作业 2：决策树与提升算法

清华大学软件学院
机器学习, 2024 年春季学期

1 介绍

本次作业需要提交说明文档（PDF 形式）。注意事项如下：

- 作业按点给分，因此请在说明文档中按点回答，方便助教批改。
- 本次作业总分为 105 分，若得分超过 100 分，将按照 100 分进行截断。
- 不要使用他人的作业，也不要向他人公开自己的作业，否则处罚很严厉，会扣至-100（倒扣本次作业的全部分值）。
- 统一文件的命名：{学号}_{姓名}_hw2.zip

2 决策树与随机森林（40pt）

2.1 构建一个决策树来决定你的周末活动 (15pt)

假设你想构建一个决策树来决定你的周末活动。你需要根据以下四个属性来决定：天气（晴朗、多云、下雨）、温度（高、中、低）、湿度（高、低）、风力（强、弱）。

你有以下历史数据：

表 1: 历史数据

天气	温度	湿度	风力	活动
晴朗	高	低	弱	去徒步
晴朗	中	高	弱	去博物馆
多云	高	低	弱	去徒步
下雨	低	高	强	看电影
下雨	低	高	弱	看电影
多云	中	高	强	去博物馆

晴朗	高	低	强	去野餐
晴朗	低	低	弱	去徒步
下雨	中	高	弱	看电影
晴朗	中	低	弱	去野餐

1. 对于根节点，试使用信息增益（ID3 算法）选择最佳分割属性。(5pt)
2. 使用 ID3 算法构建决策树。(5pt)
3. 使用决策树解释下列情况应该选择什么活动：晴朗，高温，高湿，弱风。(5pt)

2.2 代码实验 (15pt)

在本题中，你将使用决策树解决二分类问题和回归问题。

1. 补全 tree.py 中 DecisionTree 类的 fit 函数。(**提示**：递归调用决策树的构造与 fit 函数,4pt)。
2. 根据决策树熵的定义，完成 tree.py 中 compute_entropy 函数 (3pt)。
3. 根据基尼系数的定义，完成 tree.py 中 compute_gini 函数 (3pt)。
4. 运行 tree.py，在实验文档中记录决策树在不同数据集上运行的结果，包括

   (a) DT_entropy.pdf，使用决策树在二分类问题上的结果。

   (b) DT_regression.pdf，使用决策树在回归问题上的结果。

   并简要描述实验现象, 例如超参数对于决策树的影响 (5pt)。

2.3 随机森林 (10pt)

假设你有一个包含 100 个样本的数据集，你需要使用随机森林算法建模。随机森林配置如下：

- 每棵树使用有放回的抽样选择样本（bootstrap sampling）。
- 每棵树训练时选取的样本数量等于原始数据集的样本数量。
- 在每个分裂点，随机选取一半的特征进行最佳分割点的寻找。

回答以下问题：

1. 当构建一棵树时, 计算任意一个样本被选中至少一次的概率。如果随机森林中有 10 棵树，计算任意一个样本在至少一棵树中被选中至少一次的概率。(5pt)
2. 假设每个样本有 20 个特征，计算任意一个特征在单次分裂时被选中进行最佳分割点的寻找的概率。如果一棵树平均有 10 个分裂点，计算任意一个特征在整棵树的生长过程中至少被选中一次的概率。(5pt)

   (**注：**本题可以不计算出最终结果，列出表达式即可)

3 Gradient Boosting Machines(30pt)

总结课件中的 Gradient Boosting Machine 的算法流程如下：

1. 令 $f_0(\mathbf{x}) = 0$。
2. For t=1 to T：
   (a) 计算在各个数据点上的梯度 $\mathbf{g}_t = \left(\frac{\partial}{\partial \widehat{\mathbf{y}}_i}\ell(\mathbf{y}_i, \widehat{\mathbf{y}}_i)|_{\widehat{\mathbf{y}}_i = f_{t-1}(\mathbf{x}_i)}\right)_{i=1}^{n}$。
   (b) 根据 $-\mathbf{g}_t$ 拟合一个回归模型，$h_t = \arg\min_{h\in\mathcal{F}}$ _____。
   (c) 选择合适的步长 α_t, 最简单的选择是固定步长 $\eta \in (0, 1]$。
   (d) 更新模型，$f_t(\mathbf{x}) =$ _____。

请完成以下题目：

1. 完成上述算法中的填空 (5pt)。
2. 考虑回归问题，假设损失函数 $\ell(\mathbf{y}, \widehat{\mathbf{y}}) = \frac{1}{2}(\mathbf{y} - \widehat{\mathbf{y}})^2$。直接给出第 t 轮迭代时的 $\mathbf{g}_t$ 以及 h_t 的表达式。（使用 f_{t-1} 表达）(5pt)。
3. 考虑二分类问题，假设损失函数 $\ell(\mathbf{y}, \widehat{\mathbf{y}}) = \ln(1 + e^{-\mathbf{y}\widehat{\mathbf{y}}})$。直接给出第 t 轮迭代时的 $\mathbf{g}_t$ 以及 h_t 的表达式。（使用 f_{t-1} 表达）(5pt)。
4. 完成 boosting.py 中 GradientBoosting 类的 fit 函数及 predict 函数 (5pt)。
5. 完成 boosting.py 中函数 gradient_l2 以及 gradient_logistic(5pt)。
6. 运行 boosting.py，在实验文档中记录 GBM 在不同数据集上运行的结果，包括
   (a) GBM_l2.pdf，使用 L2 loss 在二分类问题上的结果。
   (b) GBM_logistic.pdf，使用 logistic loss 在二分类问题上的结果。
   (c) GBM_regression.pdf，使用 L2 loss 在回归问题上的结果。

   并简要描述实验现象（例如超参数和损失函数对于 GBM 的影响等）(5pt)。

4 个人年收入预测 (35pt)

Adult 数据集（也称为 “Census Income” 数据集）是一个常用于评估机器学习算法性能的标准数据集。该数据集来源于 1994 年美国人口普查局数据库，主要任务是根据一系列变量预测个体的年收入是否超过 50,000 美元。该数据集中的特征变量如表2所示。在本题中，你需要基于特征变量预测个体年收入是否超过 50,000 美元。本题代码文件为 `adult_classification.py`。

表 2: Adult 数据集字段解释

字段名称	**描述**
`age`	个人的年龄。
`workclass`	个人的工作类型。
`fnlwgt`	人口普查专家认为该人所代表的人群数量。
`education`	个人的最高教育程度，如 Bachelors, Some-college 等。
`education-num`	受教育的年数。
`marital-status`	婚姻状况，如 Married-civ-spouse, Divorced 等。
`occupation`	工作职业，如 Tech-support, Craft-repair 等。

relationship	个体相对于家庭的角色，如 Wife, Own-child 等。
race	种族。
sex	性别。
capital-gain	资本收益。
capital-loss	资本损失。
hours-per-week	每周工作时数。
native-country	个体的原籍国家。
income	年收入，标记为 <=50K 或 >50K。

你需要完成以下任务：

1. 数据预处理与特征工程 (10pt)：
   (a) 处理缺失值，可以使用填充或者删除等策略;
   (b) 对类别型特征进行编码，选择合适的编码方法 (如独热编码、标签编码);
   (c) 归一化数值型特征。
2. 使用决策树算法进行个人年收入的预测，汇报模型在训练集和验证集上的正确率。本题可以直接使用 `sklearn.tree` 算法包中的 `DecisionTreeClassifier` 算法 (5pt)。
3. 使用随机森林算法进行个人年收入的预测，汇报模型在训练集和验证集上的正确率，并解释你的参数选择，如树的数量、树的深度、特征数量等。本题可以直接使用 `sklearn.tree` 算法包中的 `RandomForestClassifier` 算法 (5pt)。
4. 对决策树进行可视化，分析数据集不同字段对决策过程的影响。在本题中，你可以使用 `sklearn.tree` 算法包中的 `plot_tree` 函数方便地对决策树进行可视化。当使用不同的节点分裂策略或不同的树深度时，得到的决策树有什么变化？(**提示**: 用该函数可视化得到的每个决策节点都有两个分支，通常左边是特征的值小于等于阈值的分支，右边是特征的值大于阈值的分支)(10pt)。
5. 使用 XGBoost 算法进行个人年收入的预测，汇报模型在训练集和验证集上的正确率，并解释你的参数选择，如树的深度、学习率等。(5pt)。