机器学习导论
习题三

181220031, 李惟康, liwk@smail.nju.edu.cn

2020 年 5 月 5 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 [20 pts] Decision Tree I

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

(2) **[5 pts]** 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。

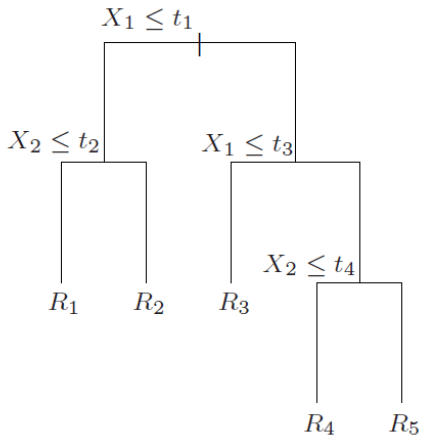


图 1: 回归决策树

(3) **[10 pts]** 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) 以最小训练误差作为决策树划分的依据，由于训练集的情况和真实的测试或验证集之间总是会存在一定偏差，而主要依靠“最小训练误差”划分决策树可能存在过拟合的情况，即只在训练集上的表现较好而泛化能力较差。因此最小训练误差不适合用来作为决策树划分的依据。

(2) 根据上述回归决策树示意图可以得到:

$$f(X_1,X_2)=\begin{cases}c_1 & X_1\le t_1, X_2\le t_2\\ c_2 & X_1\le t_1, X_2> t_2\\ c_3 & X_1> t_1, X_2\le t_3\\ c_4 & X_1> t_1, X_1> t_3, X_2\le t_4\\ c_5 & X_1> t_1, X_1> t_3, X_2> t_4\end{cases}$$

(3) 注意到，回归树的本质在于被划分的输入空间以及相应的输入空间之上的值。举例来说，若选择第 j 个特征以及相应的切分点 s，可划分为：

$$R_1(j,s)=\{\boldsymbol{x}|x_j\le s\}, R_2(j,s)=\{\boldsymbol{x}|x_j>s\}$$

这里采取平方误差 $\sum_{x_i\in R_k}(y_i-f(x_i))^2$ 表示回归树的预测误差，因此需要采取贪心的算法遍历所有的输入变量来确定最优切分变量 j 和最优切分点 s, 即:

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

(采取遍历的方式，我们可以将 j 和 s 找出来：先固定第 j 个特征再选出在该特征下的最佳划分 s；对每一个特征都这样做，那么有 m 个特征，我们就能得到 m 个特征对应的最佳划分，从这 m 个值中取最小值即可得到令全局最优的 (j,s))
求出最优划分 (j,s) 之后继续对两个子区域重复递归的调用上述步骤，直到满足条件，最终将输入空间划分为 M 个区域 $R_1,R_2,\ldots,R_m$ 生成决策树：

$$f(x)=\sum_{m=1}^{M}\hat{c}_m I(x\in R_m)$$

其中，

$$\hat{c}_m=\frac{1}{N_m}\sum_{x_i\in R_m(j,s)}y_i$$
$$x\in R_m, m=1,2$$

2 [25 pts] Decision Tree II

(1) **[5 pts]** 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) **[5 pts]** 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) **[10 pts]** 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) **[5 pts]** 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) 若不存在与训练集一致的决策树，那么训练集生成的决策树至少有一个结点上有多个数据无法划分。这与前提条件矛盾，因此必存在与训练集一致的决策树。

(2) 显然: $|\mathcal{Y}| = 2$, 起始状态下，根节点包含训练集当中的所有样例，其中正例占 $p_1 = \frac{3}{5}$, 负例占 $p_2 = \frac{2}{5}$。因此根节点的信息熵为:

$$\mathrm{Ent}(D) = -\sum_{k=1}^{2} p_k \log_2 p_k = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0.971$$

接下来首先考虑“性别”属性，以该属性对样本集 D 进行划分，可得 D^1(性别 = 男)，D^2(性别 = 女)。其中子集 D^1 中，正例占 $p_1 = \frac{1}{3}$，负例占 $p_2 = \frac{2}{3}$；D^2 中全为正例；因此用“性别”划分后所获得的分支结点的信息熵为:

$$\mathrm{Ent}(D^1) = -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) = 0.918$$
$$\mathrm{Ent}(D^2) = 0$$

因此，属性“性别”的信息增益为:

$$\begin{aligned}\mathrm{Gain}(D, \text{性别}) &= \mathrm{Ent}(D) - \sum_{v=1}^{2}\frac{|D^v|}{|D|}\mathrm{Ent}(D^v) \\ &= 0.971 - (\frac{3}{5}\times 0.918 + \frac{2}{5}\times 0) \\ &= 0.420\end{aligned}$$

类似地，我们可以计算出“喜欢 ML 作业”属性的信息增益:

$$\mathrm{Gain}(D, \text{喜欢 ML 作业}) = 0.420$$

因此，选择“性别”或“喜欢 ML 作业”作为划分属性均可，这里选择“喜欢 ML 作业”为划分属性，下一步对每个分支节点进行进一步的划分。以图 [**??**] 第一个分支结点 (喜欢 ML 作业 = 是) 为例，该结点包含有 $\{1, 2\}$ 2 个样例，且全部为“ML 成绩高”的样例，则划分结束；第

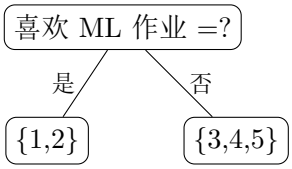


图 2: 基于喜欢 ML 作业属性对根节点进行划分

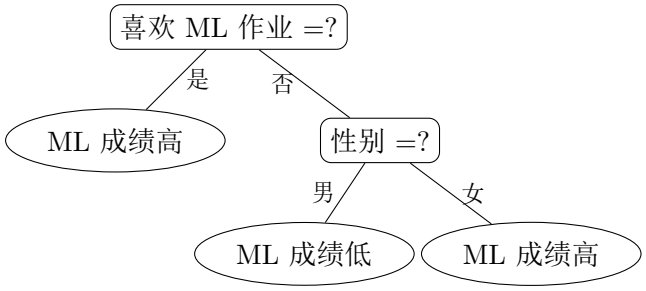
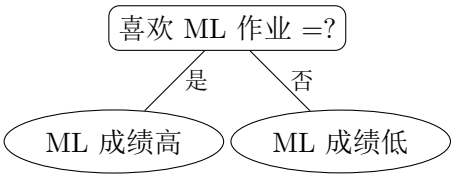


图 3: 在训练集上基于信息增益生成的决策树

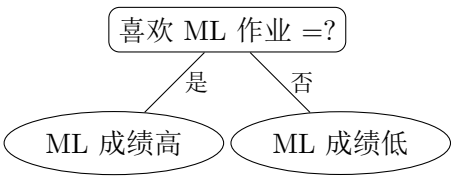
二个分支结点 (喜欢 ML 作业 = 否)，注意到该结点包含有 $\{3,4,5\}$ 3 个样例，再以“性别”作为划分属性进行划分，最终得到的决策树如图 [**??**] 所示。

(3) i) 预剪枝: 首先考虑根节点，若不进行划分，该节点被标记为叶结点，且类别应当被标记为“是”(ML 成绩高)，因此单结点决策树此时的验证集精度为: $\frac{1}{4} \times 100\% = 25\%$。

在用属性“喜欢 ML 作业”划分之后得到如图 [**??**] 所示的结果，生成的两个结点分别被标记为“是”和“否”，此时的验证集中编号为 $\{6,8,9\}$ 的样例被分类正确，精度为 $\frac{3}{4} \times 100\% = 75\% > 25\%$, 故确定以“喜欢 ML 作业”进行划分。注意到左侧结点已经划分结束，接下来考虑右侧结点，此时的验证集精度为 75%。以属性“性别”进行划分后，编号为 $\{9\}$ 的样本结果将由正确转为错误，此时的验证集精度下降为 50%，因此禁止这一结点被划分，最终得到的决策树如下图所示：



ii) 后剪枝: 先从训练集生成一颗如图 [**??**] 所示的完整决策树，易知该决策树的验证集精度为 50%。下面自底向上的对非叶结点进行考察，首先考虑图 [**??**] 中以“性别”属性进行划分的这一非叶结点，若将其领衔的分支删除，即将这一结点替换为叶结点。替换后的叶节点包含编号为 $\{3,4,5\}$ 的训练样例，故类别应标记为“否”(即 ML 成绩低)，此时的验证集精度提升为 75%, 于是应当剪枝，得到如下图所示的决策树。接下来考虑根结点，类似地，如果将其所有的分支删除易见此时的验证集精度下降为 25%, 因此不应进行剪枝。最终得到的决策树仍如下图所示:



(4) 根据 (3) 中的分析可得，预剪枝在训练集和验证集上的准确率分别为 80%，75%; 后剪枝在训练集和验证集上的准确率分别为 80%，75%。

3 [25 pts] 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) **[10 pts]** 请给出相应的 SVM 优化问题。

(2) **[15 pts]** 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) 考虑对负类分类错误的样本施加 $k > 0$ 倍于正例样本被分错的得到的惩罚。由此可得:

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

其中 $\mathcal{P}$ 是所有正例样本的上标集合，$\mathcal{N}$ 是所有负例样本的下标集合。

(2) 令 α, μ 表示 Lagrange 乘子，那么有:

$$L(\mathbf{w}, \mathrm{b}, \xi, \alpha, \mu) = \frac{1}{2}\|\mathbf{w}\|^2 + C\left(\sum_{i\in\mathcal{P}}^{m}\xi_i + k\sum_{i\in\mathcal{N}}^{m}\xi_i\right) + \sum_{i=1}^{m}\alpha_i(1 - \xi_i - y_i(\mathbf{w}\mathbf{x_i} + b)) - \sum_{i=1}^{m}\mu_i\xi_i$$

令 $\nabla_{\mathbf{w}}L = \nabla_{\mathbf{b}}L = \nabla_{\xi_\mathbf{i}}L = 0$, 则:

$$\begin{cases} \mathbf{w} & = \sum_{i=1}^{m} a_i y_i \mathbf{x}_i \\ 0 & = \sum_{i=1}^{m} a_i y_i \\ C & = (a_i + \mu_i) \cdot (\frac{1}{k}\mathbb{I}(i \in \mathcal{P} + \mathbb{I}(i \in \mathcal{N}) \end{cases}$$

由此我们可以得到相应的对偶问题:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i\alpha_j y_i y_j x_i^T x_j) \\ \text{s.t.} \quad & \sum_{i=1}^{m} y_i\alpha_i = 0 \\ & 0 \leq \alpha_i \leq C \cdot (k\mathbb{I}(i \in \mathcal{P} + \mathbb{I}(i \in \mathcal{N})) \end{aligned}$$

相应的 KKT 条件为:

$$\begin{cases} \alpha_i, \mu_i, \xi_i \geq 0 \\ \xi_i - 1 + y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 0 \\ \alpha_i(1 - \xi_i - y_i(\mathbf{w}^T \mathbf{x}_i + b)) = 0 \\ \mu_i \xi_i = 0 \end{cases}$$

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：

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i x_i \tag{3}$$

$$\mathbf{b} = \frac{1}{|S|} \sum_{s \in S} (y_s - \sum_{i \in S} \alpha_i y_i x_i^T x_s) \tag{4}$$

请完成以下任务：

(1) **[5 pts]** 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.
编程题详细实验结果见附件。