STA101: 魅力统计 Fascinating Statistics

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**:

   (1) 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)

   (2) 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)

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)

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。

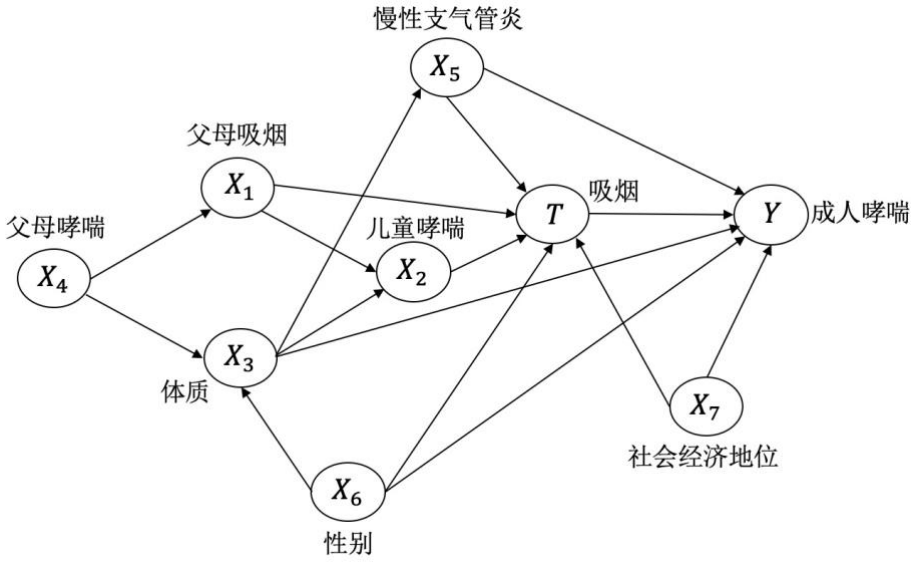
4. (10 分) 除第七章 7.2 节中提到的 Geoffrey Hinton, Yann LeCun, Yoshua Bengio, 李飞飞外, 请你自行上网了解, 另外选择两个在 21 世纪后对神经网络发展具有重大贡献的科学家, 并分别用两到三句话详细描述他们的主要贡献。

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。

6. (10 分) 假设一个卷积神经网络的输入图片是一个227×227像素矩阵, 经过一个11×11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**。

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。

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)