燕山大学

Python 机器学习实验指导书

Python machine learning Experiment Instruction Book

实验三：决策树

教 务 处

2023 年 5 月

实验三　决策树

一、实验目的

1.理解并掌握编经典的决策树分类算法。

2.能够基于决策树分类算法实现鸢尾花分类与预测。

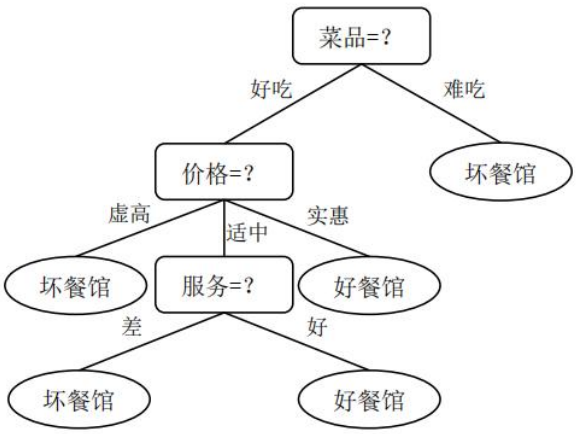
3.能够举一反三，基于决策树分类算法实现肿瘤、购买服装等数据的分析与预测。

二、实验原理

（一）决策树

决策树（decision tree）是一种应用广泛的机器学习方法。顾名思义，决策树算法的表现形式可以直观理解为一棵树（可以是二叉树或非二叉树）。 一棵决策树一般包含一个根节点、一系列内部节点和叶节点，一个叶节点对应一个决策结果，一个内部节点则对应一个属性测试。 算法的目标是根据训练集中的样本和标签，自动生成一棵具有泛化能力的决策树。它反映了人对分类问题的判断过程，即类别的判断由对一系列特征的区分（属性测试）组成。

以二分类问题为例：我们在判断一间餐馆是不是好餐馆时，我们将会从它的某些特征来进行判断决定：餐馆的菜品是否好吃？价格是否合理？服务态度是否良好？根据这些问题，我们可以构建一棵决策树，用以判断最终的问题：是不是一间好餐馆？但是，只考虑某个单一的属性进行提问并不能很好地对正类与负类进行划分，例如服务员态度好的餐馆不一定就好，也存在部分不好的餐馆符合这一判断。因此需要与其它特征一同进行综合判断。



决策树的算法在自动构建之后，能还原出每一项特征的具体划分标准。因此，决策树具有解释性较好的特点。

决策树算法的目标是根据由特征和标签组成的训练数据，自动生成一棵能对未见数据进行分类的决策树。

决策树算法学习的关键在于如何在每一次进行决策判断时选取到最优划分属性，其中一个基本的准则是：决策树按照某个属性进行划分后，得到的分支节点所包含的样本应该尽可能属于同一类别,相当于分类后同个分支节点所包含的样本类别”越纯“。而要达到这一点，我们需要一个衡量节点“纯度”的函数。 当某一属性划分的“纯度”最高时，我们选择对这一属性进行划分来建立下一个分支。理想的，通过对更多的属性进行划分，结点的 “纯度”会越来越高。现有的计算“纯度”的主流方法有：信息增益、增益率、基尼指数。

1.信息增益

信息增益的诞生基于“信息熵”的思想。“信息熵”(Information Entropy) 概念是由克劳德·香农(Claude Shannon) 于 1948 年提出，是一个以变量的不确定性来量化信息的指标。而在决策树算法的语境下，“信息熵”则可以看作是分类的不确定性，即进行某个属性进行划分之后，对叶节点中的样本类别的不确定程度。

信息熵的值越小，则样本的纯度越高。当我们准备选择新的属性来建立分支时，就需要计算信息增益，即信息熵的变化。信息增益可以作为选择用以测试的属性，著名的 ID3 便是采用的是信息增益。

2.增益率

信息增益的最大缺点就是分叉越多，信息增益就越大，换句话说，它倾向于选择属性值比较多的特征来分叉。但是，有时这样的属性对分类器的泛化性是有害的。例如，假如我们将样本的“编号”也作为一种属性，即假如有v个样本，这个属性就有v个可能值。使用该属性进行测试则产生v个分支，此时根据信息增益公式，得到的纯度提升非常大，但是却对分类模型性能没有实质提升，导致过拟合训练样本泛化性能低。因此另一个著名的决策树算法 C4.5 提出了“增益率”概念。

但是，增益率指标偏好于可能值较少的属性，因此有时也无法取得最佳效果。较好的方法可以将信息增益与增益率两种指标结合使用，例如先找出信息增益高于一定阈值的那些属性，再根据增益率的高低来筛选剩下的属性。

3.基尼指数

基尼指数是另一衡量经过划分的数据集纯度的指标,最先被 CART 决策树算法使用。由于其优异的性能及计算效率,这一指标被 sklearn 等实现作为默认的决策树属性选择指标。

直观上理解，基尼指数代表着从数据集中随机抽取的两个样本类别标记不一致的概率。因此基尼指数越小，则划分后的数据集纯度越高。因此与其它之前介绍的指标不同，

我们选取基尼指数最小的属性作为最优划分属性。

4.剪枝处理

在决策树算法的训练中，往往可能会发生“过拟合”。也就是决策树为了提高划分数据集的纯度，生成过多的分支，但是每个分支下的叶节点仅有相当少的样本。符合该分支情况的样本可能并不会出现在实际测试数据中，而仅仅出现在训练集中。为了避免这种把训练集自身的一些特点当作所有数据都具有的普遍性质的有害做法，决策树需要进行“剪枝”操作。

决策树的剪枝操作可以分为两种——“预剪枝”和“后剪枝”：

（1）预剪枝：设定一些规则来避免树的过度生长。

- 信息增益（率）少于阈值就不再生长。
- 节点的样本数少于阈值（例如 1%）就不再生长。
- 不允许叶子节点的样本数少于某个阈值（例如 0.5%）。
- 不允许树的深度超过阈值（例如 8 层）。

（2）后剪枝：先让决策树无限的生长成一棵很大的树（这个时候肯定过拟合了），然后再剪枝。

从下往上依次判断：如果剪掉子树（把父节点作为叶子节点）能否让验证集的误差下降，如果可以，那么就减掉，不断重复此操作。

这里的关键问题就变成了：如何判断性能是否可以得到提升？这时就需要验证集的帮助。在进行是否要进行剪枝操作的判断时，使用一组没有出现在训练集中的数据，分别测试剪枝前后的决策树性能。

（二）鸢尾花分类数据集 iris

鸢尾花分类是机器学习中比较经典的入门式教学课程。

鸢尾花数据集总共包含 150 行数据。每一行数据由 4 个特征值及一个目标值组成。

4 个特征值分别为：萼片长度（SepalLengthCm）、萼片宽度（SepalWidthCm）、花瓣长度（PetalLengthCm）、花瓣宽度（PetalWidthCm）。

目标值为三种不同类别的鸢尾花，分别为：Iris-setosa（山鸢尾），Iris-versicolor（杂色鸢尾），Iris-virginica（维吉尼亚鸢尾）。

```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
```

鸢尾花分类数据集

鸢尾花分类

（三）昆明 04、06-17 年部分气象数据集

该数据集采集自昆明 04、06-17 年的部分气象数据。

数据集共有 4869 行样本数据，每行样本数据包括如下标签：平均气压、日最低气压、平均气温、日最低气温、平均相对湿度、累计降水量、大型蒸发量、平均风速、最大风速、日照时数、平均地表气温、日最低地表气温、是否下雪。

年	月	日	平均气压	日最低气压	平均气温	日最低气温	平均相对湿度	累计降水量	大型蒸发量	平均风速	最大风速	日照时数	平均地表气温	日最低地表气温	是否下雪
2004	1	1	811.2	808.6	10.8	4.9	6	0	0.8	1.5	4.6	9.2	10.4	-0.2	NO
2004	1	2	812.1	808.7	10.6	6	6.1	0	0.7	1.5	5.3	8.2	11.5	1.6	NO
2004	1	3	811.1	809.1	10.7	5.4	5.3	0	1	1.8	5.2	9.4	10.3	-1.2	NO
2004	1	4	812.1	810.5	10.5	4.1	5.3	0	3.2	2.2	6.2	9.3	10	-1.2	NO
2004	1	5	814.1	812.3	10.3	3.9	5.1	0	3.3	1.8	6.3	9.4	10.1	-2.1	NO
2004	1	6	813.7	811.5	11.1	3.9	4.7	0	2.9	1.5	4.6	9.3	10.4	-2.2	NO
2004	1	7	813.5	811.8	11.4	5.2	4.7	0	3.3	2.2	6.5	9.5	10	-1.3	NO
2004	1	8	814.1	812.6	11.2	5.6	5.6	0	2.4	2	6.1	8.7	10	-0.2	NO
2004	1	9	813.6	810.3	8.7	6.2	7.9	25.5	1.3	2.5	6.4	1.4	8.4	3.6	NO
2004	1	10	813.9	812.5	8.3	3.1	6.5	0	2.4	1.6	2.5	9.5	7.6	-3.3	NO
2004	1	11	814.2	811.7	8.4	2	6.4	0	2.5	1.2	4.6	9.1	7.1	-2.9	NO
2004	1	12	813.6	811.7	9.4	2.8	6.5	0	2.4	1.1	4	8.8	8.3	-2.4	NO
2004	1	13	810.5	807.7	10.6	6	6.4	0	2.6	2.2	6.9	8.7	9.5	0.9	NO
2004	1	14	808.4	806.6	10.1	5.2	6.2	0	2.7	2.8	5.8	7.9	9.2	-0.4	NO
2004	1	15	808.7	806.6	10.4	7.4	5.2	0	3.4	3	7.1	9.1	7.8	1.1	NO
2004	1	16	811.1	808.7	9.2	3.4	5.5	0	2.7	1.5	4	8.7	6.9	-2.8	NO
2004	1	17	809.3	806.1	8.5	3.3	6.1	0	2.4	2.5	6.2	8.5	6.5	-2.4	NO
2004	1	18	808.7	807.3	9.3	4.4	6.1	0.001	1.8	1.4	4	3.7	7	-1.1	NO
2004	1	19	808.3	806.1	9.9	4.2	5.5	0	2.9	1.6	5.8	9	9.2	-2	NO
2004	1	20	808.6	806.6	11	5.5	5	0.001	2.9	2.6	5.6	8.8	11.4	-0.7	NO

昆明 04、06-17 年部分气象数据集

（四）威斯康星乳腺癌数据集

威斯康星乳腺癌数据集来自美国威斯康星州的乳腺癌诊断数据集。医疗人员采集了患者乳腺肿块经过细针穿刺（FNA）后的数字化图像，并且对这些数字图像进行了特征提取，这些特征可以描述图像中的细胞核呈现。

该数据集中肿瘤是一个非常经典的用于医疗病情分析的数据集，包括 569 个病例的数据样本，每个样本具有 30 个特征。样本共分为两类：恶性（Malignant）和良性（Benign）。

字段	含义
ID	ID标识
diagnosis	M/B（M：恶性，B：良性）
radius_mean	半径（点中心到边缘的距离）平均值
texture_mean	文理（灰度值的标准差）平均值
perimeter_mean	周长 平均值
area_mean	面积 平均值
smoothness_mean	平滑程度（半径内的局部变化）平均值
compactness_mean	紧密度（=周长*周长/面积-1.0）平均值
concavity_mean	凹度（轮廓凹部的严重程度）平均值
concave points_mean	凹缝（轮廓的凹部分）平均值
symmetry_mean	对称性 平均值
fractal_dimension_mean	分形维数（=海岸线近似-1）平均值
radius_se	半径（点中心到边缘的距离）标准差
texture_se	文理（灰度值的标准差）标准差
perimeter_se	周长 标准差
area_se	面积 标准差
smoothness_se	平滑程度（半径内的局部变化）标准差
compactness_se	紧密度（=周长*周长/面积-1.0）标准差
concavity_se	凹度（轮廓凹部的严重程度）标准差
concave points_se	凹缝（轮廓的凹部分）标准差
symmetry_se	对称性标准差
fractal_dimension_se	分形维数（=海岸线近似-1）标准差
radius_worst	半径（点中心到边缘的距离）最大值
texture_worst	文理（灰度值的标准差）最大值
perimeter_worst	周长 最大值
area_worst	面积 最大值
smoothness_worst	平滑程度（半径内的局部变化）最大值
compactness_worst	紧密度（=周长*周长/面积-1.0）最大值
concavity_worst	凹度（轮廓凹部的严重程度）最大值
concave points_worst	凹缝（轮廓的凹部分）最大值
symmetry_worst	对称性 最大值
fractal_dimension_worst	分形维数（=海岸线近似-1）最大值

威斯康辛乳腺癌数据集特征

属性信息：

ID——身份证号码

diagnose——诊断结果（M=恶性，B=良性）

其中’B’代表良性，包含 357 例；’M’代表恶性，包含 212 例。

计算每个细胞核的 10 个实值特征：

1） radius_mean：半径（从中心到周界各点的平均距离）

2） texture_mean：纹理（灰度值的标准偏差）

3） perimeter_mean：周长

4） area_mean：面积

5） smoothness_mean ：平滑度（半径长度的局部变化）

6） compactness_mean：密实度/紧密度（周长^2/面积-1.0）

7） concavity_mean：凹度（轮廓凹陷部分的严重程度）

8） concave points_mea ：凹点（轮廓凹面部分的数量）

9） symmetry_mean：对称性

10）fractal_dimension_mean：分形维数（“海岸线近似值”-1）

Mean、se、worst：为每个图像计算这些特征，产生了 30 个特征。所有特征值用四个有效数字重新编码。

- 包含 mean 的数据——平均值。
- 包含 se 的数据——标准误差。

包含 worst 的数据——最差值或最大值（三者中的平均值最大值），是最严重的数据样例(最坏值)。

（五）顾客购买服装数据集

顾客购买服装数据集包含 4 个特征值和 1 个目标值。

- 特征值：
  - review（商品评价变量）、
  - discount（打折程度）、
  - needed（是否必需）、
  - shipping（是否包邮）。
- 目标值：buy（是否购买）。

review	discount	needed	shipping	buy
3	3	0	1	1
3	3	0	0	1
2	3	0	1	0
1	2	0	1	0
1	1	1	1	0
1	1	1	0	1
2	1	1	0	0
3	2	0	1	1
3	1	1	1	0
1	2	1	1	0
3	2	1	0	0
2	2	0	0	1
2	3	1	1	0
1	2	0	0	1

顾客购买服装数据集

三、实验环境

计算机；网络环境。

四、实验内容及步骤

（一）实验内容

1.对于给定的例题，基于决策树分类算法进行鸢尾花分类与预测的练习。

2.对于给定的项目，自行编写程序，使用决策树分类算法对威斯康星乳腺癌数据集、顾客购买服装数据集进行分析与预测。

（二）实验步骤

1.【选做】进入指定实验课程，可通过阅览“知识讲解”进行实验相关知识的查缺补漏。

- 知识讲解：决策树

2.【必做】在熟悉了实验原理的基础上，进行“实验三：决策树”例题部分的实操练习。

- 例题 1：鸢尾花分类（决策树）【必做】
- 例题 2：预测降雪【扩展.选做】

3.【必做】在步骤 2 例题练习的基础上，对给定的实操项目，自行编写程序，使用决策树分类算法对威斯康星乳腺癌数据集、顾客购买服装数据集进行分析与预测。

- 实操项目 1——肿瘤预测（决策树）
- 实操项目 2——顾客购买服装的分析与预测

4.【必做】完成实验报告。

请严格基于实验报告的模板撰写实验报告。

本课程所有实验全部结束后再统一打印。

附：实操项目要求

- **实操项目 1——肿瘤预测（决策树）**

基于威斯康辛乳腺癌数据集，采用决策树的方法进行肿瘤预测。

【实验要求】

1.加载 sklearn 自带的威斯康星乳腺癌数据集，探索数据。

2.进行数据集分割。

3.配置决策树模型。

4.训练决策树模型。

5.模型预测。

6.模型评估。

7.参数调优。可以根据评估结果，对模型设置或调整为更优的参数，使评估结果更准确。

- **实操项目 2——顾客购买服装的分析与预测**

采用决策树算法，对“双十一”期间顾客是否买服装的数据集进行分析与预测。

顾客购买服装数据集：包含 review（商品评价变量）、discount（打折程度）、needed（是否必需）、shipping（是否包邮）、buy（是否购买）。

【实验要求】

1.读取顾客购买服装的数据集（数据集路径：data/data76088/3_buy.csv），探索数据。

2.分别用 ID3 算法和 CART 算法进行决策树模型的配置、模型的训练、模型的预测、模型的评估。

3.扩展内容（选做）：对不同算法生成的决策树结构图进行可视化。

封面设计： 贾丽

地 址：中国河北省秦皇岛市河北大街 438 号
邮 编：066004
电 话：0335-8057068
传 真：0335-8057068
网 址：http://jwc.ysu.edu.cn