$做一些限制。一般情况下，我们可以要求$\min_{i\in[m]}|\mathbf{w}^\top\mathbf{x}+b|=1$（或者等价地，$\min_{i\in[m]}y_i(\mathbf{x}^\top\mathbf{x}_i+b)=1$），此时$\mathcal{P}^1$将转化为优化问题$\mathcal{P}^2$。试写出$\mathcal{P}^2$；

(6) **[5pts]** 问题$\mathcal{P}^2$的优化目标中涉及参数$\mathbf{w}$的范数的倒数。如果将最大化$\frac{1}{\|\mathbf{w}\|}$转化为最小化$\frac{\|\mathbf{w}\|^2}{2}$，我们就可以得到问题$\mathcal{P}^3$。试写出$\mathcal{P}^3$。

(7) **[5pts]** 试推导问题$\mathcal{P}^3$的对偶问题$\mathcal{P}^4$，给出过程；

(8) **[10pts]** 描述“凸优化问题”的定义，并证明$\mathcal{P}^3$和$\mathcal{P}^4$都是凸优化问题；

(9) **[5pts]** 既然$\mathcal{P}^3$和$\mathcal{P}^4$都是凸优化问题，为什么我们要对$\mathcal{P}^3$做对偶操作？或者说，在这里使用对偶有什么好处？

(10) **[10pts]** 设$\{\boldsymbol{\alpha}_i^\star\}_{i=1}^m$是对偶问题$\mathcal{P}^4$的最优解，设$(\mathbf{w}^\star, b^\star)$是原问题$\mathcal{P}^3$的最优解。请使用$\{\boldsymbol{\alpha}_i^\star\}_{i=1}^m$和$\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ 表达$\mathbf{w}^\star$和$b^\star$，给出过程；

(11) **[10pts]** 利用上一小问中的结果，经过一些代数运算，我们发现：可以只用$\{\boldsymbol{\alpha}_i^\star\}_{i=1}^m$简洁地表达$\|\mathbf{w}^\star\|^2$。请写出这个表达，给出推导过程；

(12) **[5pts]** 我们注意到，在问题$\mathcal{P}^2$中，分类器的间隔由式子$\frac{1}{\|\mathbf{w}\|}$表达。再结合上一小问的结果，你可以得到何种启发？

Solution. 此处用于写解答(中英文均可)

(1)

若样本$(\mathbf{x}, y)$被分类正确，则有$y = +1$时，$\mathbf{w}^\top\mathbf{x} + b > 0$；$y = -1$时，$\mathbf{w}^\top\mathbf{x} + b < 0$，即：

$$y(\mathbf{w}^\top\mathbf{x} + b) > 0$$

(2)

设点$\mathbf{x}_0$为点$\mathbf{x}$在超平面上的投影点，则$|\mathbf{x}-\mathbf{x}_0|$即为点$\mathbf{x}$到由$h$确定的超平面的欧式距离，即$\rho_h(\mathbf{x})$。由于向量$\mathbf{x} - \mathbf{x}_0$与超平面法向量$\mathbf{w}$平行，所以有：

$$\begin{aligned} |\mathbf{w}\cdot(\mathbf{x}-\mathbf{x}_0)| &= |\mathbf{w}||\mathbf{x}-\mathbf{x}_0| \\ &= ||\mathbf{w}||\rho_h(\mathbf{x}) \end{aligned} \tag{1}$$

由向量内积的运算规则可知$\mathbf{w}^\top\mathbf{x}_0 = \mathbf{w}\cdot\mathbf{x}_0$，$\mathbf{w}^\top\mathbf{x} = \mathbf{w}\cdot\mathbf{x}$。又点$\mathbf{x}_0$在超平面上，故有$\mathbf{w}^\top\mathbf{x}_0 + b = 0$，即$\mathbf{w}\cdot\mathbf{x}_0 + b = 0$，所以有：

$$\begin{aligned} \mathbf{w}\cdot(\mathbf{x}-\mathbf{x}_0) &= \mathbf{w}\cdot\mathbf{x} - \mathbf{w}\cdot\mathbf{x}_0 \\ &= \mathbf{w}\cdot\mathbf{x} + b \\ &= \mathbf{w}^\top\mathbf{x} + b \\ |\mathbf{w}\cdot(\mathbf{x}-\mathbf{x}_0)| &= |\mathbf{w}^\top\mathbf{x} + b| \end{aligned} \tag{2}$$

联立(1)式和(2)式，则有：

$$\begin{aligned} ||\mathbf{w}||\rho_h(\mathbf{x}) &= |\mathbf{w}^\top\mathbf{x} + b| \\ \rho_h(\mathbf{x}) &= \frac{|\mathbf{w}^\top\mathbf{x} + b|}{||\mathbf{w}||} \end{aligned}$$

(3)

在$\mathcal{H}$中寻找“能将所有样本分类正确且间隔最大”的分类器，即在$\mathcal{H}$中寻找一个分类器h使间隔$\rho_h = \min_{i\in[m]} \rho_h(\mathbf{x}_i)$最大化，并附加约束$y_i(\mathbf{w}^\top \mathbf{x}_i + b) > 0$：

$$\begin{aligned} &\max_{h\in\mathcal{H}} \min_{i\in[m]} \rho_h(\mathbf{x}_i) \\ &s.t. \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) > 0, \quad i \in [m] \end{aligned}$$

又线性分类器h实际由参数$\mathbf{w}$和b决定，且$\rho_h(\mathbf{x}) = \frac{|\mathbf{w}^\top \mathbf{x}+b|}{||\mathbf{w}||}$，所以上述优化问题可以改写为：

$$\begin{aligned} &\max_{\mathbf{w},b} \min_{i\in[m]} \frac{|\mathbf{w}^\top \mathbf{x}_i + b|}{||\mathbf{w}||} \\ &s.t. \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) > 0, \quad i \in [m] \end{aligned}$$

此即问题$\mathcal{P}^1$。

(4)

对于优化问题$\mathcal{P}^1$，由参数空间和优化目标函数自身的性质可知，其一定存在最优解。又其优化目标函数$\frac{|\mathbf{w}^\top \mathbf{x}+b|}{||\mathbf{w}||}$是一个齐次式，即当$(\mathbf{w}^*, b^*)$为该优化问题的最优解时，对于任意常数$c > 0$，$(c\mathbf{w}^*, cb^*)$也为该优化问题的最优解，因为有：

$$\frac{|c\mathbf{w}^{*\top} \mathbf{x} + cb^*|}{||c\mathbf{w}^*||} = \frac{|\mathbf{w}^{*\top} \mathbf{x} + b^*|}{||\mathbf{w}^*||}$$

所以该问题有无穷多组最优解。

(5)

对优化问题$\mathcal{P}^1$添加限制条件$\min_{i\in[m]} |\mathbf{w}^\top \mathbf{x}_i+b| = 1$，由于$|y| = 1$，此条件等价于$\min_{i\in[m]} |y_i(\mathbf{w}^\top \mathbf{x}_i+b)| = \min_{i\in[m]} |y_i||\mathbf{w}^\top \mathbf{x}_i+b| = 1$，又原限制条件有$y_i(\mathbf{w}^\top \mathbf{x}_i+b) > 0$，则$\min_{i\in[m]} y_i(\mathbf{w}^\top \mathbf{x}_i+b) = \min_{i\in[m]} |y_i(\mathbf{w}^\top \mathbf{x}_i + b)| = 1$，得到优化问题$\mathcal{P}^2$：

$$\begin{aligned} &\max_{\mathbf{w},b} \frac{1}{||\mathbf{w}||} \\ &s.t. \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \quad i \in [m] \end{aligned}$$

(6)

将最大化$\frac{1}{\|\mathbf{w}\|}$转化为最小化$\frac{\|\mathbf{w}\|^2}{2}$，我们就可以得到问题$\mathcal{P}^3$：

$$\begin{aligned} &\min_{\mathbf{w},b} \frac{\|\mathbf{w}\|^2}{2} \\ &s.t. \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \quad i \in [m] \end{aligned} \tag{3}$$

(7)

对问题$\mathcal{P}^3$中的优化目标函数使用拉格朗日乘子法，对每条约束添加拉格朗日乘子$\alpha_i \geq 0$，则该问题的拉格朗日函数可写为：

$$\mathbf{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{\|\mathbf{w}\|^2}{2} + \sum_{i=1}^{m} \alpha_i \left(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\right) \tag{4}$$

其中$\boldsymbol{\alpha} = (\alpha_1; \alpha_2; ...; \alpha_m)$.令$\mathbf{L}(\mathbf{w}, b, \boldsymbol{\alpha})$对$\mathbf{w}$和$b$的偏导为零可得：

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \tag{5}$$

$$0 = \sum_{i=1}^{m} \alpha_i y_i \tag{6}$$

将(5)式和(6)式代入(4)式，原拉格朗日函数可以化简为：

$$\begin{aligned}
\mathbf{L}(\mathbf{w}, b, \boldsymbol{\alpha}) &= \frac{\|\mathbf{w}\|^2}{2} + \sum_{i=1}^{m} \alpha_i \left(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\right) \\
&= \frac{\mathbf{w}^\top \mathbf{w}}{2} + \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \alpha_i y_i \mathbf{w}^\top \mathbf{x}_i - \sum_{i=1}^{m} \alpha_i y_i b \\
&= \frac{1}{2}\mathbf{w}^\top \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i + \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \alpha_i y_i \mathbf{w}^\top \mathbf{x}_i - \sum_{i=1}^{m} \alpha_i y_i b \\
&= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \alpha_i y_i \mathbf{w}^\top \mathbf{x}_i - b \sum_{i=1}^{m} \alpha_i y_i \\
&= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \alpha_i y_i (\sum_{j=1}^{m} \alpha_j y_j \mathbf{x}_j^\top) \mathbf{x}_i \\
&= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \mathbf{x}_j^\top \mathbf{x}_i
\end{aligned} \tag{7}$$

将(7)式作为优化目标函数，(6)式作为约束，即得到问题$\mathcal{P}^3$的对偶问题$\mathcal{P}^4$：

$$\begin{aligned}
&\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \mathbf{x}_j^\top \mathbf{x}_i \\
&s.t. \quad \sum_{i=1}^{m} \alpha_i y_i = 0, \\
&\qquad \alpha_i \geq 0, \quad i \in [m]
\end{aligned} \tag{8}$$

(8)

“凸优化问题”是形如：

$$\begin{aligned}
&minimize \quad f_0(x) \\
&subject\ to \quad f_i(x) \leq 0, \quad i = 1, ..., m \\
&\qquad\qquad\ \ a_i^T x = b_i, \quad i = 1, ..., p
\end{aligned}$$

的问题，其中$f_0, f_1, ..., f_m$均为凸函数。

对于$\mathcal{P}^3$

$$\begin{aligned}
&\min_{\mathbf{w}, b} \frac{\|\mathbf{w}\|^2}{2} \\
&s.t. \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \quad i \in [m]
\end{aligned}$$

目标函数为$f_0(w,b) = \frac{\|\mathbf{w}\|^2}{2}$，约束函数为$f_i(w,b) = y_i(\mathbf{w}^\top \mathbf{x}_i + b)$。目标函数$f_0(w,b)$为一个开口向上的二次函数，二阶导数大于零，故目标函数为凸函数。又由于y_i的取值只有+1和−1，且i确定时y_i的取值不变，所以约束函数$f_i(w,b)$为线性函数。故约束函数也为凸函数，该优化问题为凸优化问题。

对于$\mathcal{P}^4$，

$$\begin{aligned}\max_{\boldsymbol{\alpha}} & \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \mathbf{x}_j^\top \mathbf{x}_i \\ s.t. \quad & \sum_{i=1}^{m}\alpha_i y_i = 0, \\ & \alpha_i \geq 0, \quad i \in [m]\end{aligned}$$

目标函数为$f_0(\boldsymbol{\alpha}) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \mathbf{x}_j^\top \mathbf{x}_i$，约束函数为$f_i(\boldsymbol{\alpha}) = \alpha_i$和$f(\boldsymbol{\alpha}) = \sum_{i=1}^{m}\alpha_i y_i$。最大化目标函数即最小化目标函数取负得到的函数，$f_0(\boldsymbol{\alpha})$为线性函数减去一个开口向上的类二次函数，则$-f_0(\boldsymbol{\alpha})$为一个开口向上的类二次函数减去线性函数，二阶导数大于等于零，故目标函数取负得到的函数为凸函数。约束函数$f_i(\boldsymbol{\alpha})$和$f(\boldsymbol{\alpha})$均为线性函数，故约束函数也为凸函数，该优化问题为凸优化问题。

(9)

使用对偶操作后可以对对偶问题应用SMO算法求解，故不需要优化计算包即可高效求解；对偶操作将原问题中的不等式约束条件转换为了对偶问题中的等式约束条件，更利于求解最优解；在对偶问题里，原问题的最优解只与$\alpha_i > 0$的样本（即支持向量）有关，所以在求解模型参数时只需考虑支持向量而不需要全部样本，大幅提高了参数求解的效率；在对偶问题中，$\mathbf{x}$总是以内积形式成对出现，有利于在SVM中引入核函数，从而解决线性不可分问题。

(10)

$\{\boldsymbol{\alpha}_i^\star\}_{i=1}^m$是对偶问题$\mathcal{P}^4$的最优解，则由(5)式可直接得到$\mathbf{w}^\star$:

$$\mathbf{w}^\star = \sum_{i=1}^{m}\boldsymbol{\alpha}_i^\star y_i \mathbf{x}_i$$

又原优化问题满足KKT条件：

$$\begin{cases}\alpha_i \geq 0 \\ y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1 \geq 0 \\ \alpha_i\left(y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1\right) = 0\end{cases}$$

由第三个条件可以求出可以求出$b^\star$.由于对任意支持向量都有$\alpha_i > 0$，由KKT条件中的第三条，可知对任意支持向量$(\mathbf{x}_s, y_s)$都有$y_s(\mathbf{w}^\top \mathbf{x}_s + b) = 1$。设$S = \{i \mid \boldsymbol{\alpha}_i^\star > 0,\ i = 1, 2, ..., m\}$为所有支持向量的下标集，则$b^\star$可表示为：

$$b^\star = \frac{1}{y_s} - \sum_{i\in S}\boldsymbol{\alpha}_i^\star y_i \mathbf{x}_i^\top \mathbf{x}_s \tag{9}$$

为提高鲁棒性，也可使用所有支持向量求解出b的平均值作为$b^\star$:

$$b^\star = \frac{1}{|S|}\sum_{s\in S}\left(\frac{1}{y_s} - \sum_{i\in S}\boldsymbol{\alpha}_i^\star y_i \mathbf{x}_i^\top \mathbf{x}_s\right) \tag{10}$$

(11)

$$\begin{aligned}\|\mathbf{w}^\star\|^2 &= \mathbf{w}^{\star\top}\mathbf{w}^\star \\ &= \sum_{i=1}^{m}\boldsymbol{\alpha}_i^\star y_i\mathbf{x}_i^\top \sum_{i=1}^{m}\boldsymbol{\alpha}_i^\star y_i\mathbf{x}_i \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m}\boldsymbol{\alpha}_i^\star\boldsymbol{\alpha}_j^\star y_i y_j\mathbf{x}_i^\top\mathbf{x}_j \end{aligned} \tag{11}$$

由于只有支持向量对应的$\boldsymbol{\alpha}_i^\star > 0$，否则$\boldsymbol{\alpha}_i^\star = 0$，所以(11)式等价于：

$$\|\mathbf{w}^\star\|^2 = \sum_{i\in S}\sum_{j\in S}\boldsymbol{\alpha}_i^\star\boldsymbol{\alpha}_j^\star y_i y_j\mathbf{x}_i^\top\mathbf{x}_j \tag{12}$$

设$S = \{i \mid \boldsymbol{\alpha}_i^\star > 0,\ i = 1, 2, ..., m\}$为所有支持向量的下标集，由(9)式：

$$b = \frac{1}{y_s} - \sum_{i\in S}\boldsymbol{\alpha}_i^\star y_i\mathbf{x}_i^\top\mathbf{x}_s \tag{13}$$

对(13)式左右两边同时左乘$\boldsymbol{\alpha}_s^\star y_s$然后对$s$求和:

$$\begin{aligned}\sum_{s\in S}\boldsymbol{\alpha}_s^\star y_s b &= \sum_{s\in S}\boldsymbol{\alpha}_s^\star y_s\frac{1}{y_s} - \sum_{s\in S}\boldsymbol{\alpha}_s^\star y_s\sum_{i\in S}\boldsymbol{\alpha}_i^\star y_i\mathbf{x}_i^\top\mathbf{x}_s \\ &= \sum_{s\in S}\boldsymbol{\alpha}_s^\star - \sum_{s\in S}\sum_{i\in S}\boldsymbol{\alpha}_s^\star\boldsymbol{\alpha}_i^\star y_s y_i\mathbf{x}_i^\top\mathbf{x}_s \end{aligned} \tag{14}$$

又对偶问题$\mathcal{P}^4$有约束条件$\sum_{i=1}^m \alpha_i y_i = 0$，由于只有支持向量对应的$\alpha_i > 0$，否则$\alpha_i = 0$，所以$\sum_{s\in S}\boldsymbol{\alpha}_s^\star y_s = 0$，(14)式等价于：

$$\begin{aligned} 0 &= \sum_{s\in S}\boldsymbol{\alpha}_s^\star - \sum_{s\in S}\sum_{i\in S}\boldsymbol{\alpha}_s^\star\boldsymbol{\alpha}_i^\star y_s y_i\mathbf{x}_i^\top\mathbf{x}_s \\ \sum_{s\in S}\boldsymbol{\alpha}_s^\star &= \sum_{s\in S}\sum_{i\in S}\boldsymbol{\alpha}_s^\star\boldsymbol{\alpha}_i^\star y_s y_i\mathbf{x}_i^\top\mathbf{x}_s \end{aligned} \tag{15}$$

联立(12)式与(15)式，可得：

$$\|\mathbf{w}^\star\|^2 = \sum_{s\in S}\boldsymbol{\alpha}_s^\star = \sum_{i=1}^{m}\boldsymbol{\alpha}_i^\star \tag{16}$$

(12)

分类器的间隔由$\frac{1}{\|\mathbf{w}\|}$表达，即可看作只与$\|\mathbf{w}\|$有关。而由上一小问的结果，$\|\mathbf{w}^\star\|^2$可以只用$\{\boldsymbol{\alpha}_i^\star\}_{i=1}^m$表达，故可看作$\|\mathbf{w}^\star\|^2$只与$\{\boldsymbol{\alpha}_i^\star\}_{i=1}^m$有关，即$\|\mathbf{w}^\star\|$只与$\{\boldsymbol{\alpha}_i^\star\}_{i=1}^m$有关。故最大间隔$\frac{1}{\|\mathbf{w}^\star\|}$只与$\{\boldsymbol{\alpha}_i^\star\}_{i=1}^m$有关，而对于$\{\boldsymbol{\alpha}_i^\star\}_{i=1}^m$，只有当$(\mathbf{x}_i, y_i)$为支持向量时才有$\boldsymbol{\alpha}_i^\star > 0$，否则$\boldsymbol{\alpha}_i^\star = 0$。所以分类器的最大间隔只与支持向量有关，即支持向量机的最终模型只与支持向量有关。且$\sum_{i=1}^m \boldsymbol{\alpha}_i^\star$越小，最大间隔越大。

3 参考文献

(1) 《机器学习》周志华著

(2) 《模式识别》吴建鑫著

(3) 《Convex Optimization》by Stephen Boyd and Lieven Vandenberghe

(4) 《The Elements of Statistical Learning》 by Trevor Hastie, Robert Tibshirani and Jerome Friedman

(5) 《Foundations of Machine Learning》 by Mehryar Mohri, Afshin Rostamizadeh and Ameet Talwalkar