

# 机器学习导论

## 习题三

191300020, 黄彦骁, AdrianHuang@smail.nju.edu.cn

2021 年 4 月 23 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。<sup>1</sup>

- (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**。

<sup>1</sup>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。
- (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  $A = \{a_1, \dots, a_k, \dots, a_K\}$  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  $p_k = \text{Prob}[y = +1|A = a_k]$ ，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  $p_1 \leq \dots \leq p_k \leq \dots \leq p_K$ 。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  $\{a_1, \dots, a_{\bar{k}}\}$  和  $\{a_{\bar{k}+1}, \dots, a_K\}$ ，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) 二分划分策略可以有效的减少对于特征不必要的划分，减少建立决策树的开销以及判定样本的开销，同时这样的划分方法也具有更好的泛化性能，更有可能得到一颗有着更高正确率的决策树。

- (2) 对其进行划分得到的方案数  $\Omega$  为：

$$\Omega = \frac{C_K^1 + C_K^2 + \dots + C_K^{K-1}}{2} = 2^{K-1} - 1$$

- (3) 假设我们找到的最优的  $\bar{k}$  存在如下性质，即：

$$p(\bar{k}) \leq 0.5, p(\bar{k} + 1) > 0.5$$

那么划分出来的两个子集一个概率均小于 0.5，一个概率均大于 0.5，那么在此概率下便可将一个集合划分为正类，另一个划分为负类，这样情况下两个集合的 *Gini\_index* 都会相对其他划分更高，如果直接采用找到最小 *Gini\_index* 来进行集合划分，也会倾向于找到这样的划分方式（划分的两集合大小一致的话）所以这样的划分与使用 *Gini\_index* 遍历所有情况进行划分是一致的，具有合理性。

## 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  $h: \mathbb{X} \rightarrow \mathbb{Y}$  表示某个线性分类器，即  $h \in \mathcal{H} = \{\mathbf{x} \rightarrow \text{sign}(\mathbf{w}^\top \mathbf{x} + b) : \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}\}$ 。

- (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  $\rho_h = \min_{i \in [m]} \rho_h(\mathbf{x}_i)$ 。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  $\min_{i \in [m]} |\mathbf{w}^\top \mathbf{x}_i + b| = 1$ （或者等价地， $\min_{i \in [m]} y_i(\mathbf{x}_i^\top \mathbf{x}_i + b) = 1$ ），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  $\{\alpha_i^*\}_{i=1}^m$  是对偶问题  $\mathcal{P}^4$  的最优解，设  $(\mathbf{w}^*, b^*)$  是原问题  $\mathcal{P}^3$  的最优解。请使用  $\{\alpha_i^*\}_{i=1}^m$  和  $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$  表达  $\mathbf{w}^*$  和  $b^*$ ，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  $\{\alpha_i^*\}_{i=1}^m$  简洁地表达  $\|\mathbf{w}^*\|^2$ 。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) 得到不等式为：

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

(2)

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

(3)  $\mathcal{P}^1$  为:

$$\begin{aligned} \max_{\mathbf{w}, b} \quad & \frac{|\mathbf{w}^\top \mathbf{x} + b|}{\|\mathbf{w}\|} \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) > 0 \quad i = 1, 2, \dots, m \end{aligned}$$

(4) 在优化问题  $\mathcal{P}^1$  中, 对于求到的  $\frac{|\mathbf{w}^\top \mathbf{x} + b|}{\|\mathbf{w}\|}$  的最大值对应的解为  $\mathbf{w}^*, b^*$ , 去任意不等于 0 的实数  $\lambda$ , 如果取值为  $\lambda \mathbf{w}^*, \lambda b^*$ , 那么需要优化的表达式取值不变, 所以解也可以为  $\lambda \mathbf{w}^*, \lambda b^*$ , 故该优化问题可以有无穷多组解。

(5)  $\mathcal{P}^2$  为:

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

(6)  $\mathcal{P}^3$  为:

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

(7) lagrange 函数为:

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

其中  $\boldsymbol{\alpha}$  为拉格朗日乘子满足  $\alpha_i \geq 0$ , 令函数对  $\mathbf{w}, b$  求偏导等于 0 有:

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

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

将上述两式代入得到原问题的对偶问题为:

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

(8) 凸优化问题定义: 在最小化(最大化)目标函数的情况下, 目标函数是凸函数(凹函数), 不等式约束函数是凸函数, 等式约束函数是仿射的。

对于  $\mathcal{P}^3$  而言, 最小化的目标函数为  $\frac{\|\mathbf{w}\|^2}{2}$  为二次函数是凸函数, 不等式约束函数为  $1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)$  为线性函数, 也为凸函数, 故  $\mathcal{P}^3$  为凸优化问题。

再考虑  $\mathcal{P}^4$ , 最大化的目标函数为  $\sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j$ , 设为  $G$ , 对其求二阶导可以得到  $\nabla^2 G = -\mathbf{A}$ , 其中  $\mathbf{A} = (y_i \mathbf{x}_i * y_j \mathbf{x}_j)_{[i, j]}$ , 所以我们得到  $\nabla^2 G$  为半负定的, 所以有  $G$  为凹函数, 而不等式约束和等式约束函数均为线性仿射的, 所以  $\mathcal{P}^4$  为凸优化问题。

(9) 使用对偶将原来的双变量优化问题转化为单变量优化问题, 减少了计算复杂度, 降低求解问题的时间开销。

(10) 由 *lagrange* 函数的偏导数得:

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

再求该问题得 *KKT* 条件得:

$$\begin{cases} \alpha_i \geq 0 \\ y_i f(\mathbf{x}_i) - 1 \geq 0 \\ \alpha_i (y_i f(\mathbf{x}_i) - 1) \geq 0 \end{cases}$$

则  $\alpha_i = 0$  和  $y_i f(\mathbf{x}_i) = 1$  必有一项成立, 当  $\alpha_i = 0$  时, 对应的情况对于  $f(\mathbf{x})$  没有影响当  $\alpha_i > 0$  时, 有  $y_i f(\mathbf{x}_i) = 1$ , 这表示该项为一个支持向量, 我们从支持向量来推导  $b^*$ 。对任意的支持向量  $\mathbf{x}_i$ , 有:

$$\mathbf{w}_i^\top \mathbf{x}_i + b = y_i$$

所以我们得到:

$$b^* = y_i - \sum_{i=1}^m \alpha_i^* y_i (\mathbf{x}_i * \mathbf{x}_j)$$

更鲁棒一点的结果我们对所有的支持向量取平均得到  $b^*$  值

(11) 由于结果  $b^* = y_i - \sum_{i=1}^m \alpha_i y_i (\mathbf{x}_i * \mathbf{x}_j)$  对所有  $\alpha_i^* \neq 0$  都成立, 我们将等式两边乘以  $\alpha_i^* y_i$  并累加有:

$$\sum_{i=1}^m \alpha_i^* y_i b = \sum_{i=1}^m \alpha_i^* y_i^2 - \sum_{i,j=1}^m \alpha_i^* \alpha_j^* y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

由于  $y_i^2 = 1$ , 得到结果为:

$$\sum_{i=1}^m \alpha_i^* - \|\mathbf{w}\|^2 = 0$$

所以  $\|\mathbf{w}\|^2$  表示为  $\sum_{i=1}^m \alpha_i^*$

(12) 分类器的间隔表示为  $\frac{1}{\|\mathbf{w}\|}$ , 由上一问可表示为  $\frac{1}{\|\alpha\|_1}$ , 说明原问题最终只与 *lagrange* 乘子有关, 所以对偶问题的最优解也对应着原问题的最优解, 两问题本质上一致。