机器学习导论
习题三

学号, 姓名, 邮箱

2020 年 5 月 3 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LaTeX模板中第一页填写个人的学号、姓名、邮箱；

(2) 本次作业需提交该pdf文件、问题4可直接运行的源码(.py文件)、问题4的预测结果(.csv文件)，将以上三个文件压缩成zip文件后上传。注意：pdf、预测结果命名为“学号_姓名”(例如“181221001_张三.pdf”)，源码、压缩文件命名为“学号”，例如“181221001.zip”；

(3) 未按照要求提交作业，提交作业格式不正确，**作业命名不规范**，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23日23:59:59。除非有特殊情况（如因病缓交），否则截止时间后不接收作业，本次作业记零分。

[1] 参考尹一通老师高级算法课程中对学术诚信的说明。

1 [20pts] Decision Tree I

(1) **[5pts]** 试分析使用“最小训练误差”作为决策树划分选择的缺陷。

(2) **[5pts]** 树也是一种线性模型，考虑图(1)所示回归决策树，X_1, X_2均在单位区间上取值，t_1, t_2, t_3, t_4满足$0 < t_1 < t_3 < 1, 0 < t_2, t_4 < 1$，试绘制出该决策树对于特征空间的划分。假设区域$R_i$上模型的输出值为$c_i$，试用线性模型表示该决策树。

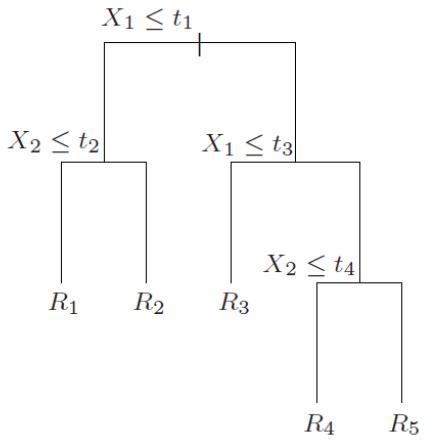

图 1: 回归决策树

(3) **[10pts]** 对于回归树，我们常采用平方误差来表示回归树对于训练数据的预测误差。但是找出平方误差最小化准则下的最优回归树在计算上一般是不可行的，通常，我们采用贪心的算法计算切分变量j和分离点s。CART回归树在每一步求解如下优化问题

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right]$$

其中$R_1(j,s)=\{\boldsymbol{x}|x_j\le s\}, R_2(j,s)=\{\boldsymbol{x}|x_j>s\}$。试分析该优化问题表达的含义并给出变量$j, s$的求解思路。

Solution. 此处用于写解答(中英文均可)

(1) 更容易导致决策树过拟合；且对于节点概率变化的敏感程度不如熵和基尼值。

(2) 特征空间划分如图(2)所示：

该决策树可以表示为

$$\text{DecisionTree}(x)=\sum_{i=1}^{5}c_i\mathbb{I}(x\in R_i)$$

(3) 该优化问题求解参数对(j,s)使得在切分后的两个子区域上决策树平方误差的总和最小。参数对的求解通过穷举进行。

设输入一共包含n个实例，每个实例包含m维特征，第i维特征上的取值为$x_{ij_1}\le x_{ij_2}\le\cdots\le x_{ij_n}$，则$j$遍历$1,2,\cdots,m$，当$j=i(i=1,2,\cdots,m)$时，$s$遍历$x_{ij_1}, x_{ij_2},\cdots, x_{ij_{n-1}}$，选

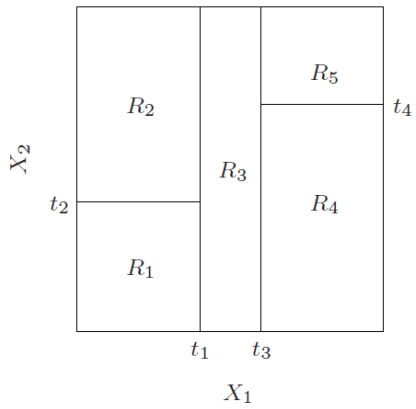

图 2: 决策树对特征空间的划分

择使得$f(j,s) = \min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2$最小的参数对$(j,s)$即可。对于函数$f(j,s)$，$c_1, c_2$的最小值点分别为$c_1 = \text{avg}(y_i | x_i \in R_1(j,s)), c_2 = \text{avg}(y_i | x_i \in R_2(j,s))$。

Feedback.

1. 对于第一问，使用“最小训练误差”作为决策树划分选择的缺陷主要是容易导致过拟合。“对于节点概率变化的敏感程度不如熵和基尼值”可以通过简单的举例体会。此问答出“过拟合”即给满分。
2. 对于第二问，通过**示性函数**可以将决策树表达成线性模型。
3. 对于第三问，s遍历的时候可以取$x_{ij_1}, x_{ij_2}, \cdots, x_{ij_{n-1}}$，也可以取$\frac{x_{ij_1}+x_{ij_2}}{2}, \frac{x_{ij_2}+x_{ij_2}}{3}, \cdots, \frac{x_{ij_{n-1}}+x_{ij_n}}{2}$。其他的取法也可以，言之有理即可。

2 [25pts] Decision Tree II

(1) **[5pts]** 对于不含冲突数据（即特征向量相同但标记不同）的训练集，必存在与训练集一致（即训练误差为0）的决策树。如果训练集可以包含无穷多个数据，是否一定存在与训练集一致的深度有限的决策树？证明你的结论。（仅考虑单个划分准则仅包含一次属性判断的决策树）

(2) **[5pts]** 考虑如表1所示的人造数据，其中“性别”、“喜欢ML作业”是属性，“ML成绩高”是标签。请画出使用信息增益为划分准则的决策树算法所有可能的结果。（需说明详细计算过程）

(3) **[10pts]** 考虑如表2所示的验证集，对上一小问的结果基于该验证集进行预剪枝、后剪枝，剪枝结果是什么？（需给出详细计算过程）

(4) **[5pts]** 比较预剪枝、后剪枝的结果，每种剪枝方法在训练集、验证集上的准确率分别为多少？哪种方法拟合能力较强？

表 1: 训练集

编号	性别	喜欢ML作业	ML成绩高
1	男	是	是
2	女	是	是
3	男	否	否
4	男	否	否
5	女	否	是

表 2: 验证集

编号	性别	喜欢ML作业	ML成绩高
6	男	是	是
7	女	是	否
8	男	否	否
9	女	否	否

Solution. 此处用于写解答(中英文均可)

(1) 不一定。考虑数据集$\{(1/i,(-1)^i)\}_{i=1}^{\infty}$。有限深度的决策树有有限次属性判断，只能将区间划分为有限个区域，而该数据集将区间划分为无穷个区域。

(2) 结果如图3所示。

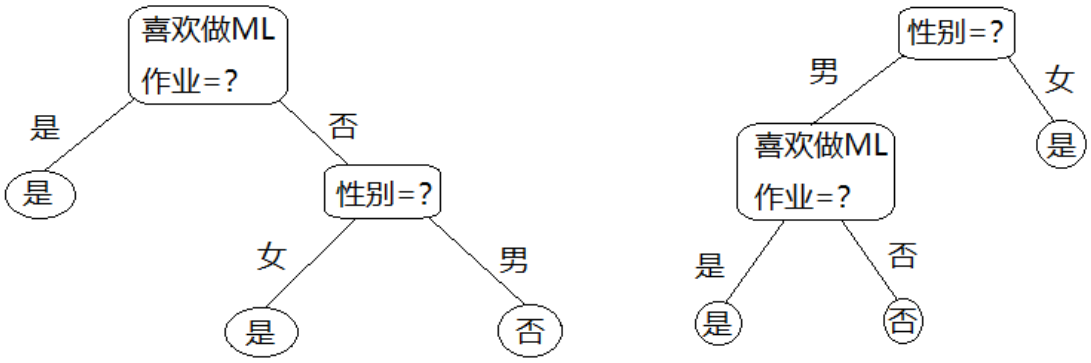

图 3: 决策树结果

(3) 预剪枝结果如图4所示。 后剪枝结果如图5所示。

(4) 预剪枝分别为7/9与4/9，后剪枝分别为7/9与7/9。后剪枝更强。

Feedback.

2.1

1. 题目问的是“是否一定存在”，负面的回答应该是“不一定存在”，而不是“一定不存在”，后者扣1分。

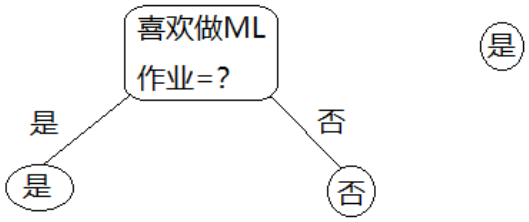

图 4: 预剪枝结果

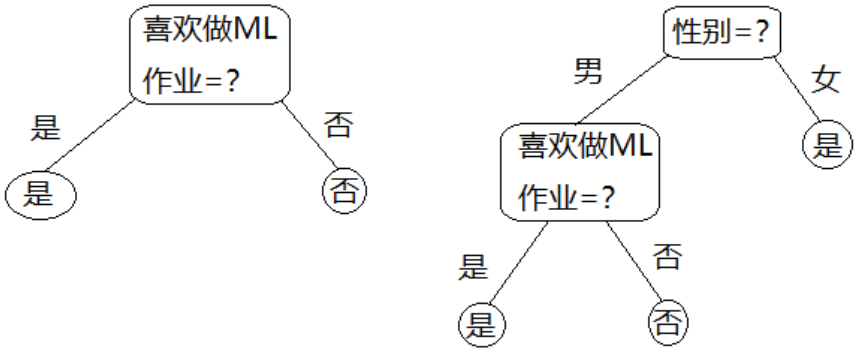

图 5: 后剪枝结果

2. 构造含有无穷多个数据的数据集时，要注意有限和无限的本质区别，可能无法定义“第x大”、“相邻”的概念。有限时，可以把所有数据按某特征排序，并说第x大的数据；无限时，排序可以排，但可能无法说第x大的数据（考虑全体整数的排序）。无限也可能无法定义相邻，考虑0和所有正整数倒数组成的集合，无法定义0和谁相邻。该错误扣1分。
3. “无穷个数据”与“无穷个属性”不是一个含义，只讨论后者的不得分（扣5分）。
4. 注意数据集与树的顺序问题。应该是先有数据集，然后看是否有可以分类正确的树；而不是有了一棵树，构造一个数据集说明这棵树无法完美分类。犯该错误不得分（扣5分）。
5. “离散”不代表“有限”，酌情扣分。
6. 证明过程中不应出现模糊的概念，如“随机的”、“没有规律的”、“没有数学关联的”；应使用具有明确定义的语言论证，酌情扣分。

2.2

1. 题目要求画出树，两棵树每棵2分。
2. 题目要求画出树，两棵树每棵2分。

2.3

1. 两棵树的剪枝均要讨论，只考虑一种扣5分。

2.4

1. 两棵树的结果都要给出，只考虑一种扣2分。
2. 结论错误，或无法通过数据支撑的，扣1分。

3 [25pts] SVM with Weighted Penalty

考虑标准的SVM优化问题如下(即课本公式(6.35))，

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, 2, \cdots, m. \end{aligned} \tag{1}$$

注意到，在(1)中，对于正例和负例，其在目标函数中分类错误的“惩罚”是相同的。在实际场景中，很多时候正例和负例分错或分对但置信度较低的“惩罚”是不同的，比如癌症诊断等。

现在，我们希望对负例分类错误(即false positive)或分对但置信度较低的样本施加$k > 0$倍于正例中被分错的或者分对但置信度较低的样本的“惩罚”。对于此类场景下，

(1) **[10pts]** 请给出相应的SVM优化问题。

(2) **[15pts]** 请给出相应的对偶问题及KKT条件，要求详细的推导步骤。

Solution. 此处用于写解答(中英文均可)

(1) 我们设所有正例的下标构成的集合为$\mathcal{P}$，所有负例的下标构成的集合为$\mathcal{N}$。则SVM优化问题可以写成如下形式：

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\left(\sum_{i\in\mathcal{P}}\xi_i + k\sum_{i\in\mathcal{N}}\xi_i\right) \\ \text{s.t.} \quad & y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, 2, \cdots, m. \end{aligned} \tag{2}$$

(2) 拉格朗日函数(Lagrangian)为

$$L(\mathbf{w}, b, \xi_i, \alpha_i, \beta_i) = \frac{1}{2}\|\mathbf{w}\|^2 + C\left(\sum_{i\in\mathcal{P}}\xi_i + k\sum_{i\in\mathcal{N}}\xi_i\right) - \sum_{i=1}^{m}\alpha_i(y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - 1 + \xi_i) - \sum_{i=1}^{m}\beta_i\xi_i$$

易知拉格朗日函数是凸函数，因此可以通过一阶条件求解最小值点，即求解相应的偏导并设为0

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i = 0$$

$$\frac{\partial L}{\partial \mathbf{w}} = \sum_{i=1}^{m}\alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i,\ i \in \mathcal{P}$$

$$\frac{\partial L}{\partial \xi_i} = kC - \alpha_i - \beta_i,\ i \in \mathcal{N}$$

因此，拉格朗日对偶函数(Lagrange dual function)为

$$\begin{aligned} g(\alpha_i, \beta_i) &= \inf_{\mathbf{x},b,\xi_i} L(\mathbf{w}, b, \xi_i, \alpha_i, \beta_i) \\ &= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j^T + \sum_{i=1}^{m}\alpha_i \end{aligned}$$

其中，$\alpha_i \geq 0, \beta_i \geq 0,\ i = 1, 2, \cdots, m$

拉格朗日对偶问题(Lagrange dual problem)即是最大化拉格朗日对偶函数：

$$\begin{aligned} \min_{\alpha_i, \beta_i} \quad & \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j - \sum_{i=1}^{m}\alpha_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i \\ & 0 \leq \alpha_i \leq C \quad i \in \mathcal{P} \\ & 0 \leq \alpha_i \leq kC \quad i \in \mathcal{N} \end{aligned} \tag{3}$$

容易验证原始的凸优化问题满足Slatter条件(Slatter condition)，因此，强对偶性成立，原始问题的最优解$(\mathbf{w}^*, b^*, \xi_i^*)$和对偶问题的最优解$(\alpha_i^*, \beta_i^*)$满足KKT条件(KKT condition)：

$$y_i({\mathbf{w}^*}^{\mathrm{T}}\mathbf{x}_i + b^*) \geq 1 - \xi_i^*,\ \xi_i^* \geq 0, i = 1, 2, \cdots, m.$$

$$\alpha_i^* \geq 0,\ \beta_i^* \geq 0, i = 1, 2, \cdots, m.$$

$$\alpha_i(y_i({\mathbf{w}^*}^{\mathrm{T}}\mathbf{x}_i + b^*) - 1 + \xi_i^*) = 0, i = 1, 2, \cdots, m.$$

$$\beta_i^*\xi_i^* = 0, i = 1, 2, \cdots, m.$$

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

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

$$\alpha_i^* + \beta_i^* = C, i \in \mathcal{P}$$

$$\alpha_i^* + \beta_i^* = kC, i \in \mathcal{N}$$

Feedback.

1. 分错或者分对但置信度较低的惩罚其实就是$y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i$中的$\xi_i$，因此该问题按照样本的实际标记对$\xi_i$进行不同的处理即可。关于$k$倍的惩罚有多种写法，言之有理即可。
2. 有的同学原始问题写对了，但是对偶问题求解错了。拉格朗日对偶函数求导和化简相对麻烦，推导的时候请细心细心再细心!

4 [30 pts] 编程题, Linear SVM

请结合编程题指南进行理解

SVM转化成的对偶问题实际是一个二次规划问题，除了SMO算法外，传统二次规划方法也可以用于求解对偶问题。求得最优拉格朗日乘子后，超平面参数$\mathbf{w}, \mathbf{b}$可由以下式子得到：

$$\mathbf{w} = \sum_{i=1}^{m}\alpha_i y_i x_i \tag{4}$$

$$\mathbf{b} = \frac{1}{|S|}\sum_{s \in S}(y_s - \sum_{i \in S}\alpha_i y_i x_i^T x_s) \tag{5}$$

请完成以下任务：

(1) **[5pts]** 使用QP方法求解训练集上的SVM分类对偶问题(不考虑软间隔情况)。

(2) **[10 pts]** 手动实现SMO算法求解上述对偶问题。

(3) **[15 pts]** 对测试数据进行预测，确保预测结果尽可能准确。

Solution. 此处用于写解答(中英文均可)

Feedback.

1. 这组数据在原问题上做硬间隔对偶问题是一定会得到不满足KKT条件的对偶问题，此时用QP优化包求解一定会报错，之所以把这个问题放在第一问就是想提醒同学此问题需要用软间隔，此时只需要如实回答硬间隔报错，然后在后续问题中应用软间隔即可，但是很多同学的重点在于如何调整得到一个结果（注意加对角小量和软间隔原理是类似的）。
2. 在朴素贪心框架下就能实现，大家做的都还可以，唯一缺少的一步是在求解之前需要验证对偶问题是否满足KKT条件，否则SMO解出来的解很可能错得离谱。
3. 大部分预测正确率在80%左右，应用高斯核+软间隔的同学正确率在90%以上。