机器学习导论
习题三

学号, 作者姓名, 邮箱

2021 年 4 月 25 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无序的。

(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. (1) 第一，二分划分会以缓慢的速度划分节点，而多路划分会在某个节点将数据按照某个属性完全分开。多路划分往往会使得下一层的数据量迅速减少，增加过拟合的风险；第二，多次二分划分可以模拟一次多路划分，所以二分划分更加灵活，该策略允许我们在后续需要的时候进一步使用某属性进行划分。

(2) 所有子集的个数为2^K，除去全集、空集，并考虑对称性，所有合法的划分数为$2^{K-1} - 1$。

(3) 如果使用Gini Index作为评价节点混淆度（impurity）的指标，那么这种策略可以最大程度地减少平均混淆度。具体证明可以参考“第一题第三问.pdf”。

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$h \in \mathcal{H} = \{\mathbf{x} \to \text{sign}(\mathbf{w}^\top \mathbf{x} + b) : \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}\}$。

(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. (1) $y(\mathbf{w}^\top\mathbf{x}+b)\geq 0$

(2) $\rho_h(\mathbf{x})=\frac{|\mathbf{w}^\top\mathbf{x}+b|}{\|\mathbf{w}\|}$

(3) 优化问题$\mathcal{P}^1$如下：

$$\begin{aligned}\max_{\mathbf{w},b}\ \min_{i\in[m]}\quad & \frac{|\mathbf{w}^\top\mathbf{x}_i+b|}{\|\mathbf{w}\|}\\ s.t.\quad & y_i(\mathbf{w}^\top\mathbf{x}_i+b)\geq 0,\forall i\in[m]\end{aligned}\tag{1}$$

(4) 若$(\mathbf{w},b)$是问题$\mathcal{P}^1$的一组最优解，那么对于任意的正数$s\in\mathbb{R}_{++}$，容易验证$(s\mathbf{w},sb)$也是$\mathcal{P}^1$的一组最优解。

(5) 问题$\mathcal{P}^2$如下：

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

(6) 问题$\mathcal{P}^3$如下：

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

(7) 问题$\mathcal{P}^4$如下：

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j \\ s.t. \quad & \alpha_i \geq 0, \forall i \in [m] \\ & \sum_{i=1}^{m} \alpha_i y_i = 0, \forall i \in [m] \end{aligned} \tag{4}$$

(8) 首先，一个标准的优化问题（最小化问题）包含目标函数f_0、不等式约束$f_i \leq 0$和等式约束$h_i = 0$。如果f_0是凸函数、所有不等式约束f_i都是凸函数、所有等式约束h_i都是仿射函数，则该问题为凸优化问题。对照定义，容易证明$\mathcal{P}^3$和$\mathcal{P}^4$都是凸优化问题。

(9) 第一，在该问题中获得的对偶问题满足强对偶性，可以获得原问题的最优值；第二，对偶可以使优化问题的求解复杂度不依赖于样本维度，而依赖于样本数；第三，便于引入核技巧。

(10) $\mathbf{w}^\star = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$；对于任意一个支持向量$\mathbf{x}_i$，有$b = y_i - \sum_{j=1}^{m} \alpha_j y_j \mathbf{x}_j^\top \mathbf{x}_i$。

(11) $\|\mathbf{w}^\star\|^2 = \|\boldsymbol{\alpha}^\star\|_1$

(12) 最大化间隔相当于最小化$\boldsymbol{\alpha}$的ℓ_1范数，可以解释支持向量的稀疏性。