

## 作业 4

1. (15 分) 基于第七章例 7.1 的数据, 考虑使用信贷情况作为构建决策树的第一层内部节点。由于信贷情况有一般、好、非常好三个取值, 为构建二叉树, 需要确定用信贷情况的哪个值作为划分。以下计算请给出计算细节:
- 将一般作为一个分支, 好与非常好作为另一分支, 计算此判别条件的信息增益。(10 分)
  - 将一般与好作为一个分支, 非常好作为另一分支, 计算此判别条件的信息增益, 并与(1)中的结果对比, 哪个更好? (5 分)

ID	年龄	有工作	有房	信贷情况	贷款	ID	年龄	有工作	有房	信贷情况	贷款
1	青年	否	否	一般	否	9	中年	否	是	非常好	是
2	青年	否	否	好	否	10	中年	否	是	非常好	是
3	青年	是	否	好	是	11	老年	否	是	非常好	是
4	青年	是	是	一般	是	12	老年	否	是	好	是
5	青年	否	否	一般	否	13	老年	是	否	好	是
6	中年	否	否	一般	否	14	老年	是	否	非常好	是
7	中年	否	否	好	否	15	老年	否	否	一般	否
8	中年	是	是	好	是						

## (1) 信贷情况

一般 ↘ 好 | 非常好  
 5人 10人 8人 贷  
 1人 贷 2人 不贷  
 4人 不贷

设  $X_1$ : 信贷好 / 非常好,  $X_2$ : 信贷一般,  $Y$ : 贷款

$$H(Y) = -\frac{9}{15} \log_2 \frac{9}{15} - \frac{6}{15} \log_2 \frac{6}{15} = 0.971$$

$$H(Y|X_1) = -\frac{2}{10} \log_2 \frac{2}{10} - \frac{8}{10} \log_2 \frac{8}{10} = 0.722$$

$$H(Y|X_2) = -\frac{1}{5} \log_2 \frac{1}{5} - \frac{4}{5} \log_2 \frac{4}{5} = 0.722$$

$$\therefore H(Y|\text{信贷}) = \frac{1}{3} H(Y|X_2) + \frac{2}{3} H(Y|X_1) = 0.722$$

$$g(Y|\text{信贷}) = H(Y) - H(Y|\text{信贷}) = 0.249$$

## (2) 信贷情况

一般 / 好  $X_2$  ↘ 非常好  $X_1$   
 11人 4人 4人 贷  
 5人 贷 6人 不贷

$$H(Y|X_1) = 0 \quad H(Y|X_2) = -\frac{5}{11} \log_2 \frac{5}{11} - \frac{6}{11} \log_2 \frac{6}{11} = 0.994$$

$$\therefore H(Y|\text{信贷}) = \frac{4}{15} H(Y|X_1) + \frac{11}{15} H(Y|X_2) = 0.729$$

$$\therefore g'(Y|\text{信贷}) = H(Y) - H(Y|\text{信贷}) = 0.242 < g(Y|\text{信贷})$$

∴ (2) 中更好。

2. (10 分) 在第七章 PPT 的 Page 9 中提到了随机变量的熵的特点, 请证明(给出证明步骤)

(1) 当随机变量有  $K$  个取值时, 这  $K$  个值是等概率时, 熵最大; (5 分)

(2)  $K$  越大时, (1) 中得到的熵的最大值越大. (5 分)

证明: (1)  $H(Y) = -\sum_{i=1}^K p_i \log_2 p_i = \log_2 \prod_{i=1}^K \left(\frac{1}{p_i}\right)^{p_i} \leq \log_2 \left(\frac{\sum_{i=1}^K p_i \frac{1}{p_i}}{\sum_{i=1}^K p_i}\right)^{\sum_{i=1}^K p_i} = \log_2 K$

其中  $p_1 = \dots = p_K = \frac{1}{K}$  时取得最大值为  $H(Y) = \log_2 K$

(2)  $\frac{dH(Y)}{dk} = \frac{1}{k \ln 2} > 0 \quad \therefore K$  越大, 熵最大值越大.  $\square$

3. (10 分) 在第七章 7.1 节中我们提到了集成学习常用的方法有 Bagging 和 Boosting, 而我们只介绍了 Bagging 的思想. 请你自学 Boosting 方法, 并制作 1-2 页 PPT 介绍 Boosting 的基本思想和优缺点。

### Algorithm

#### Boosting 算法 训练弱分类器 $\rightarrow$ 加权组合成强分类器

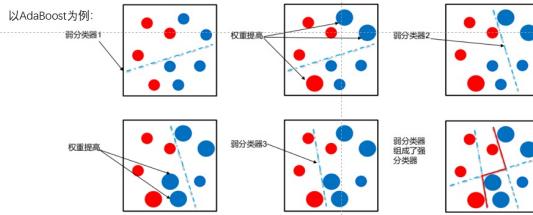
算法思想:

步骤1: 所有分布下的基础学习器对于每个观测值都应该有相同的权重

步骤2: 如果第一个基础的学习算法预测错误, 则该点在下一次的基础学习算法中有更高的权重

步骤3: 迭代第2步, 直到到达预定的学习器数量或预定的预测精度。

以 AdaBoost 为例:



### Algorithm

#### 优点

① 提高分类精度

② 处理复杂数据

③ 灵活性

④ 可解释性

#### 缺点

训练次数太多  
考虑到了每个偏差数据!

4. (10 分) 除第七章 7.2 节中提到的 Geoffrey Hinton, Yann LeCun, Yoshua Bengio, 李飞飞外, 请你自行上网了解, 另外选择两个在 21 世纪后对神经网络发展具有重大贡献的科学家, 并分别用两到三句话详细描述他们的主要贡献。

Ian Goodfellow: 他提出了生成对抗网络 (GAN) 这一重要的深度学习框架, 该框架可以用来生成具有逼真度的图像和其他类型的数据。他的贡献使得深度学习在生成数据、对抗性学习和无监督学习方面取得了重大进展, 成为了深度学习领域的重要技术之一。

Andrej Karpathy: 作为计算机视觉和深度学习领域的专家, 他在图像识别和自然语言处理方面做出了突出贡献。他的工作涵盖了深度学习模型的训练和应用, 尤其是在循环神经网络 (RNN) 和卷积神经网络 (CNN) 的研究中取得了显著进展, 推动了深度学习技术在实际应用中的发展。

5. (10 分) 神经网络模型的一个重要组成部分是激活函数(activation function), 请你自行上网了解, 选择两个常用的激活函数, 并分别用两到三句话详细介绍它们各自的优缺点和适用场景。

ReLU (Rectified Linear Unit) : 公式为  $f(x) = \max(0, x)$ 。

优点: 具有简单的形式, 计算速度快, 在深度神经网络中表现良好。可以有效地解决梯度消失问题, 并且在训练过程中加速了收敛速度。

缺点: 当输入为负数时, 导数为0, 可能导致神经元死亡。

适用场景: 图像处理、自然语言处理和各种深度学习任务。

Tanh 函数: 公式为  $f(x) = (\exp(x) - \exp(-x)) / (\exp(x) + \exp(-x))$ , 将输入映射到 -1 到 1 之间。

优点: 输出范围广, 并且具有零中心化的特点, 使得其在训练过程中的收敛速度更快。

缺点: 存在梯度消失的问题, 尤其是在深度神经网络中。

适用场景: 循环神经网络 (RNN) 和长短期记忆网络 (LSTM) 等需要零中心化和输出范围广的情况。

6. (10 分) 假设一个卷积神经网络的输入图片是一个  $227 \times 227$  像素矩阵, 经过一个  $11 \times 11$  卷积核矩阵的卷积操作(卷积步长为 4, 无边距扩展), 得到的特征矩阵的维度是多少? 请给出计算细节。

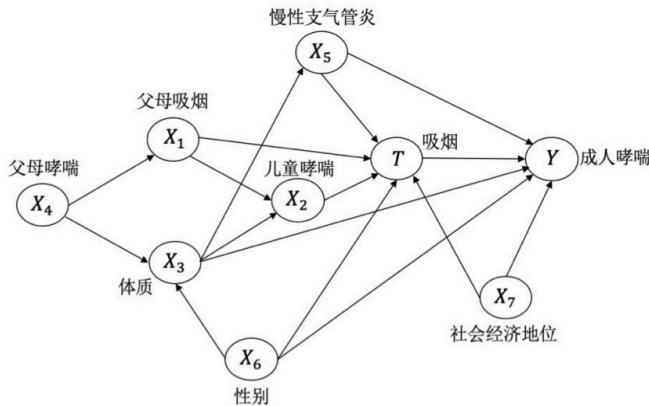
朝一个方向可移动的步数为  $227 - 11 = 216$ , 步长为 4.

故特征矩阵一边步数为  $\frac{216}{4} + 1 = 55$ , 即维度为  $55 \times 55$ .

7. (15分) OpenAI于北京时间2024年5月14日凌晨举办了春季发布会，发布会最大的亮点是新旗舰多模态模型GPT-4o，可接受文本、音频、图像、视频的任意组合作为输入。请你点击[这里](#)观看完整的发布会，并发挥你的想象能力，设想一个未来GPT-4o的应用场景，并回答是该模型的什么新特点/功能使得这样的应用场景成为可能。示例回答：未来可以使用GPT-4o模型进行体育/电竞比赛的实时解说。这是因为GPT-4o中新增了视频、音频等输入和输出功能，它解决了以往语音助手需要2-3秒延时的问题，能够实时地根据视频输入进行解说，并且与以往语音助手不同，GPT-4o的语音输出可带有强烈的情感，是体育/电竞比赛解说中非常需要的特点。

应用场景可以是使用GPT-4o模型进行创意和艺术创作的辅助。这是因为GPT-4o模型在自然语言生成方面具有更高的创造性和想象力，能够生成富有表现力和情感的文本。同时，GPT-4o还具备了对图像、音频以及视频等多种媒体的理解和生成能力，可以与其他艺术创作工具结合，例如图像、音乐生成模型等，共同创作出更加丰富多彩的艺术作品。这使得艺术家和创作者可以借助GPT-4o拓展创作思路，探索更加创新的艺术表现形式，推动艺术领域的发展和创新。

8. (20分) 在一项研究中，研究人员感兴趣的是吸烟对成人哮喘的影响。变量 $T$ 表示某人的吸烟行为， $Y$ 表示此人是否为成人哮喘患者， $X_1$ 表示此人父母的吸烟行为， $X_2$ 表示此人儿童时期是否患有哮喘， $X_3$ 表示此人的体质(无法被观测到的潜在变量)， $X_4$ 表示此人的父母是否患有哮喘， $X_5$ 表示此人是否患有慢性支气管炎， $X_6$ 表示此人的性别， $X_7$ 表示此人的社会经济地位。研究人员构建了如下的因果图：



- (1) 请列出所有从 $T$ 到 $Y$ 的路径，并标明每条路径是打开还是关闭状态，以及是否为因果关联路径，以如下表格形式作答，表中已给出其中两条路径的结果。(提示： $T$ 到 $Y$ 的总路径条数=20。) (10分)

路径编号	路径	状态	是否为因果关联路径
1	$T \rightarrow Y$	打开	是
2	$T \leftarrow X_1 \rightarrow X_2 \leftarrow X_3 \rightarrow Y$	关闭	否

(2) 研究目标是基于观察性研究收集的数据得到  $T$  对  $Y$  (吸烟行为对成人哮喘) 的因果关系，那么在建立  $Y$  关于  $T$  的回归模型中应该加入哪些变量作为协变量？注意： $X_3$  表示人的体质，是无法被观测到的潜在变量，不能被加入模型。请给出你判断每个变量是否应该被加入模型的依据。(提示：加入的变量要能关闭所有打开状态的非因果关联路径，并保证所有关闭状态的非因果关联路径保持关闭。) (10 分)

编号	路径	状态	是否为因果关联路径
1	$T \rightarrow Y$	打开	是
2	$T \leftarrow X_1 \rightarrow X_2 \leftarrow X_3 \rightarrow Y$	关闭	否
3	$T \leftarrow X_1 \rightarrow X_2 \leftarrow X_3 \rightarrow X_5 \rightarrow Y$	关闭	否
4	$T \leftarrow X_1 \rightarrow X_2 \leftarrow X_3 \leftarrow X_6 \rightarrow Y$	关闭	否
5	$T \leftarrow X_1 \leftarrow X_4 \rightarrow X_3 \rightarrow Y$	打开	否
6	$T \leftarrow X_1 \leftarrow X_4 \rightarrow X_3 \rightarrow X_5 \rightarrow Y$	打开	否
7	$T \leftarrow X_1 \leftarrow X_4 \rightarrow X_3 \leftarrow X_6 \rightarrow Y$	关闭	否
8	$T \leftarrow X_2 \leftarrow X_3 \rightarrow Y$	打开	否
9	$T \leftarrow X_2 \leftarrow X_3 \rightarrow X_5 \rightarrow Y$	打开	否
10	$T \leftarrow X_2 \leftarrow X_3 \leftarrow X_6 \rightarrow Y$	打开	否
11	$T \leftarrow X_2 \leftarrow X_1 \leftarrow X_4 \rightarrow X_3 \rightarrow Y$	打开	否
12	$T \leftarrow X_2 \leftarrow X_1 \leftarrow X_4 \rightarrow X_3 \rightarrow X_5 \rightarrow Y$	打开	否
13	$T \leftarrow X_2 \leftarrow X_1 \leftarrow X_4 \rightarrow X_3 \leftarrow X_6 \rightarrow Y$	关闭	否
14	$T \leftarrow X_5 \rightarrow Y$	打开	是
15	$T \leftarrow X_5 \leftarrow X_3 \rightarrow Y$	打开	否
16	$T \leftarrow X_5 \rightarrow X_6 \rightarrow Y$	打开	否
17	$T \leftarrow X_6 \rightarrow Y$	打开	是
18	$T \leftarrow X_6 \rightarrow X_3 \rightarrow Y$	打开	否
19	$T \leftarrow X_6 \rightarrow X_3 \rightarrow X_5 \rightarrow Y$	打开	否
20	$T \leftarrow X_7 \rightarrow Y$	打开	是

(2) 由路径 2、3、4 可知， $X_2$  为叶结点，若不加入模型，则路径 8 无法关闭，故  $X_2$  应加入模型。为了路径 2 保持关闭，则  $X_1$  也应加入模型。

再对比路径 6、7 和 12、13 可知， $X_5$  应加入模型，而  $X_6$  不应加入，再由 5、6、7， $X_4$  不应加入。综上， $X_1, X_2, X_5$  应加入模型。