机器学习导论
习题三

181220010, 丁豪, 181220010@smail.nju.edu.cn

2020 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。[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 23:55:00**。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 (**??**) 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。

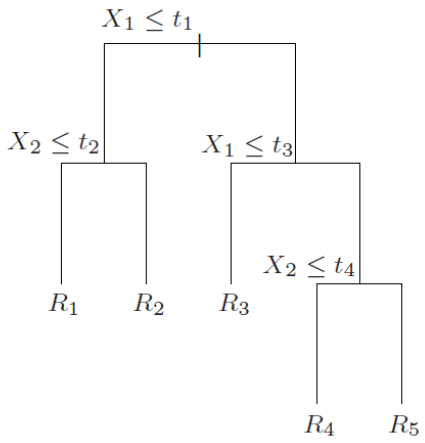


图 1: 回归决策树

(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

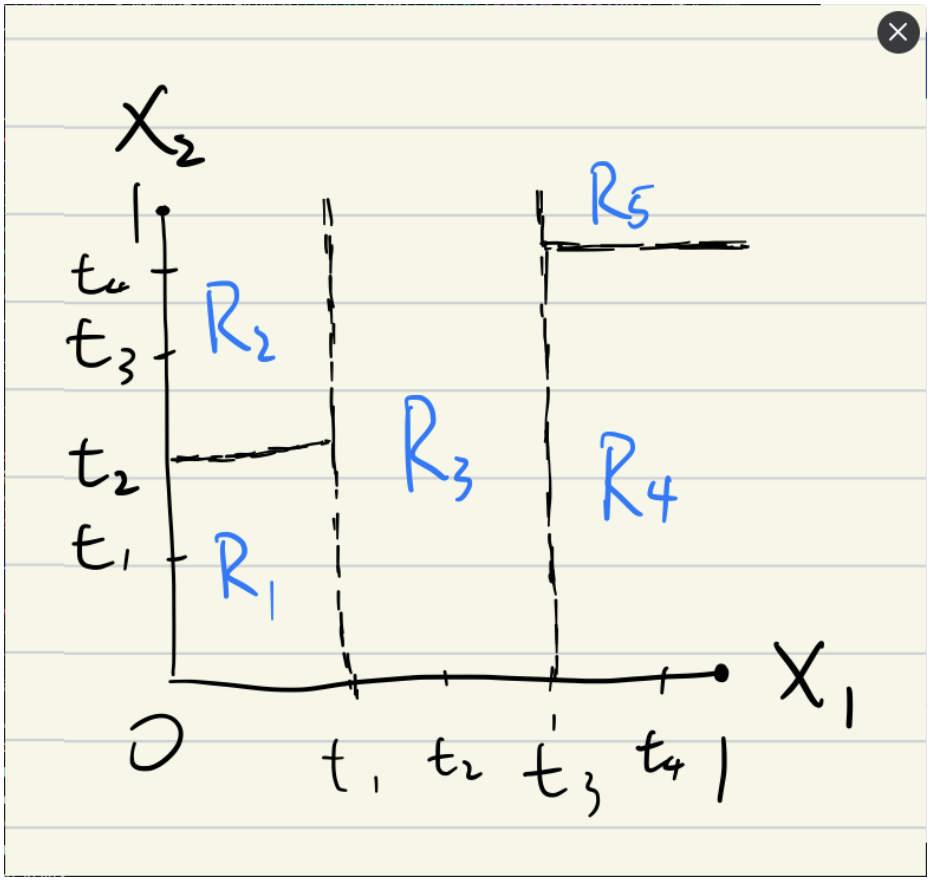
$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right]$$

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.

(1) 将训练集同时当做评价指标，会使得划分时有可能过分考虑所选训练集内部特殊性的影响，将会导致过拟合

(2) 如图



用线性模型来表示，使用 4 个分类器来对空间进行划分：

$$\begin{cases} f(x_1,x_2)_{45,123} = x_1 - t_3 \\ f(x_1,x_2)_{345,12} = x_1 - t_1 \\ f(x_1,x_2)_{2,1} = x_2 - t_2 \\ f(x_1,x_2)_{5,4} = x_2 - t_4 \end{cases}$$

*. 参考机器学习导论课程群中助教对部分同学针对此题讨论的解答，这里给出另外一种依靠示性函数的线性模型分类方案

$$\begin{aligned} f(X) = &\sum_{i=1}^{5} i\mathbb{I}(X \in R_i) \\ =&\mathbb{I}(X_1 \le t_1, X_2 \le t_2) + 2\mathbb{I}(X_1 \le t_1, X_2 > t_2) \\ &+ 3\mathbb{I}(t_1 < X_1 \le t_3) + 4\mathbb{I}(X_1 > t_3, X_2 \le X_4) \\ &+ 5\mathbb{I}(X_1 > t_3, X_2 > t_4) \end{aligned}$$

此时我们有$f(X) = i \iff X \in R_i$

(3) 表达的含义即求解在切分变量 j 和分离点 s 下，训练集上 x 关于变量 j 的“正类”、“负类”其对应 y 的类内方差之和最小的 j,s。也就是寻找某个变量 j 和划分点 s，使得这样划分后，两类的聚类程度最高。

求解 j,s 的方法可以采用双重遍历，j 自然是可以遍历的，s 虽然是离散值，但是当其取值

在 $[x_k, x_{k+1}), x_1 \leq x_2 \leq x_3 ...$ 时，他的作用是完全相同的，因此 s 可以离散化，于是就可以使用双重遍历找到最佳 j 与 s 的组合。

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**??**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）

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

(3) **[10pts]** 考虑如表**??**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）

表 2: 验证集

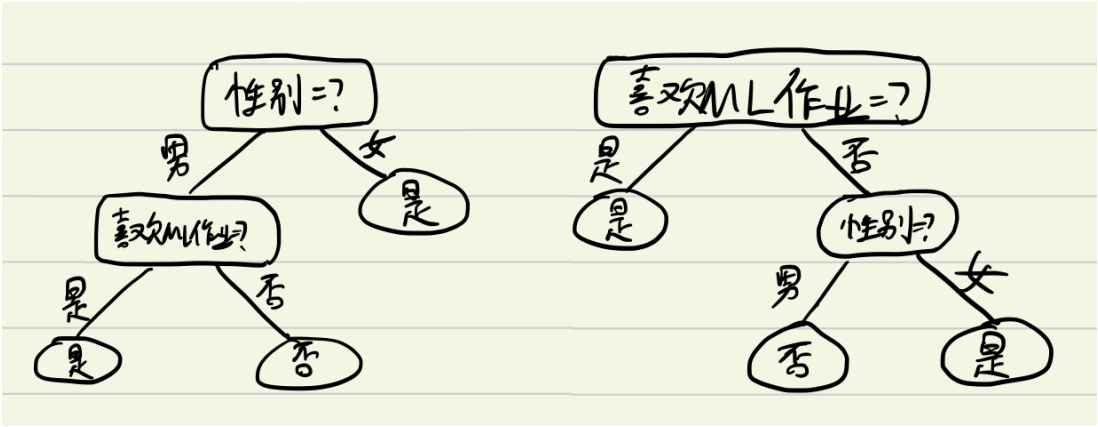
编号	性别	喜欢 ML 作业	ML 成绩高
6	男	是	是
7	女	是	否
8	男	否	否
9	女	否	否

(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？

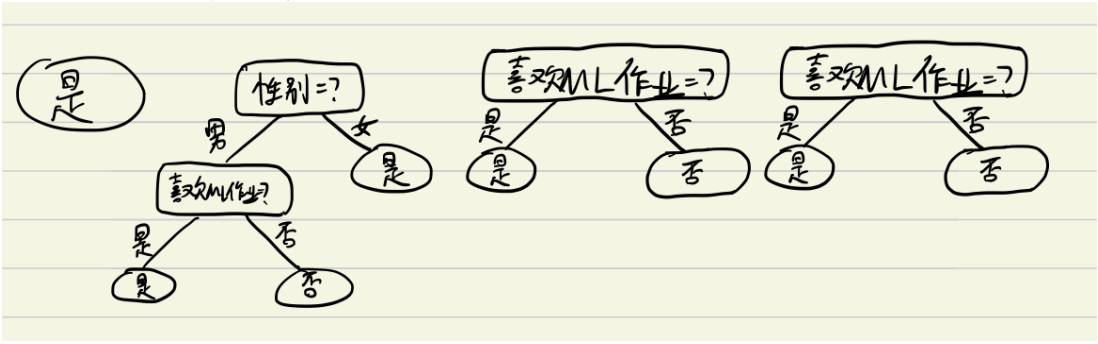
Solution.

(1) 假设数据属性值数量有限为 n，因为不含冲突数据，所以即使有无穷多个数据，其中不重复的至多有 2^n 个。根据决策树的训练方法可知，每一个不重复的训练数据在完全展开的决策树上都对应到一个叶子节点，因此至多有 2^n 个叶子节点。对于有限叶子节点的树，其深度自然是有限的。但如果数据的属性值也是无限的，则无法产生有限深度决策树。

(2) 若不进行划分时 ML 成绩高占多数，因此单独一个节点为 ML 成绩高，此时信息熵
$Ent(D) = -(\frac{3}{5}lg\frac{3}{5} + \frac{2}{5}lg\frac{2}{5}) = 0.97$
此时若对性别进行划分，男:1,3,4: 否, 女:2,5: 是，信息增益为
$Gain(D, sex) = Ent(D) - \frac{3}{5}(-\frac{1}{3}lg\frac{1}{3} - \frac{2}{3}lg\frac{2}{3}) - \frac{2}{5}(-1lg1) = 0.42$
此时若对喜欢 ML 作业进行划分，是:1,2: 是，否:3,4,5: 否, 信息增益为
$Gain(D, love) = Ent(D) - \frac{2}{5} * 0 - \frac{3}{5}(-\frac{1}{3}lg\frac{1}{3} - \frac{2}{3}lg\frac{2}{3}) = 0.42$
两者信息增益相同，因此任意选择某一个作为第一划分属性，另一个作为下一个划分属性，所以共有如图两种决策树。



(3) 左图，预剪枝：首先对于根节点，若不划分则“是”为多数，因此将所有判定为是，正确率 $\frac{1}{4}$，若展开，则男“否”女“是”，正确率 $\frac{1}{4}$, 正确率不提高，因此不展开。
左图，后剪枝：对于最底部两个节点，若不剪枝则正确率为 $\frac{1}{2}$, 若剪枝则正确率为 $\frac{1}{4}$, 因此不剪枝。
右图，预剪枝：对根节点如果不展开，正确率 $\frac{1}{4}$, 展开正确率为 $\frac{3}{4}$，因此此处展开。再看第二个划分点，若不划分则正确率 $\frac{3}{4}$，若划分则正确率 $\frac{1}{4}$，因此不展开。
右图，后剪枝：对底下两个节点，由上述分析可知剪枝导致正确率上升，因而剪枝，而根节点剪枝导致正确率下降，因而不剪枝。



(4) 对左图，预剪枝在训练集与测试集上准确率分别为 $\frac{3}{5}, \frac{1}{4}$, 后剪枝为 $1, \frac{1}{2}$
对右图，预剪枝后剪枝均为 $\frac{4}{5}, \frac{3}{4}$
经过对比可以发现后剪枝方法在训练集上拟合能力较强。

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(**??**)中, 对于正例和负例, 其在目标函数中分类错误或分对但置信度较低的 “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” 往往是不同的, 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.

(1) 相应的 SVM 优化问题为:

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

(2) 对偶问题及 KKT 条件推导如下

$$L(w,b,\xi,\alpha,\mu)=\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{m}k^{\frac{1-y_i}{2}}\xi_i+\sum_{i=1}^{m}\alpha_i(1-\xi_i-y_i(w^Tx_i+b))-\sum_{i=1}^{m}\mu_i\xi_i$$

其中$\alpha_i\geq 0,\mu_i\geq 0$是拉格朗日乘子

令$L(w,b,\xi,\alpha,\mu)$对w,b,ξ_i偏导为零可得

$$w=\sum_{i=1}^{m}\alpha_i y_i x_i \quad \sum_{i=1}^{m}\alpha_i y_i=0 \quad Ck^{\frac{1-y_i}{2}}=\alpha_i+\mu_i$$

将此三式带入可得

$$\begin{aligned}\theta(\alpha)&=\min_{w,b,\xi}L(w,b,\xi,\alpha,\mu)\\&=\frac{1}{2}\|\sum_{i=1}^{m}\alpha_i y_i x_i\|+\sum_{i=1}^{m}\alpha_i-\sum_{i=1}^{m}\alpha_i y_i(\sum_{i=1}^{m}\alpha_i y_i x_i)^T x_i\\&=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\|\sum_{i=1}^{m}\alpha_i y_i x_i\|\\&=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j\end{aligned}$$

所以对偶问题为

$$\begin{aligned}\max_{\alpha}\quad&\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j\\\text{s.t.}\quad&\sum_{i=1}^{m}\alpha_i y_i=0,\\&0\leq\alpha_i\leq Ck^{\frac{1-y_i}{2}},\quad i=1,2,\ldots,m.\end{aligned}$$

KKT 条件为

$$\begin{cases}w=\sum_{i=1}^{m}\alpha_i y_i x_i \quad \sum_{i=1}^{m}\alpha_i y_i=0 \quad Ck^{\frac{1-y_i}{2}}=\alpha_i+\mu_i\\y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i+b)+\xi_i-1\geq 0 \quad \xi_i\geq 0\\\alpha\geq 0 \quad \mu_i\geq 0\\\alpha_i(y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i+b)+\xi_i-1)=0 \quad \mu_i\xi_i=0\end{cases}$$

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：

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

$$\mathbf{b}=\frac{1}{|S|}\sum_{s\in S}(y_s-\sum_{i\in S}\alpha_i y_i x_i^T x_s) \tag{3}$$

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.

(1) 对偶问题为

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ s.t. \quad & \sum_{i=1}^{m} \alpha_i y_i = 0, \\ & \alpha_i \geq 0, \quad i = 1, 2, \ldots, m. \end{aligned}$$

将其化为标准二次规划问题为

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \alpha^T P \alpha + q^T \alpha \\ s.t. \quad & G\alpha \preceq h \\ & A\alpha = b \end{aligned}$$

其中

$$\begin{aligned} P_{ij} &= y_i y_j x_i^T x_j \\ q_i &= -1, \ i = 1, 2, \ldots, m \\ G &= -I \\ h &= 0, \ i = 1, 2, \ldots, m \\ A_i &= y_i, \ i = 1, 2, \ldots, m \\ b &= 0 \end{aligned}$$

使用 cvxopt 包进行求解即得答案

(2) 参考课本内容以及 https://zhuanlan.zhihu.com/p/29212107 上对于 SMO 算法的讲解，手动实现了 SMO 算法优化上述对偶问题，并根据 w,b 的表达式进行最终模型生成

(3) 针对前两问的数据，将 xtrain 拆分出一部分为 xvalid，以对训练的模型进行评估，并选取正确率较高的方法作为最终预测 ytest 的评估