



# 燕山大学

## Python 机器学习实验指导书

---

Python machine learning Experiment Instruction Book

### 实验五：AdaBoost

教 务 处

2021 年 2 月

# 实验五 AdaBoost

## 一、实验目的

1. 理解并掌握集成学习中 AdaBoost 算法。
2. 能够基于 AdaBoost 算法实现鸢尾花分类。
3. 能够举一反三，基于 AdaBoost 算法实现肿瘤预测。

## 二、实验原理

集成学习是结合多个个体学习器（弱学习器），从而获得精确率很高的学习器（强学习器）。它是通过对多个模型进行整合，从而获得更好的学习效果的过程。

集成学习的两种重要策略是自举汇聚（Bagging）和提升（Boosting）。提升是一个逐步优化集成学习器的过程，即每个个体学习器都在弥补集成学习器的欠缺，从而达到整体的优化。Boosting 的经典算法是 AdaBoost（Adaptive Boosting，自适应增强）。

### （一）AdaBoost 算法

AdaBoost 是一种迭代算法，通过每次降低个体学习器的分类误差，加大效果好的个体学习器的重要性，得到最终的集成学习器。

#### 1. 核心思想

针对同一个训练集训练不同的分类器（弱分类器），然后把这些弱分类器集合起来，构成一个更强的最终分类器（强分类器）。

#### 2. 算法流程

该算法其实是一个简单的弱分类算法提升过程，这个过程通过不断的训练，可以提高对数据的分类能力。过程如下：

- （1）先通过对  $N$  个训练样本的学习得到第一个弱分类器；
- （2）将分错的样本和其他的新数据一起构成一个新的  $N$  个的训练样本，通过对这个样本的学习得到第二个弱分类器；
- （3）将（1）和（2）都分错了的样本加上其他的新样本构成另一个新的  $N$  个的训练样本，通过对这个样本的学习得到第三个弱分类器；
- （4）最终经过提升的强分类器。即某个数据被分为哪一类要由各分类器权值决定。

由 AdaBoost 算法的描述过程可知，该算法在实现过程中根据训练集的大小初始化样本权值，使其满足均匀分布，在后续操作中通过公式来改变和规范化算法迭代后样本的权值。样本被错误分类导致权值增大，反之权值相应减小，这表示被错分的训练样本集包括一个更高的权重。这就会使在下轮时训练样本集更侧重于难以识别的样本，针对被

错分样本的进一步学习来得到下一个弱分类器，直到样本被正确分类。在达到规定的迭代次数或者预期的误差率时，则强分类器构建完成。

总而言之，误差率低的弱分类器在最终分类器中占的权重较大，否则较小。

### 3. 应用

对 AdaBoost 算法的研究以及应用大多集中于分类问题，同时也出现了一些在回归问题上的应用。就其应用 adaBoost 系列主要解决了：两类问题、多类单标签问题、多类多标签问题、大类单标签问题、回归问题。

#### （二）鸢尾花数据集 iris

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）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	纹理 (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	纹理 (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	纹理 (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（ $\text{周长}^2/\text{面积}-1.0$ ）
- 7) concavity\_mean: 凹度（轮廓凹陷部分的严重程度）
- 8) concave points\_me : 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)。

### 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 AdaBoost 集成学习算法进行鸢尾花分类的练习。
2. 对于给定的项目，自行编写程序，使用 AdaBoost 集成学习算法实现肿瘤预测。

#### （二）实验步骤

1. 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：AdaBoost
2. 在熟悉了实验原理的基础上，进行“实验五：AdaBoost”例题部分的实操练习。
  - 例题：鸢尾花分类（AdaBoost）
3. 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 AdaBoost 算法实现肿瘤预测。
  - 实操项目——肿瘤预测（AdaBoost）
4. 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——肿瘤预测 (AdaBoost)

基于威斯康星乳腺癌数据集，使用 AdaBoost 算法实现肿瘤预测。

#### 【实验要求】

1. 加载 sklearn 自带的数据集，使用 DataFrame 形式探索数据。
2. 划分训练集和测试集，检查训练集和测试集的平均癌症发生率。
3. 配置模型，训练模型，模型预测，模型评估。
  - (1) 构建一棵最大深度为 2 的决策树弱学习器，训练、预测、评估。
  - (2) 再构建一个包含 50 棵树的 AdaBoost 集成分类器（步长为 3），训练、预测、

评估。

参考：将决策树的数量从 1 增加到 50，步长为 3。输出集成后的准确度。

(3) 将 (2) 的性能与弱学习者进行比较。

4. 绘制准确度的折线图，x 轴为决策树的数量，y 轴为准确度。

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>