Artificial Intelligence
人工智能实验

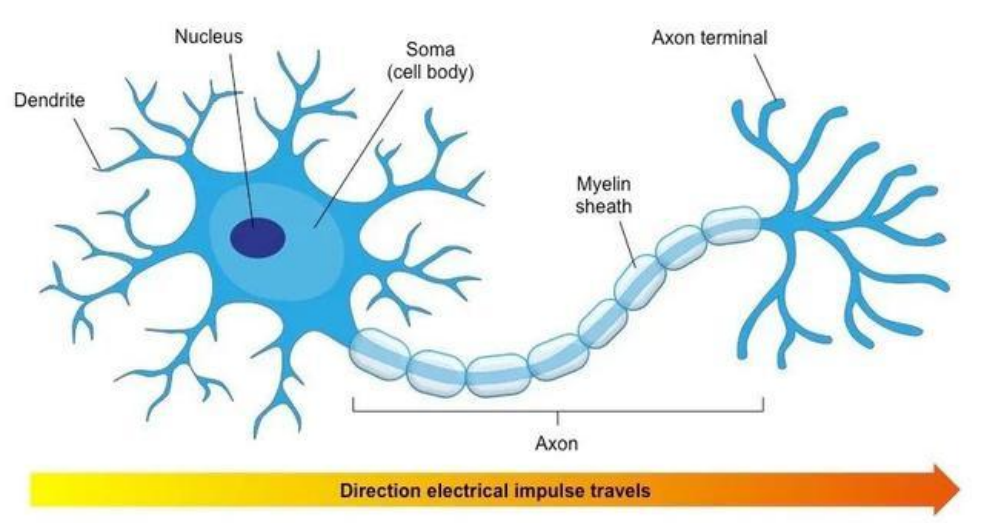
机器学习基础

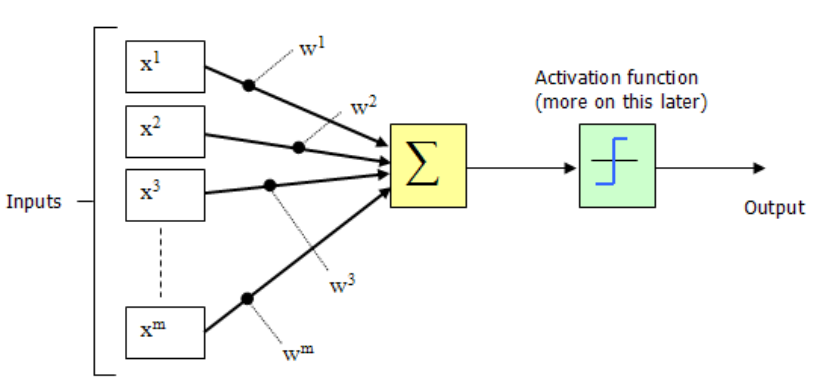
中山大学计算机学院
2025年春季

目录

1.1 人工神经网络介绍

- 神经网络采用了仿生学的思想，通过模拟生物神经网络的结构和功能来实现建模



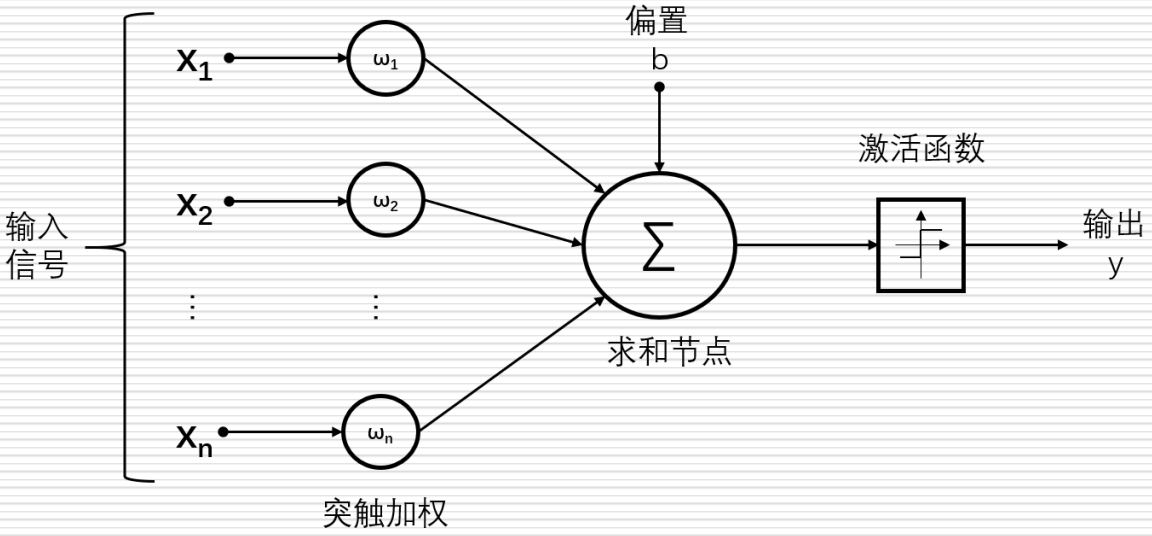


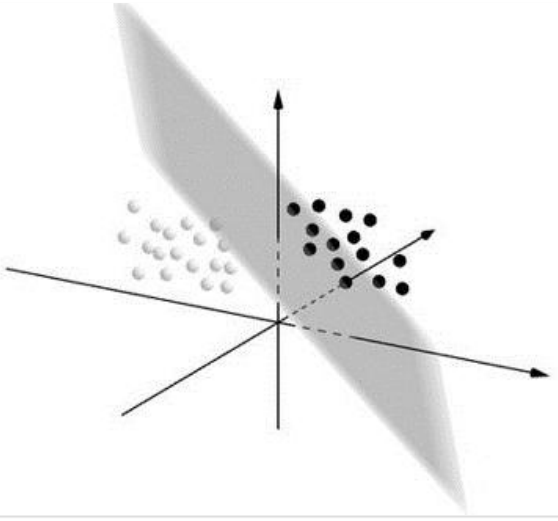
$$y = f(\sum_{i=1}^{n} w_i x_i - \theta)$$

1.1 人工神经网络介绍

单层感知机

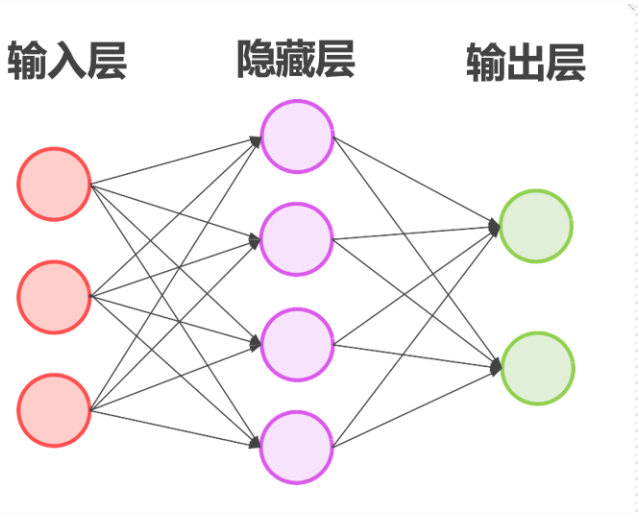
- 由于M-P神经元模型参数需要事先设定好，为了能够自适应学习出所需要的参数，研究人员就提出了单层感知机(Perceptron)
- 感知机的基本公式为：$y(x)=sign(wx+b)$
- sign为符号函数，当自变量为正数时取值为1，否则取值为0

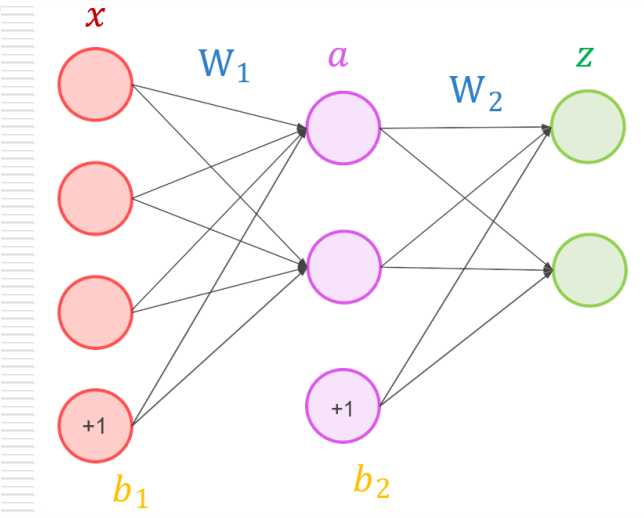




1.1 人工神经网络介绍

- 多层感知机（Multi-Layer Perceptron，简称MLP）
  - 包含三个层次：一个输入层，一个或多个中间层 (也叫隐藏层，hidden layer) 和一个输出层
  - 输入层与输出层的节点数是固定的，中间层则可以自由指定
  - MLP通常还会引入偏置单元***b***





$$a = g_1(W_1 \times x + b_1)$$

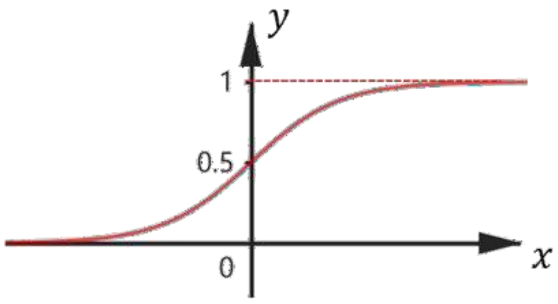
$$z = g_2(W_2 \times a + b_2) = g_2(W_2 \times g_1(W_1 \times x + b_1) + b_2)$$

1.1 人工神经网络介绍

激活函数

- 常用激活函数：
  - sigmoid、Relu、tanh

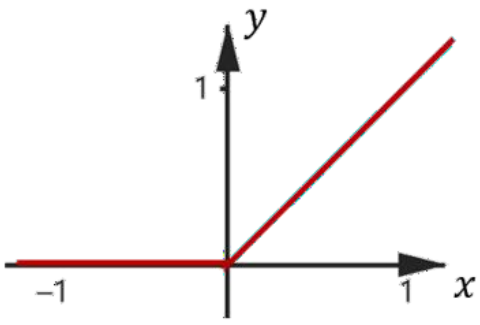
$$sigmoid(x) = \frac{1}{1+e^{-x}}$$



(a) sigmoid

多用于多层神经网络(MLP)

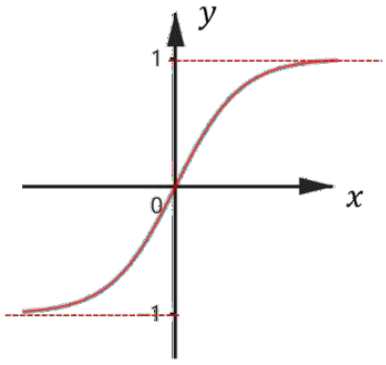
$$ReLU(x) = \max(0, x)$$



(b) ReLU

多用于深层神经网络(DNN)

$$tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$$



(c) tanh

多用于深层神经网络(DNN)

1.1 人工神经网络介绍

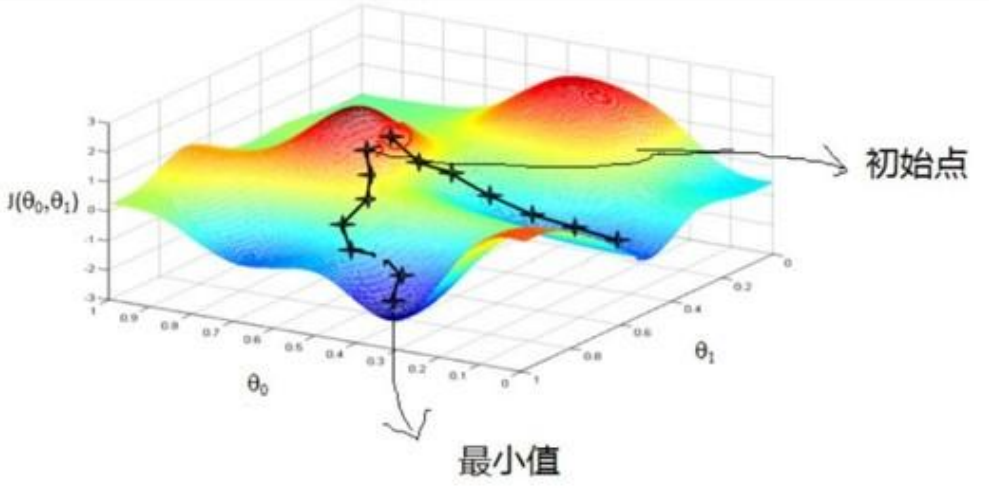
损失函数

- 作用：为了衡量网络表现是否良好，并为之后的网络参数优化提供指导。
- 常见的用在分类任务上的损失函数：
  - 均方误差(MSE): $L_{MSE} = \frac{1}{2n}\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2$
  - 交叉熵: $L_{CE} = -\sum_{i}^{n} y^{(i)} \log \hat{y}^{(i)} + \left(1 - y^{(i)}\right) \log(1 - \hat{y}^{(i)})$

1.1 人工神经网络介绍

梯度下降

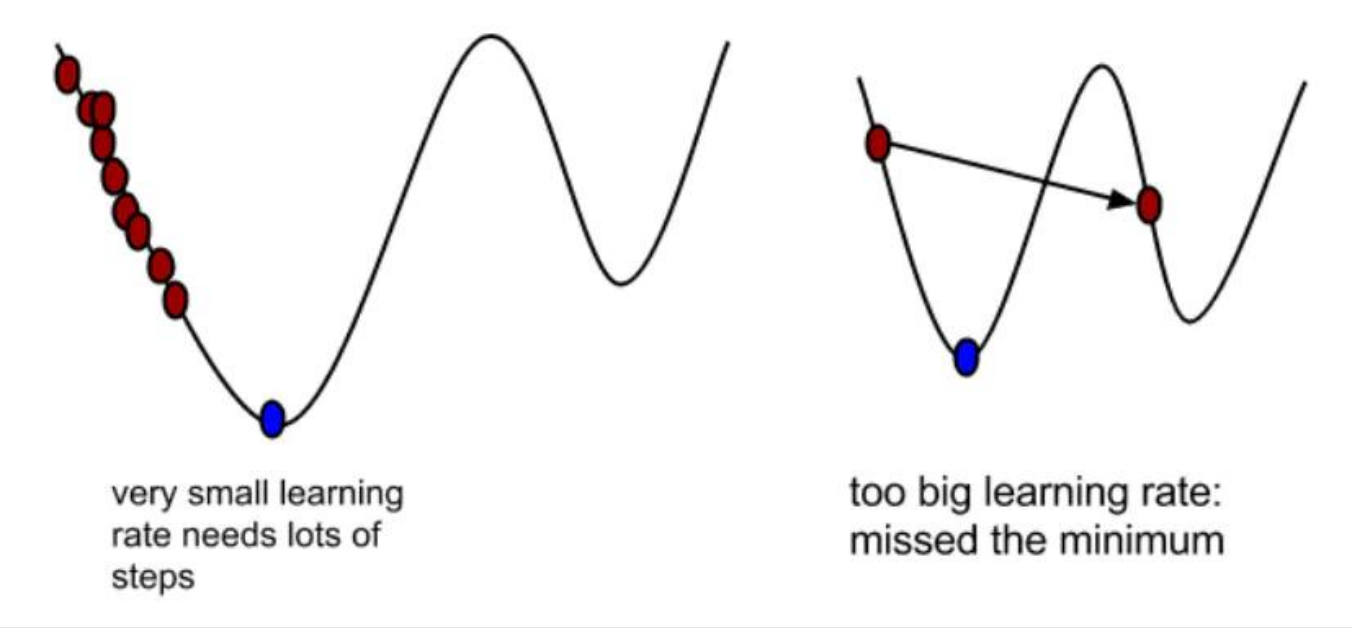
- 梯度定义：梯度是一个向量，表示某一函数在该点出的方向导数沿着该方向取得最大值。
- 也就是说该点处沿着梯度的方向变化最快，变化率最大
  - 沿着梯度方向容易找到函数最大值
  - 沿着梯度方向的反方向，容易找到函数最小值
- 梯度下降的一般公式为：$\theta = \theta - \eta \nabla_\theta L(\theta)$，其中，$\eta$是学习率，$\nabla_\theta$是对$\theta$的梯度，$\theta$是参数



1.1 人工神经网络介绍

学习率

- 学习率限制了下一步能到达的地方
- 如果学习率太小，可能很难达到最小值
- 如果学习率太大，则会错过最小值，无法收敛



1.2 决策树

基本概念

- 决策树基于“树”结构进行决策
  - 每个“内部结点”对应于某个属性上的“测试”(test)
  - 每个分支对应于该测试的一种可能结果（即该属性的某个取值）
  - 每个“叶结点”对应于一个“预测结果”
- 学习过程：通过对训练样本的分析来确定“划分属性”（即内部结点所对应的属性）
- 预测过程：将测试示例从根结点开始，沿着划分属性所构成的“判定测试序列”下行，直到叶结点

1.2 决策树

基本流程

- 策略：“分而治之”(divide-and-conquer)
  - 自根至叶的递归过程；
  - 在每个中间结点寻找一个“划分”(split or test)属性。
- 三种停止条件：
  - 当前结点包含的样本全属于同一类别，无需划分；
  - 当前属性集为空，或是所有样本在所有属性上取值相同，无法划分；
  - 当前结点包含的样本集合为空，不能划分。

1.2 决策树

□ 基本算法

输入： 训练集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$;
属性集 $A = \{a_1, a_2, \ldots, a_d\}$.

过程： 函数 TreeGenerate(D, A)

1: 生成结点 node;
2: **if** D 中样本全属于同一类别 C **then**
3: 将 node 标记为 C 类叶结点; **return** 终止条件1
4: **end if**
5: **if** $A = \varnothing$ **OR** D 中样本在 A 上取值相同 **then**
6: 将 node 标记为叶结点, 其类别标记为 D 中样本数最多的类; **return** 终止条件2
7: **end if**
8: 从 A 中选择最优划分属性 a_*;
9: **for** a_* 的每一个值 a_*^v **do**
10: 为 node 生成一个分支; 令 D_v 表示 D 中在 a_* 上取值为 a_*^v 的样本子集;
11: **if** D_v 为空 **then**
12: 将分支结点标记为叶结点, 其类别标记为 D 中样本最多的类; **return** 终止条件3
13: **else**
14: 以 TreeGenerate(D_v, $A \setminus \{a_*\}$)为分支结点 递归处理
15: **end if**
16: **end for**

输出： 以 node 为根结点的一棵决策树

1.2 决策树

常用划分属性的方法

- 信息增益（ID3）：若以属性a来进行划分，属性a可取值为$a^1, a^2, \ldots, a^V$，属性集D在a^v上的样本为D^v，那么以属性a对样本进行划分的信息增益为

$$Gain(D,a) = Ent(D) - \sum_{v=1}^{V} \frac{|D_v|}{|D|} Ent(D^v),\ Ent(D) = -\sum_{k=i}^{|y|} p_k \log_2 p_k$$

- 增益率（C4.5）：在使用信息增益率的时候，一个属性的取值越多，信息增益越高，为此引入增益率来进行属性划分

$$Gain\_ratio(D,a) = \frac{Gain(D,a)}{IV(a)}, IV(a) = -\sum_{i=1}^{V} \frac{|D_v|}{|D|} \log_2 \frac{|D_v|}{|D|}$$

- 基尼指数（CART）：CART分类树是一个二分类树，在所有属性的所有划分点的里面寻找具有最小基尼指数的点作为划分点

$$Gini(D) = \sum_{k=1}^{|y|} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{|y|} p_k^2, Gini\_index(D,a) = \sum_{v=1}^{V} \frac{|D_v|}{|D|} Gini(D^v)$$

1.2 决策树

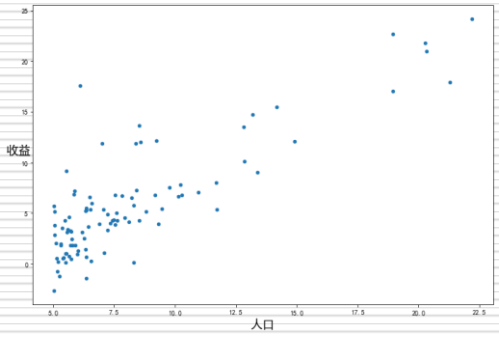
剪枝

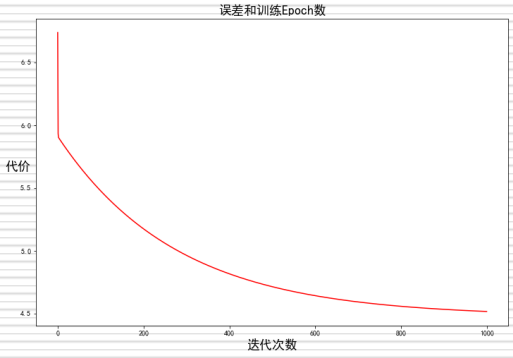
- 剪枝是为了获得更好的泛化性能，剪枝分为预剪枝与后剪枝。
  - 预剪枝：提前终止某些分支的生长。
  - 后剪枝：在决策树已经建立的基础上，把某些分割的点用叶子节点来替代。
- 剪枝评估：剪枝即剪去不必要的、不应该得到的分支，剪枝的过程需要采用模型评估的方法去评估剪枝前后的优劣
- 对比：
  - 时间开销：预剪枝测试时间开销降低，训练时间开销降低；后剪枝测试时间开销降低，训练时间开销增加
  - 过/欠拟合风险：预剪枝过拟合风险降低，欠拟合风险增加；后剪枝过拟合风险降低，欠拟合风险基本不变
  - 泛化性能：后剪枝通常优于预剪枝

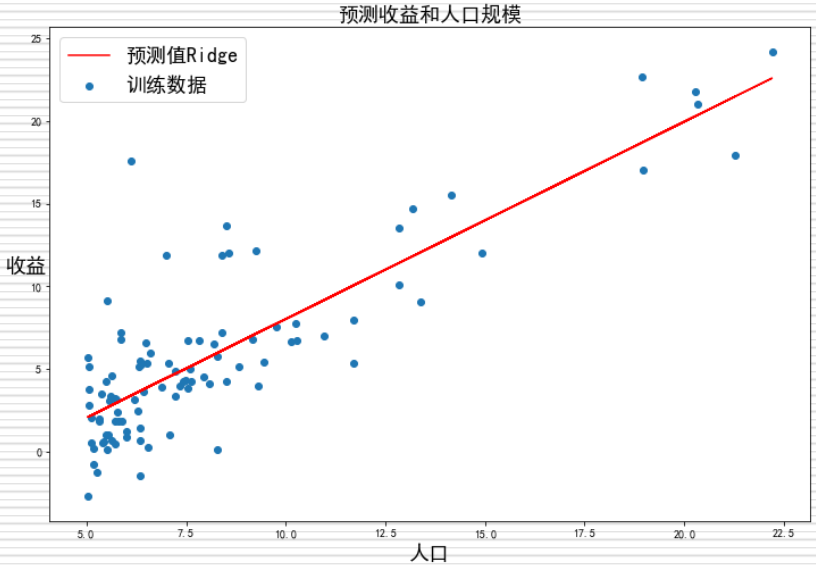
2.1 购房预测分类任务

购房预测分类任务

- 利用**感知机算法**在给定数据集完成购房预测训练。
- 要求：
  - 选择合适的损失函数，利用训练集完成网络训练，画出数据可视化图、loss曲线图。
  - 单层感知机示例：







2.1 购房预测分类任务

购房预测分类任务

- 利用**感知机算法**在给定数据集完成购房预测训练

1. 假定网络为多层感知机，网络输出为 $\hat{y} = MLP(X_{train})$，X为房子的特征， MLP为多层神经网络， W为MLP的参数， $\hat{y}$为预测的房价
2. 设置损失函数为L_{MSE}，并随机初始化网络参数
3. 当满足终止条件时，终止优化，否则继续
4. 计算网络输出$\hat{y} = MLP(X_{train})$，以及损失$L_{MSE} = (MLP(X_{train}) - Y)^2$，Y为真实房价
5. 对所有网络参数求导$\frac{\partial L_{MSE}}{\partial W}$
6. 根据$W = W - \eta \frac{\partial L_{MSE}}{\partial W}$更新参数$W$， η为学习率（步长）
7. 跳转到3

2.2 信誉度分类任务

- 信誉度分类任务（无需提交）
  - 利用决策树算法在给定数据集完成信誉度分类训练。
  - 要求：
    - 选择合适的决策树算法以及剪枝方法，利用训练集完成决策树的构建，计算决策树模型的分类准确率。

3. 作业提交说明

- 压缩包命名为：“学号_姓名_作业编号”，例：20250414_张三_实验4。
- 每次作业文件下包含两部分：code文件夹和实验报告PDF文件。
  - code文件夹：存放实验代码；
  - PDF文件格式参考发的模板。
- 如果需要更新提交的版本，则在后面加_v2，_v3。如第一版是“学号_姓名_作业编号.zip”，第二版是“学号_姓名_作业编号_v2.zip”，依此类推。
- 截至日期：**2025年4月28日晚24点**。
- 提交邮箱：zhangyc8@mail2.sysu.edu.cn。