机器学习导论
习题三

181220031, 李惟康, liwk@smail.nju.edu.cn

2020 年 4 月 15 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 (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。

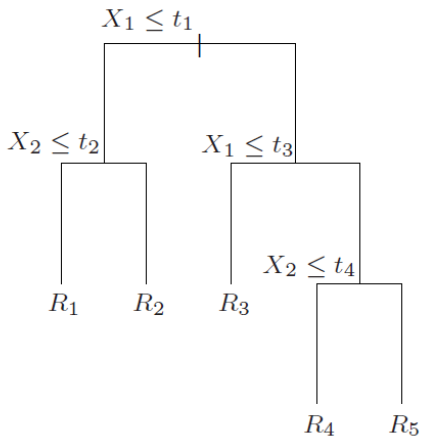

图 1: 回归决策树

(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)

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)

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）

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. TODO;

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. 此处用于写解答 (中英文均可)

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. 此处用于写解答 (中英文均可)