

人工智能基础课后习题答案

第一章：初始人工智能

一、判断题

1. 扫地机器人、无人机和自动驾驶都是人工智能的应用。

正确 (√)

解析：这些设备都集成了人工智能技术。例如，扫地机器人使用路径规划算法（属于问题求解）和传感器数据处理；无人机涉及计算机视觉和自主飞行控制；自动驾驶则综合了计算机视觉、传感器融合、决策制定等多种复杂的AI技术。它们都是AI在现实世界中的具体应用。
2. 1956年的达特茅斯会议标志着人工智能的诞生，但并没有确立人工智能作为一个正式的研究领域。

错误 (×)

解析：这句话前半部分正确，后半部分错误。1956年的达特茅斯会议正是标志着人工智能作为一个正式研究领域诞生的事件。在这次会议上，“人工智能”这个术语被首次提出，并汇集了该领域的早期先驱，确立了其研究目标和方向。
3. 符号主义认为计算机就是一个可以用来模拟人的智能行为的物理符号系统。

正确 (√)

解析：这是对符号主义核心思想的准确描述。符号主义学派认为，人类的认知和思维过程本质上是对符号的运算和处理，因此，只要计算机这个物理系统能够操作和处理符号，就可以模拟甚至实现人类的智能。
4. 人工智能发展两次进入低谷期的共同原因是机器算力限制，政府削减人工智能投资。

正确 (√)

解析：这基本概括了两次“AI冬天”的主要原因。第一次低谷期是因为早期的算法（如机器翻译）无法处理更复杂的问题，技术瓶颈加上承诺无法兑现导致资金削减。第二次低谷期也与技术未达预期和政府及风险投资的资金撤出有关。算力限制和资金支持减弱是贯穿两次低谷期的共性问题。
5. 第三次浪潮开启的标志性事件是辛顿提出深度学习神经网络。

正确 (√)

解析：2006年，杰弗里·辛顿（Geoffrey Hinton）及其学生提出的深度信念网络和“逐层初始化”的训练方法，有效解决了深度神经网络的训练难题，这被广泛认为是开启人工智能第三次浪潮，即深度学习时代的关键性事件。
6. 人工智能的幸存者偏差总是存在的。

正确 (√)

解析：幸存者偏差是指我们过度关注“幸存”下来的样本（如成功案例、返回的数据），而忽略了那些“未幸存”的样本（如失败案例、未返回的数据），从而得出片面结论。由于AI模型依赖数据进行训练，而训练数据几乎不可能完美地覆盖所有情况，因此总会存在某种程度的采样偏差，幸存者偏差是其中一种，它是一个需要持续警惕和缓解的问题。
7. 人工智能的可解释性无法做到。

错误 (×)

解析：虽然提高可解释性是AI领域的一大挑战（特别是对于深度学习的“黑箱”问题），但这并不意味着“无法做到”。目前已经有很多研究和技术致力于提升AI的可解释性（XAI, Explainable AI），例如通过可视化特征、简化模型、输出决策依据等方法。所以说“无法做到”过于绝对。
8. 人工智能只能在虚拟环境中应用，无法对现实世界的物理系统产生影响。

错误 (×)

解析：这是一个明显的错误。第一题中提到的扫地机器人、无人机、自动驾驶汽车，以及工业机器人、智能家居设备等，都是直接在物理世界中运行并产生影响的AI应用。
9. 人工智能技术主要应用于消费领域，而在工业和科学研究中的应用还非常有限。

错误 (×)

解析：人工智能技术已经广泛应用于工业（如智能制造、质量检测）、科学研究（如新药发现、材料科学、天体物理数据分析）、金融、医疗等多个领域，并且正在发挥越来越重要的作用。
10. 由于算法的复杂性和透明性问题，人工智能的“黑箱”问题使得其决策过程难以解释和理解。

正确 (√)

解析：这准确地描述了AI的“黑箱”问题。特别是深度学习神经网络，其内部包含数以亿计的参数，它们之间复杂的相互作用使得我们很难直观地理解模型为什么会做出某个特定的决策。这个问题在金融、医疗等高风险领域尤其受到关注。

二、选择题

1. 以下哪个不属于人工智能应用？

A. 计算器 B. 个性化推荐 C. 扫地机器人 D. 人脸识别

答案：A. 计算器

解析：计算器执行的是固定的、由人类预先编程好的运算法则，它不具备学习或自主决策的能力。而个性化推荐、扫地机器人和人脸识别都需要通过学习数据来构建模型，属于典型的人工智能应用。
2. 哪个学派的目标是模拟人脑的神经网络结构？

A. 符号主义学派 B. 联结主义学派 C. 行为主义学派 D. 以上都不是

答案：B. 联结主义学派

解析：联结主义学派，又称仿生学派，其核心思想就是通过模拟生物大脑中神经元相互连接的方式来构建智能系统，这直接导向了人工神经网络和深度学习的发展。
3. 图灵测试的主要目的是？

A. 测试计算机的运算速度 B. 评估计算机的人工智能水平 C. 测量计算机的存储容量 D. 验证计算机的编程语言

答案：B. 评估计算机的人工智能水平

解析：图灵测试旨在通过对话来判断机器是否能表现出与人类无法区分的智能行为，是衡量机器“智能水平”的一个经典思想实验。

4. 下列哪一项不属于人工智能的核心研究领域？

A. 自然语言处理 B. 计算机视觉 C. 机器人学 D. 化学反应

答案：D. 化学反应

解析：自然语言处理（NLP）、计算机视觉（CV）和机器人学（Robotics）都是人工智能的核心研究分支。而化学反应本身是化学的研究对象，虽然AI可以被用来分析和预测化学反应，但它不是AI的核心研究领域本身。

5. 在人工智能系统的开发和使用过程中，以下哪一项最能确保伦理安全？

A. 只关注技术的准确性和效率，而忽略用户的隐私和数据保护。
B. 在开发过程中不进行伦理审查，以加快项目进度。
C. 建立透明的算法决策过程，并确保用户能够访问和理解数据使用的方式。
D. 仅依赖自动化系统进行决策，不考虑人类的伦理判断。

答案：C. 建立透明的算法决策过程，并确保用户能够访问和理解数据使用的方式。

解析：确保AI伦理安全的关键在于透明度、可解释性和人类监督。选项C是唯一积极正向的措施。选项A、B、D都描述了不负责任且会带来巨大伦理风险的做法。

三、简答题

1. 简述什么是人工智能。

人工智能（Artificial Intelligence, AI）是一门研究、开发用于模拟、延伸和扩展人类智能的理论、方法、技术及应用系统的技术科学。它的目标是了解智能的本质，并创造出一种能以类似于人类智能的方式进行感知、理解、推理、学习和决策的智能机器。

2. 简述人工智能的三大学派及基本区别。

- 符号主义 (Symbolism)

核心思想：认为智能的核心是基于逻辑规则的符号操作。

实现方式：构建知识库和推理机，通过逻辑运算来解决问题。

- 联结主义 (Connectionism)

核心思想：认为智能来自于大量简单的神经元相互连接而形成的复杂网络。

实现方式：通过构建人工神经网络，让模型从数据中自动学习，是当前深度学习的基础。

- 行为主义 (Behaviorism)

核心思想：关注智能体与环境的互动，认为智能是通过学习一系列行为来最大化奖励。

实现方式：以强化学习为代表，智能体在与环境的试错中学习最优策略。

其他简答题较为开放，这里不再一一赘述，重点是理解并能阐述AI的社会影响和伦理问题，如数据隐私、算法偏见、就业冲击等。

第二章：人工智能系统数据基础

一、判断题

1. 人工智能系统是在计算机之上模拟人类智能的技术系统。

正确 (√)

解析：这是人工智能系统的一个基本定义。它的载体是计算机硬件系统，目标是通过算法和数据来模拟、延伸和扩展人类的智能。

2. 面向特定的人工智能应用领域，智慧应用和智能应用从功能设计和特点是有所不同的，比如智慧家居与智能家居。

正确 (√)

解析：通常认为“智能（Smart）”更侧重于自动化和对预设指令的响应，如能按时开关的“智能”插座；而“智慧（Wise/Intelligent）”更强调系统的学习、理解、适应和主动决策能力，如能学习你的生活习惯并主动调节环境的“智慧”家居系统。两者在功能设计上确实有所不同。

3. 计算机在实现智能时依赖于逻辑推理规则，而这种规则是由开发人员事先明确定义的，因此无法适应新的未知问题。

错误 (×)

解析：这描述了人工智能的符号主义学派。现代人工智能，特别是基于联结主义的机器学习和深度学习，其核心思想是让计算机从数据中自动学习规律，而不是完全依赖于开发人员预先定义的规则。这使得AI系统能够处理和适应新的未知问题。

4. 计算机中并不是所有的数据都采用二进制表示。

错误 (×)

解析：在计算机的最底层，所有的数据——无论是数字、字符、图像还是声音——都必须被转换成二进制（0和1）的形式才能被处理和存储。这是现代计算机工作的基本原理。

5. 计算机的处理器、存储器和输入输出系统通过系统总线进行连接。

正确 (√)

解析：这是冯·诺依曼体系结构中的一个核心概念。处理器（CPU）、存储器和输入/输出（I/O）这三个核心子系统是通过一组称为总线（Bus）的导线来相互连接和传输信息的。

6. 存储单位的1KB就是1000个字节。

错误 (×)

解析：在计算机科学和操作系统领域，存储单位的换算是基于2的幂。1 KB (Kilobyte) = 1024 Bytes (2^{10} Bytes)。在某些领域，如硬盘制造商的宣传或网络传输速率中，可能会使用1000作为换算单位，但在计算存储容量时，标准是1024。

二、选择题

1. 下面哪一项不属于人工智能系统核心的3个要素？
A. 算力 B. 数据 C. 算法 D. 模型
答案：D. 模型
解析：人工智能的三大核心要素是算力、数据、算法。“模型”是运用“算法”在“数据”上进行训练后得到的产物或结果，而不是构成人工智能系统的基础要素之一。
2. 下列哪种数据类型通常用于表示图像中的像素？
A. 整数 B. 浮点数 C. 字符串 D. 布尔值
答案：A. 整数
解析：一幅彩色图像由像素点组成，每个像素的颜色由RGB（红、绿、蓝）三个通道的强度值表示，每个强度值的取值范围是0到255，这些值都是整数。
3. 一块CPU芯片集成了控制器和____
A. 存储器 B. 寄存器 C. 运算器 D. 程序计数器
答案：C. 运算器
解析：根据冯·诺依曼体系结构，中央处理器（CPU）主要由控制器（Controller）和运算器（ALU）两大部分组成。
4. 自然语言处理是计算机理解和生成____的技术。
A. 图像 B. 文本 C. 声音 D. 数字信号
答案：B. 文本
解析：自然语言处理（NLP）的核心是让计算机能够处理和理解人类的语言，其主要载体是文本。
5. 以下关于计算机图形数据表示描述正确的是
A. 图形和图像在计算机中的编码方法相同
B. 图形数据在计算机中是以像素点RGB数据信息的二进制来表示
C. 图像是栅格的图形与矢量显示方式一样
D. 视频由连续图像组成的表示方式跟图形一样
答案：B（题干应为“图像”更准确）
*解析：
 - A 错误：图形（Vector Graphics）是用数学公式描述的，图像（Raster Graphics）是由像素点阵描述的，编码方法完全不同。
 - B 正确（若题干为“图像”）：图像数据在计算机中是以像素点RGB信息的二进制来表示。
 - C 错误：图像是栅格（点位图）的，与矢量显示方式完全不同。
 - D 错误：视频是由连续的图像组成的，其表示方式与图形不同。*

三、简答题

1. 简述人工智能系统技术架构。
一个典型的人工智能系统技术架构通常是分层的，主要包含以下四个层次：
 - i. 硬件基础层：提供核心的计算能力（算力），包括处理器（CPU/GPU/TPU）、存储器（RAM/硬盘）等物理设备。
 - ii. 操作系统层：管理和调度硬件资源，为上层软件提供统一的接口，如Linux、Windows等。
 - iii. 智能算法层：实现“智能”的核心，包含了各种机器学习算法、深度学习模型以及实现这些算法的开源框架（如TensorFlow, PyTorch）。
 - iv. 领域应用层：将AI技术应用于具体的行业和场景，解决实际问题，如医疗诊断、金融风控、自动驾驶等。*
2. 简述人工智能的3个要素及其关系。
人工智能的三大核心要素是数据、算法、算力。它们之间的关系可以这样理解：
 - 数据是“燃料”：它是所有机器学习和深度学习模型学习和提取知识的基础。没有高质量、大规模的数据，再好的算法也无法发挥作用。
 - 算法是“引擎”：它定义了模型如何从数据中学习规律和模式。算法的优劣决定了模型智能水平的上限。
 - 算力是“马力”：它提供了执行算法和处理大规模数据所需的计算能力。强大的算力（如GPU、TPU）是训练复杂深度学习模型的保障。三者相辅相成，共同驱动了现代人工智能的发展。
3. 简述冯·诺依曼体系结构组成及工作原理。
组成部分：冯·诺依曼体系结构由五个基本部分组成：
 - i. 运算器：执行算术和逻辑运算。
 - ii. 控制器：指挥和协调计算机各部分的工作。
 - iii. 存储器：存放程序和数据。
 - iv. 输入设备：向计算机输入信息。
 - v. 输出设备：展示计算机处理的结果。其中，运算器和控制器合称为中央处理器（CPU）。
工作原理：其核心是“程序存储”原理。即程序和执行所需的数据都以二进制形式预先存放在存储器中。计算机从存储器中逐条读取指令，并按顺序执行，从而实现程序的自动化运行。
4. 简述图像数据的二进制表示方法。
计算机表示图像数据的基本方法如下：
 - i. 像素化：将一幅连续的图像分割成许多微小方格，每个方格称为一个像素（Pixel）。
 - ii. 颜色编码：彩色图像每个像素的颜色由RGB（红、绿、蓝）三原色的不同强度组合而成。
 - iii. 数字化：每个颜色通道的强度值通常用一个字节（8位二进制数）表示，取值范围是0到255。例如，纯红色可表示为(R=255, G=0, B=0)。
 - iv. 二进制存储：每个像素的RGB值都被转换成相应的二进制码（如(11111111, 00000000, 00000000)）存储在计算机中。对于灰度图像，则只需要一个通道来表示亮度即可。*

第三章：人工智能的应用开发基础

一、判断题

1. Python 是一种静态类型语言。

错误 (×)

解析：Python 是动态类型语言，变量类型在运行时确定，无需显式声明类型。
2. Jupyter Notebook 是基于 Web 的交互式开发环境。

正确 (√)

解析：Jupyter Notebook 是开源 Web 应用，支持实时代码、公式、可视化和文本，适合数据科学与机器学习探索。
3. NumPy 主要用于数值计算，Pandas 主要用于数据分析。

正确 (√)

解析：NumPy 专注高性能数值计算，Pandas 基于 NumPy，专注数据处理与分析。
4. PyTorch 是一个由 Facebook 开发的深度学习框架。

正确 (√)

解析：PyTorch 由 Facebook (Meta) AI 研究院开发。
5. TensorFlow 只能运行在 GPU 上。

错误 (×)

解析：TensorFlow 可运行于 CPU、GPU、移动设备等多平台。
6. Python 中的列表是可修改的，元组是不可修改的。

正确 (√)

解析：列表 (List) 可变，元组 (Tuple) 不可变。
7. 在 Python 中，可以使用单下划线或双下划线来表示私有属性或方法。

正确 (√)

解析：单下划线 _ 表示内部使用，双下划线 __ 触发名称改写，增强私有性。
8. Scikit-learn 是一个用于机器学习的 Python 库。

正确 (√)

解析：Scikit-learn 是主流机器学习库，提供丰富算法和数据预处理工具。
9. PaddlePaddle 是中国首个自主研发的深度学习框架。

正确 (√)

解析：PaddlePaddle (飞桨) 由百度开发，是中国首个自主研发的开源深度学习平台。
10. 在 Python 中，所有标识符都必须以字母开头。

错误 (×)

解析：标识符可由字母、数字、下划线组成，但不能以数字开头，可以以下划线或字母开头。

二、选择题

1. 以下哪个工具不是 Python 的集成开发环境 (IDE) ？

A. PyCharm B. Jupyter Notebook C. Visual Studio Code D. Anaconda

答案：D. Anaconda

解析：Anaconda 是数据科学平台和包管理器，不是 IDE。
2. 在 Python 中，哪个关键字用于定义函数？

A. def B. class C. for D. if

答案：A. def

解析：def 用于定义函数。
3. 以下哪个操作不能用于 Python 的列表？

A. append B. pop C. reverse D. Tob sort

答案：D. Tob sort

解析：append、pop、reverse 是列表方法，Tob sort 并不存在。
4. 以下哪个模块不是 Python 标准库的一部分？

A. math B. re C. NumPy D. random

答案：C. NumPy

解析：NumPy 是第三方库，需单独安装。
5. 在 Python 中，哪个运算符用于求幂运算？

A. ** B. % C. // D. *

答案：A. **

解析：** 表示幂运算。
6. 以下哪个不是 Python 中的基本数据类型？

A. int B. float C. str D. dict

答案：D. dict

解析：int、float、str、bool 是基础类型，dict 属于容器类型。
7. 以下哪个库主要用于数据可视化？

A. NumPy B. Pandas C. Matplotlib D. Scikit-learn

答案：C. Matplotlib

解析：Matplotlib 是主流数据可视化库。

8. 以下哪个不是 Pandas 中的数据结构？

A. DataFrame B. Series C. Array D. Index

答案：C. Array

解析：Pandas 核心结构为 DataFrame、Series、Index，Array 属于 NumPy。

9. 以下哪个不是处理自然语言处理（NLP）的 Python 库？

A. NLTK B. spaCy C. Scikit-learn D. Transformers

答案：C. Scikit-learn

解析：Scikit-learn 是通用机器学习库，非专门 NLP 库。

10. 以下哪个深度学习框架由百度开发？

A. PyTorch B. TensorFlow C. PaddlePaddle D. MXNet

答案：C. PaddlePaddle

解析：PaddlePaddle（飞桨）由百度开发。

三、简答题

1. 简述 Python 中环境管理的重要性及 Anaconda 平台在环境管理中的作用。

环境管理可避免不同项目间的依赖冲突，确保每个项目有独立的依赖空间。Anaconda 的 Conda 工具可轻松创建、切换、管理虚拟环境和库，极大简化环境配置。

2. 解释什么是面向对象编程，并列出其三大特性。

面向对象编程（OOP）是一种将数据和操作数据的方法封装为对象的编程范式。

三大特性：

封装（隐藏内部细节）

继承（代码复用）

多态（同一接口多种实现）。

3. 描述 PyTorch 和 TensorFlow 的主要区别，并给出各自适用的场景。

PyTorch 采用动态计算图，灵活易调试，适合研究和原型开发；TensorFlow 早期为静态图，现支持动态图，生态成熟，适合工业级部署。

4. 简述 Pandas 库的主要功能及其在数据分析中的应用。

Pandas 提供高效的数据读写、清洗、选择、转换、分组聚合等功能，是数据分析流程中的核心工具，常用于数据预处理和探索性分析。

5. 解释什么是深度学习框架，并列举 3 种主流的深度学习框架及其特点。

深度学习框架是简化神经网络开发、训练和部署的软件工具。主流框架有：

PyTorch（灵活易用，适合研究）

TensorFlow（工业级部署生态强）

PaddlePaddle（国产支持好，适合中文NLP和产业应用）。

第四章：从问题求解到机器学习

一、判断题

1. 如果两个水壶的容量互质，那么水壶问题一定可以通过一系列倒水操作来解决。

正确 (✓)

解析：若两个水壶容量互质，则可通过倒水操作量出不超过两壶容量和的任意整数体积的水。

2. 问题求解是人类智能普遍存在的活动，旨在消除当前状态和目标状态的差距。

正确 (✓)

解析：问题求解本质是从初始状态出发，通过一系列操作达到目标状态。

3. 机器学习是人工智能的一个分支，主要目的是让计算机系统能够通过经验或数据进行学习，而不需要显式编程。

正确 (✓)

解析：机器学习强调让机器从数据中自动学习规律，无需人工编写所有规则。

4. 在监督学习中，模型通过使用未标记的数据进行训练，从中寻找规律。

错误 (✗)

解析：监督学习用带标签的数据训练，无监督学习才用未标记数据。

5. 启发式搜索是一种求解优化问题的贪心算法。

错误 (✗)

解析：贪婪最佳优先搜索属于启发式搜索。但启发式搜索和贪心算法是两个概念。

二、选择题

1. 下列关于水壶问题的说法正确的是

A. 水壶问题属于约束满足问题的范畴

B. 水壶问题无法使用启发式搜索算法进行求解

C. 水壶问题必须使用动态规划方法来求解

D. 只要有足够多的水壶，总能得到任意水量

答案：A

解析：水壶问题需满足容量约束，属于约束满足问题。

2. 下列哪一种学习方法属于无监督学习？

A. 线性回归 B. 决策树 C. 支持向量机 D. K 均值聚类

答案：D

解析：K 均值聚类不需要标签，是无监督学习。

3. 以下哪个说法是正确的？

A. 暴力搜索是通过尝试所有可能的解决方案来找到最佳解决方案的搜索方法
B. 穷尽搜索是通过模拟对手的最佳策略选择自己的最佳行动的方法
C. 深度优先搜索的思路是从根节点开始，逐层探索所有节点，先探索完一层再探索下一层
D. 广度优先搜索的思路是通过深入每个分支直到不能再深入为止，然后回溯到上一个节点，继续探索其他分支

答案：A

解析：A 正确，B 是极小极大搜索，C、D描述颠倒。

4. 以下哪个方法是模型训练？

A. predict B. fit C. accuracy_score D. train_test_split

答案：B

解析：.fit() 用于训练模型。

5. 以下有关机器学习的说法不正确的是

A. 机器学习就是训练寻找模型的过程
B. 评估算法通过各种评估指标来衡量模型的性能
C. 机器学习模型就是机器算法
D. 预测是通过模型对新数据进行的，旨在获得预测和分类的结果

答案：C

解析：算法是方法，模型是算法在数据上训练得到的产物。

三、简答题

1. 假设你有一个8升的水壶和一个5升的水壶，且没有任何刻度线。你如何使用这两个水壶准确地量出6升的水？请详细描述每一步操作。

操作步骤如下：

- i. 8L壶装满 (8, 0)
- ii. 8L壶倒入5L壶 (3, 5)
- iii. 倒掉5L壶水 (3, 0)
- iv. 3L倒入5L壶 (0, 3)
- v. 8L壶再装满 (8, 3)
- vi. 8L壶倒入5L壶至满 (6, 5)
- vii. 此时8L壶剩6升水，完成目标。

2. 什么是人类智能的问题求解？请列举典型的算法方法。

人类智能的问题求解是通过算法和规则，从初始状态到目标状态的操作序列设计。常见算法有：

- 贪心法
- 分治法
- 回溯法
- 动态规划法

3. 什么是机器学习？它包括哪几种类型？

机器学习是让计算机通过数据自动学习规律，无需显式编程。主要类型有：

- 有监督学习（分类、回归）
- 无监督学习（聚类、降维）
- 半监督学习
- 强化学习

4. 什么是机器学习的3个要素？简述它们之间的关系。

机器学习的核心三要素是：数据、算法、模型。

它们之间的关系可以这样理解：

- i. 数据 (Data)：
 - 角色：输入/原材料。
 - 描述：这是机器学习的基础，是模型学习知识的来源。
- ii. 算法 (Algorithm)：
 - 角色：方法/过程。
 - 描述：算法是用于“建模”的一套具体方法或流程。它定义了如何从数据中学习规律，并最终构建出模型。
- iii. 模型 (Model)：
 - 角色：输出/产物。
 - 描述：模型是算法作用于数据之后得到的最终结果。它是一个已经学习到特定规律、可以用来进行预测或分类的函数或结构。

三者关系总结：

我们可以把这个过程看作：使用一种算法 (Algorithm)，让它去学习大量的数据 (Data)，最终训练得到一个可用的模型 (Model)。

5. 简述机器学习求解的基本过程。

基本流程：

- i. 数据准备（收集、清洗、预处理、特征提取）

- ii. 模型训练（选择算法，训练模型）
- iii. 模型评估（用测试集评估性能）
- iv. 预测/应用（模型上线，对新数据预测或分类）

第五章：回归与分类模型

一、判断题

- 1. 监督学习是机器学习的一个重要分支，主要解决回归和分类问题。
正确 (√)
解析：监督学习根据预测目标的类型分为回归（连续数值）和分类（离散类别）两大类问题。
- 2. 在线性回归中，模型参数的估计通常采用最大似然法。
正确 (√)
解析：最小二乘法与最大似然估计在正态分布假设下等价，都是常用的参数估计方法。
- 3. 决策树算法不需要对数据进行标准化处理。
正确 (√)
解析：决策树只关心特征的相对顺序和分割点，不受特征尺度影响，无需标准化。
- 4. 交叉熵损失函数只适用于二分类问题，不能用于多分类问题。
错误 (×)
解析：交叉熵有二分类和多分类两种形式，多分类交叉熵适用于多分类任务。

二、选择题

- 1. 以下哪种方法不属于监督学习？
A. 线性回归 B. 逻辑回归 C. K-means 聚类 D. 决策树
答案：C. K-means 聚类
解析：K-means 聚类是无监督学习方法。
- 2. 在线性回归中，最小二乘法的目标是最小化
A. 预测值与真实值的差
B. 预测值与真实值差的绝对值
C. 预测值与真实值差的平方和
D. 预测值与真实值的对数差
答案：C. 预测值与真实值差的平方和
解析：最小二乘法通过最小化残差的平方和来拟合模型。
- 3. 决策树算法中用于选择最佳分割特征的指标不包括
A. 信息增益 B. 基尼指数 C. 均方误差 D. 信息增益比
答案：C. 均方误差
解析：均方误差主要用于回归树，分类树常用信息增益、基尼指数、信息增益比。

三、简答题

- 1. 简述监督学习中回归和分类问题的区别，并各举一个实际应用的例子。
区别：回归预测连续数值，分类预测离散类别。
 - 回归例子：房价预测（根据特征预测具体售价）。
 - 分类例子：垃圾邮件识别（将邮件分为“垃圾”或“正常”）。
- 2. 描述决策树的基本原理，以及它在垃圾邮件分类中的应用。
基本原理：决策树通过对特征进行逐层判断，将数据划分为不同类别。每个内部节点是特征测试，分支是测试结果，叶节点是类别标签。
应用：垃圾邮件分类时，决策树可根据邮件内容、发件人、超链接数量等特征，逐步判断并最终归类为“垃圾邮件”或“非垃圾邮件”。

第六章：数据的聚类 and 降维问题

一、判断题

- 1. 无监督学习不需要带有标签的训练数据即可进行学习。
正确 (√)
解析：无监督学习直接从无标签数据中学习，发现数据结构和模式。
- 2. K-means 算法的聚类结果与初始中心点的选择无关。
错误 (×)
解析：K-means 聚类结果高度依赖初始中心点，可能导致不同的局部最优解。
- 3. 在使用 K-means 算法时，必须事先指定簇的数量 K。
正确 (√)
解析：K 是用户预先指定的超参数，算法无法自动确定。
- 4. 主成分分析算法（PCA）只能用于二维数据的降维。
错误 (×)

解析：PCA 可用于任意高维数据的降维，不限于二维。

5. 无监督学习的评估通常比监督学习更容易。

错误 (✖)

解析：无监督学习因无标签，评估更困难，常用间接指标或人工分析。

6. K-means 算法只能处理数值型特征。

正确 (✓)

解析：K-means 依赖欧氏距离和均值计算，只能直接处理数值型数据。

7. PCA 降维后的主成分之间是正交的。

正确 (✓)

解析：PCA 的主成分在数学上是正交的，彼此独立。

8. 聚类分析的目标是使簇内相似度最大，簇间相似度最小。

正确 (✓)

解析：聚类目标是簇内数据点相似，簇间数据点差异大。

9. 降维技术一定会导致数据信息的损失。

正确 (✓)

解析：降维本质是丢弃部分信息，通常损失较少且保留主要特征。

10. K-means 算法一定能找到全局最优解。

错误 (✖)

解析：K-means 是贪心算法，通常只能收敛到局部最优解，结果受初始点影响。

二、选择题

1. K-means 算法中的 K 值代表

A. 迭代次数 B. 特征数量 C. 簇的数量 D. 样本数量

答案：C. 簇的数量

解析：K 表示要划分的簇数，是需预设的参数。

2. 以下哪种距离度量方法最常用于 K-means 算法？

A. 曼哈顿距离 B. 欧氏距离 C. 切比雪夫距离 D. 马氏距离

答案：B. 欧氏距离

解析：K-means 默认使用欧氏距离衡量相似度。

3. PCA 降维的核心思想是

A. 随机投影 B. 寻找数据方差最大的方向 C. 最小化重构误差 D. 最大化类间距离

答案：B. 寻找数据方差最大的方向

解析：PCA 通过最大化方差方向选取主成分。

4. 下列哪项不是降维的主要目的

A. 数据可视化 B. 去除冗余 C. 增加特征数量 D. 降低计算复杂度

答案：C. 增加特征数量

解析：降维是减少特征数量，非增加。

5. 在鸢尾花数据集中，以下哪个不是原始特征？

A. 花萼长度 B. 花瓣宽度 C. 花朵重量 D. 花萼宽度

答案：C. 花朵重量

解析：鸢尾花数据集原始特征不包括花朵重量。

三、简答题

1. 请解释无监督学习与监督学习的主要区别。

监督学习用带标签数据训练，目标是学习输入到输出的映射（如分类、回归）；无监督学习用无标签数据训练，目标是发现数据结构和模式（如聚类、降维）。

2. K-means 算法的基本步骤是什么？它可能存在哪些局限性？

基本步骤：

i. 随机选 k 个初始质心

ii. 分配每个点到最近质心

iii. 更新每个簇的质心

iv. 重复分配和更新直到收敛

局限性：K 值需预设，对初始质心敏感，易陷入局部最优，对非球形簇和异常值不敏感。

3. 为什么需要降维？PCA 算法的基本原理是什么？

降维可用于数据可视化、降低计算复杂度、去除冗余和噪声。PCA 通过寻找数据方差最大的方向，将数据投影到这些主成分上，实现信息尽量保留的降维。

4. 在实际应用中，如何选择合适的聚类数量 K？

常用方法有肘部法则 (Elbow Method)、轮廓系数 (Silhouette Score)、结合业务知识等。

5. 请举例说明无监督学习在实际中的应用场景。

如客户细分（聚类）、金融异常检测、文本主题建模等。

第七章：深度网络基础组件

一、判断题

1. 梯度下降法可以找到函数的全局最优点。

错误 (✖)

解析：梯度下降法在非凸函数中通常只能找到局部最优点，不能保证全局最优。
2. 根据“没有免费的午餐”定律，没有一种算法在所有情况下都是最优的。

正确 (✓)

解析：“没有免费的午餐”定律说明不存在适用于所有问题的万能算法。
3. AdaGrad 算法的一个主要缺点是学习率会趋向于0，导致梯度消失。

正确 (✓)

解析：AdaGrad 累积梯度平方，学习率会不断减小，最终可能过小，训练提前停止。
4. 梯度下降法常用于人工神经网络、逻辑回归、决策树等机器学习中。

错误 (✖)

解析：梯度下降法适用于可导模型（如神经网络、逻辑回归），决策树不依赖梯度下降。
5. 感知机模型是对生物神经元功能的直接模拟。

错误 (✖)

解析：感知机模型灵感来源于生物神经元，模拟其加权求和与激活过程。并非直接。

二、选择题

1. 神经元细胞的哪种现象被称为“全或无”现象？

A. 只有当外来刺激有足够大的强度时，才能产生动作电位，且动作电位的幅度和传播范围恒定

B. 动作电位只能在神经元胞体内产生，且不能在突起部分传播

C. 动作电位的产生与刺激强度无关，但传播范围与刺激强度成正比

D. 神经元的兴奋只能持续一段时间，然后自动恢复到静息状态

答案：A

解析：刺激需达到阈值才能触发动作电位，且幅度恒定。
2. 下列说法正确的有

A. 感知机模型可以解决二分类问题

B. 感知机模型可以解决线性回归问题

C. 感知机模型可以解决异或问题

D. 感知机模型可以解决多分类问题

答案：A, B, D

解析：单层感知机可解决线性二分类和回归，多层感知机可多分类，单层无法解决异或问题。
3. 在全连接 MLP 模型中，神经元的输出是如何传输到相邻层的神经元的？

A. 通过联结强度权因子的加强或抑制

B. 直接复制

C. 通过随机选择

D. 通过固定比例缩放

答案：A

解析：输出通过权重连接到下一层神经元，权重决定信号强度。
4. 在实际应用中，对输入层的数据进行哪种处理可以使其落入0到1区间？

A. 标准化 B. 归一化 C. 离散化 D. 编码化

答案：B. 归一化

解析：归一化（Min-Max Scaling）将数据缩放到[0,1]区间。
5. 激活函数的主要作用是什么？

A. 增加神经网络的层数

B. 通过非线性变换，使神经网络具备非线性特性

C. 提高神经网络的计算速度

D. 减少神经网络的参数数量

答案：B

解析：激活函数赋予网络非线性能力，能学习复杂模式。
6. 随机梯度下降法（SGD）的基本思想是什么？

A. 每次训练时使用所有样本的梯度来更新模型参数

B. 每次训练时只使用一个样本的梯度来更新模型参数

C. 每次训练时使用一半样本的梯度来更新模型参数

D. 每次训练时使用部分样本的梯度，但不一定是单个

答案：B

解析：SGD 每次用一个样本的梯度更新参数，提升训练速度。
7. GPU 相较于传统的 CPU，在哪方面有所加强？

A. 串行处理能力

B. 并行处理能力

- C. 单线程处理能力
- D. 逻辑运算能力

答案：B

解析：GPU 设计用于大规模并行计算，适合深度学习。

8. ImageNet 与 AlexNet 的成功说明了什么？

- A. 好算法比好数据重要
- B. 好数据比好算法重要
- C. 算法和数据都不重要
- D. 算法和数据同样重要

答案：D

解析：AlexNet 的成功依赖于创新算法和大规模高质量数据。

三、简答题

1. 简述深度学习三要素及相互之间的关系。

三要素：

数据（基础和燃料）

算法（引擎和方法）

算力（动力保障）。

三者相辅相成，共同驱动人工智能发展。

2. 为了更快更好地训练出一个高质量的人工智能模型，需要对哪些参数进行调节？

常见超参数包括：学习率、批量大小、训练轮次、优化器、网络结构参数、激活函数、正则化参数等。

3. 为什么损失函数、激活函数必须是连续可导的？

神经网络训练依赖梯度下降法，需计算梯度（导数），因此损失函数和激活函数必须连续可导，否则无法有效训练模型。

四、计算题

1. 设计一个MLP模型，预测销售员到手提成（分段计税和分段提成）。

设计思路：

- 输入层：包括销售额及各分段金额等特征
- 隐藏层：1个或多个，10-20个神经元，激活函数用ReLU
- 输出层：1个神经元，线性激活，输出连续金额
- 损失函数：均方误差（MSE）
- 优化器：Adam

2. 用梯度下降法计算 $f(x) = x^2 + 7$ ，以 $x_0 = 1$ ，学习率 $\eta = 0.01$ ，迭代10步后的位置。

- $f'(x) = 2x$
 - 更新公式： $x_{n+1} = x_n - 0.01 \times 2x_n = x_n \times 0.98$
 - $x_{10} = 1 \times (0.98)^{10} \approx 0.817$
- 即10步后点的位置约为 $x \approx 0.817$

第八章：卷积神经网络

一、判断题

1. 在卷积运算中，补齐操作就是为了保证输出特征大小和输入图像一致。

错误 (×)

解析：补齐（Padding）主要用于控制输出特征图尺寸和处理边缘像素。虽然可以通过设置补齐使输出与输入尺寸一致，但这不是唯一目的。

2. “感受野”概念是由大卫·休伯尔提出来的。

正确 (√)

解析：大卫·休伯尔和托斯坦·威泽尔提出了感受野（Receptive Field）概念，对CNN设计有重要启发。

3. 池化操作的原理是用某一图像区域子块的统计信息包含该子块全局信息。

正确 (√)

解析：池化通过提取局部区域的统计特征（如最大值、平均值）代表整个区域，减小数据尺寸。

4. 通常我们将具有输入层、隐含层和输出层的网络称为深度学习网络。

错误 (×)

解析：只有一个隐含层的网络是浅层网络，深度学习网络通常有多个隐含层。

5. 因为手写体的数字识别数据集是28×28的图像数据，所以只能用CNN建模。

错误 (×)

解析：虽然CNN适合图像，但也可将图像展平成向量，用MLP等模型建模。

二、选择题

1. 池化操作有哪几种常用的算法？

- A. 最大池化
- B. 平均池化
- C. 扩大池化
- D. 随机池化

答案：A, B, D

解析：常用池化有最大池化、平均池化、随机池化。

2. 卷积运算的主要目的是什么？

- A. 对数字图像进行压缩
- B. 对数字图像进行放大
- C. 提取数字图像的特征值
- D. 改变数字图像的分辨率

答案：C

解析：卷积运算用于提取图像的局部特征。

3. 灰度图像使用多少个亮度通道来表示？

- A. 3个
- B. 2个
- C. 1个
- D. 4个

答案：C

解析：灰度图像只有一个通道。

4. 在RGB表示法中，每种颜色的强度用一个什么值来表示，其取值范围是多少？

- A. 一个比特,0或1
- B. 一个字节,0~127
- C. 一个字节,0~255
- D. 两个字节,0~65535

答案：C

解析：RGB每个通道用一个字节（0~255）表示强度。

5. CIFAR-10 是包含了10个动物分类的图像识别数据集，要对该数据集建立卷积神经网络进行图像分类识别，输出层网络应采用哪个激活函数？

- A. ReLU
- B. Sigmoid
- C. softmax
- D. Tanh

答案：C. softmax

解析：多分类任务输出层应使用softmax激活函数。

6. 以下哪些框架可以进行MLP的建模与训练？

- A. Sklearn
- B. TensorFlow
- C. PyTorch
- D. PaddlePaddle

答案：A, B, C, D

解析：上述所有框架都支持MLP建模与训练。

三、简答题

1. 画出AlexNet的拓扑结构图，并简述每个网络层的算法。

AlexNet结构：输入层 → 5个卷积层 → 3个全连接层 → 输出层。

- 输入层：接收224x224x3彩色图像
- 卷积层：用不同尺寸卷积核提取特征，激活函数为ReLU
- 池化层：最大池化，下采样特征图
- 全连接层：Flatten后接3层全连接，使用Dropout防止过拟合
- 输出层：softmax输出1000类概率（ImageNet）

2. 简述卷积运算中的3个主要参数的作用。

- 卷积核形状（Kernel Shape）：决定感受野大小，影响特征提取的局部范围
- 步长（Stride）：决定卷积核滑动步幅，影响输出特征图尺寸
- 卷积核数量（Number of Kernels）：决定可提取的特征种类，数量越多特征越丰富

四、计算和编程题

1. （编程题）用Python实现L2范数池化运算，步长为2。

```
import numpy as np

def l2_norm_pooling(input_matrix, pool_size=(2, 2), stride=2):
    m, n = input_matrix.shape
    pool_h, pool_w = pool_size
    output_h = (m - pool_h) // stride + 1
    output_w = (n - pool_w) // stride + 1
    output_matrix = np.zeros((output_h, output_w))
    for i in range(output_h):
        for j in range(output_w):
            start_i = i * stride
            start_j = j * stride
            window = input_matrix[start_i : start_i + pool_h, start_j : start_j + pool_w]
            l2_norm = np.sqrt(np.sum(window ** 2))
            output_matrix[i, j] = l2_norm
    return output_matrix

# 示例数据
input_data = np.array([
    [23, 34, 55, 32, 66, 43],
    [15, 43, 27, 30, 39, 54],
    [33, 67, 47, 28, 36, 49],
    [56, 89, 90, 35, 77, 65],
    [54, 32, 63, 48, 65, 32],
    [63, 67, 38, 43, 29, 54]
])
l2_pooled_output = l2_norm_pooling(input_data, stride=2)
print("L2范数池化结果:\n", l2_pooled_output)
```

2. (计算题) 例8-1中, 若2个骰子的概率分布函数f(m)、g(n)都是泊松分布, 求2个骰子抛出后点数相加之和为7的概率是多少?

解答: 设 $X \sim P(\lambda_1)$, $Y \sim P(\lambda_2)$, 则 $Z = X + Y \sim P(\lambda_1 + \lambda_2)$ 。点数和为7的概率为:

$$P(Z = 7) = \frac{e^{-(\lambda_1 + \lambda_2)} (\lambda_1 + \lambda_2)^7}{7!}$$

需给定 λ 值才能计算具体数值。

3. (计算题) 对图8-11(A)的网格数据进行卷积运算, 卷积核为图8-11(B), 步长为1, padding=(1,1)。

解答思路:

- 先对5x5输入矩阵四周补一圈0, 变为7x7
- 用3x3卷积核在补齐后的矩阵上滑动, 步长为1, 做点积
- 例如左上角第一个输出元素:

◦ 输入窗口:

```
0  0  0
0 20 21
0 102 35
```

◦ 卷积核:

```
1  -1  1
4   0  9
2   6  3
```

◦ 点积计算: $0 \times 1 + 0 \times (-1) + 0 \times 1 + 0 \times 4 + 20 \times 0 + 21 \times 9 + 0 \times 2 + 102 \times 6 + 35 \times 3 = 0 + 0 + 0 + 0 + 0 + 189 + 0 + 612 + 105 = 906$

- 依次滑动卷积核, 得到5x5输出

第九章：循环神经网络

一、判断题

1. 循环神经网络是一种前馈型神经网络, 能将上一时间点的输出作为当前时间点的输入。

错误 (×)

解析: RNN属于反馈神经网络, 传递的是上一时刻的隐藏状态而非输出。

2. LSTM 是一种特殊的RNN, 用于解决经典RNN不能很好处理序列数据中长时间依赖的问题。

正确 (√)

解析: LSTM通过门控机制有效缓解了RNN的长期依赖和梯度消失问题。

3. 所有的RNN 都只能捕获当前位置之前的状态信息。

错误 (×)

解析：双向RNN可同时利用序列的过去和未来信息。

4. RNN 可以应用于机器翻译、股票分析，也能用于图像处理。

正确 (√)

解析：RNN适合序列数据，也可用于图像描述生成等任务。

二、选择题

1. 下列哪个不是循环神经网络(RNN)模型？

A. Deep RNN B. LSTM C. CNN D. GRU

答案：C. CNN

解析：CNN是卷积神经网络，非RNN模型。

2. GRU 相比 LSTM，在结构上做了哪些主要改变？

A. 增加了记忆状态
B. 保留了完整的遗忘门和输入门
C. 去掉了记忆状态并简化了结构
D. 增加了额外的MLP层

答案：C

解析：GRU合并门控，结构更简单，参数更少。

3. RNN 在自然语言处理中首先将自然语言转化成什么？

A. Unicode码 B. 词向量 C. 知识图谱 D. 国标码

答案：B. 词向量

解析：文本需转为词向量才能输入神经网络。

4. LSTM 缺乏并行计算的能力，主要原因是什么？

A. 在序列的处理过程中，计算是顺序进行的
B. 使用了过多的MLP层
C. 超长的序列处理
D. 梯度消失和梯度爆炸

答案：A

解析：RNN/LSTM每步依赖前一步结果，无法并行。

5. 经典 RNN的设计思路是什么？

A. 通过记忆序列数据后面的输出作为参考信息来处理当前的输出
B. 通过记忆序列数据前面的输出作为参考信息来处理当前的输出
C. 不需要记忆序列数据的前后输出
D. 通过记忆序列数据的所有输出作为参考信息

答案：B

解析：经典RNN利用前面信息处理当前输出。

三、简答题

1. 简述前馈神经网络和反馈神经网络的特点和区别。

前馈神经网络：数据单向流动，无环路，适合静态输入（如MLP、CNN）。反馈神经网络：有环路或反馈连接，能记忆时间序列信息（如RNN、Hopfield网络）。主要区别在于是否有反馈连接和能否处理时序数据。

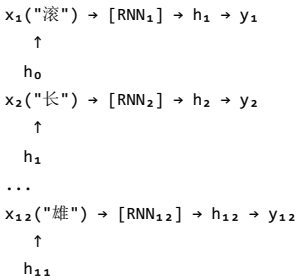
2. 简述 RNN 有哪些结构的分类，各有什么应用场景。

- 一对多（One-to-Many）：如图像描述生成
- 多对一（Many-to-One）：如文本情感分类
- 多对多（Many-to-Many）：如词性标注（同步）、机器翻译（异步）

不同结构适用于不同输入输出序列关系的任务。

3. 画图表示“滚滚长江东逝水，浪花淘尽英雄”按序列展开的简单RNN示意图，分词按字进行。

结构描述：每个字作为一个输入，前一刻的隐藏状态和当前输入一起进入RNN单元，依次传递，形成链式结构。



四、计算与编程题

1. 已知同步多对多RNN，输入层、隐含层、输出层神经元均为1，激活函数ReLU，权重W=[1.2,3.3,-5,2]，H=[1.2]，V=[5]，S₀=0，α=0，β=0，对x=[1,2,3,4],[5,6,7,8],[9,10,11,12]，计算输出序列Y。

- t=1: $W \cdot X_1^T = 0.8$, $S_1 = ReLU(0.8) = 0.8$, $Y_1 = ReLU(5 \times 0.8) = 4.0$

- $t=2: W \cdot X_2^T = 6.8, S_2 = ReLU(6.8 + 1.2 \times 0.8) = 7.76, Y_2 = ReLU(5 \times 7.76) = 38.8$
- $t=3: W \cdot X_3^T = 12.8, S_3 = ReLU(12.8 + 1.2 \times 7.76) = 22.112, Y_3 = ReLU(5 \times 22.112) = 110.56$
- 输出序列Y = [4.0, 38.8, 110.56]

2. (编程题) 用前50天的收盘价和成交量预测后2天的收盘价 (LSTM思路)。

- 数据准备：读取收盘价和成交量，归一化
- 构建序列：输入为50天×2特征，输出为后2天收盘价
- 模型：LSTM输入shape为(50,2)，输出2个神经元，线性激活
- 损失函数：均方误差MSE，优化器Adam
- 训练后预测并反归一化输出

3. (编程题) 用前60天的收盘价和换手率预测后2天股票涨跌 (二分类LSTM思路)。

- 数据准备：读取收盘价和换手率，归一化
- 标签生成：后2天收盘价高于前一天为“涨”(1)，否则为“跌”(0)
- 模型：LSTM输入shape为(60,2)，输出1个神经元，Sigmoid激活
- 损失函数：二元交叉熵，评估指标为准确率
- 训练后预测，输出概率大于0.5为“涨”，否则为“跌”

第十章：完整的人工智能应用开发实践

简答题

1. 简述一个人工智能应用系统项目综合开发应遵循的流程。

一个完整的AI应用开发项目通常包括以下七个关键步骤：

- i. 问题定义与数据收集：明确业务问题，将其转化为机器学习问题，并收集相关原始数据。
- ii. 数据预处理与特征工程：对数据进行清洗、整理、转换（如归一化、标准化），提取有用特征，并划分训练集、验证集和测试集。
- iii. 模型选择：根据任务类型选择合适的机器学习或深度学习模型（如线性回归、决策树、CNN等）。
- iv. 模型训练：用训练集训练模型，通过优化算法（如梯度下降）调整参数以最小化损失函数。
- v. 模型评估：用验证集或测试集评估模型性能，选择合适的评估指标（如准确率、MSE等）。
- vi. 模型调优与部署：根据评估结果调整超参数，优化模型后部署到生产环境。
- vii. 结果解释与持续改进：解释模型输出，持续监控系统表现，并根据新数据迭代优化模型。

2. 结合专业人工智能应用场景，完成一个人工智能应用的综合实践项目，并撰写一份实验报告。

本题为开放性实践题，需结合自身专业领域选择实际问题并完成AI项目开发。

- 选择场景：如医学影像肿瘤分类、金融欺诈检测、土木工程健康监测等。
- 按照上述七步流程执行项目开发。
- 撰写实验报告，详细记录每一步的操作、遇到的问题、决策过程及最终结果和分析，对开发过程进行总结和复盘。

第十一章：自然语言处理中的模型

一、判断题

1. 19世纪70年代，IBM科学家弗雷德里克·贾里尼克团队首次采用了基于统计的方法来解决语音识别的问题。

错误 (×)

解析：应为20世纪70年代。

2. 在序列生成模型中，输入的长度和输出的长度必须是一样的。

错误 (×)

解析：如机器翻译等异步多对多模型，输入输出长度可不同。

3. 本质上讲，词向量可以理解多维空间中的一个点，这个多维空间表示了该词的多个特征值。

正确 (√)

解析：词向量是高维空间的点，每一维代表词的语义特征。

4. 在 Jaccard 相似度计算中，值越大，相似度越小。

错误 (×)

解析：Jaccard 值越大，相似度越高。

5. 在 Transformer 结构中，编码器的输出词向量仅作为第一层解码器输入的词向量参与解码器的运算。

错误 (×)

解析：编码器输出参与解码器每一层的注意力机制。

二、选择题

1. NLP 最初起源于哪方面的需求？

A. 语音识别 B. AI写作 C. 翻译 D. 聊天机器人

答案：C. 翻译

解析：NLP起源于机器翻译需求。

2. NLP 的技术演变经历了哪些阶段？

A. 基于规则的算法 B. 统计语言模型 C. 序列生成模型 D. 预训练-微调大模型

答案：A, B, C, D

解析：四个阶段完整概括了NLP技术发展。

3. 基于规则的 NLP 研究方法陷入困境的主要原因是什么？

- A. 规则过于简单，无法处理复杂的语言现象
- B. 规则很多且复杂，无法涵盖所有语言现象
- C. 语言在不断演变，规则无法跟上变化
- D. 机器翻译的质量已经足够高，无须进一步改进

答案：BC

解析：B.复杂性和完备性,C.动态性和维护性

4. “熟读唐诗三百首，不会作诗也会吟”这句话所体现的是哪个模型的思想？

- A. 基于规则的模型
- B. 统计语言模型
- C. 神经网络模型
- D. 语义分析模型

答案：B. 统计语言模型

解析：通过统计规律建模语言。

5. 以下哪个模型不属于统计生成模型？

- A. Skip-Gram模型
- B. CBOW模型
- C. 词袋模型
- D. 支持向量机模型

答案：D. 支持向量机模型

解析：SVM是判别式分类模型。

6. 在自然语言处理中，“Token”可以是什么？

- A. 一个字
- B. 一个词
- C. 一个短语
- D. 一个句子

答案：A, B, C, D

解析：Token粒度灵活，可为字、词、短语、句子。

7. 词向量表示文本后，哪些方法可以用来计算文本相似度？

- A. 曼哈顿距离
- B. 欧几里得距离
- C. 余弦相似度
- D. Jaccard相似系数

答案：A, B, C, D

解析：这些都是常用的相似度计算方法。

8. Transformer结构中的哪个组件使得模型可以并行处理序列中的每个Token？

- A. 编码器
- B. 解码器
- C. 自注意力机制
- D. 词嵌入

答案：C. 自注意力机制

解析：自注意力机制支持并行计算。

9. BERT系列模型是由哪个公司的AI研究团队提出的？

- A. OpenAI
- B. 微软
- C. 谷歌
- D. 阿里巴巴

答案：C. 谷歌

解析：BERT由Google提出。

10. 以下哪个模型同时采用了编码器和解码器结构？

- A. BERT系列
- B. T5系列
- C. GPT系列
- D. LLaMA系列

答案：B. T5系列

解析：T5是编码器-解码器结构。

三、简答题

1. 简述 Word2Vec 模型中两个不同模型的区别。

- CBOW：用上下文预测中心词，输入为上下文，输出为中心词，训练速度快。
- Skip-Gram：用中心词预测上下文，输入为中心词，输出为上下文，适合小数据集和生僻词。

2. 简述 Transformer 结构的主要特征，并列3种不同架构的Transformer模型。

- 主要特征：自注意力机制、多头注意力、位置编码、编码器-解码器结构。
- 三种架构：仅编码器（如BERT）、仅解码器（如GPT）、编码器-解码器（如T5、BART）。

四、计算编程题

1. 将英文“He likes to read the book about books that I like” 按空格分割后，词汇表大小有多少个Token？ 这个句子有多少个Token？

- 句子有11个Token。
- 词汇表大小也是11（所有词都唯一）。

2. 有以下3个句子，请计算并判断哪两个句子之间的语义最相似？（应用思路题）

- 用预训练模型（如BERT）提取句子向量，计算三组余弦相似度，值最大的一对即为最相似。

3. 在《西游记》这本小说中，与牛魔王关系最密切的8个人物是哪些？（应用思路题）

- 步骤：获取文本→命名实体识别→关系提取（如共现分析）→构建知识图谱→统计与牛魔王共现最多的8个人物。

4. 计算以下两句话的Jaccard相似度：句子1=“天行健君子以自强不息”，句子2=“地势坤君子以厚德载物”。

- 句子1集合A: {天, 行, 健, 君, 子, 以, 自, 强, 不, 息}
- 句子2集合B: {地, 势, 坤, 君, 子, 以, 厚, 德, 载, 物}
- 交集大小=3（君, 子, 以），并集大小=17
- Jaccard相似度=3/17≈0.1765

第十二章：生成式AI与大模型

一、判断题

1. 随着AIGC的高速发展和技术升级，AIGC将完全替代人类在所有领域的工作。

错误 (×)

解析：AIGC会替代部分工作，但难以完全替代需要创造力、情感、复杂决策和物理操作的人类工作，更可能是人机协作。
2. 大语言模型在海量的数据上进行训练，所以不可能有偏见。

错误 (×)

解析：大语言模型会从数据中学习并可能放大原有的偏见。
3. 同一个GAI产品在不同的时间回答同一个问题，答案可能会不一样。

正确 (√)

解析：生成式AI生成内容时带有随机性，结果可能不同。
4. 目前，AI造假大量存在，但是AI本身不知道自己是否在造假。

正确 (√)

解析：AI没有判断真伪的意识，生成内容时可能产生“幻觉”。
5. LLaMA3是目前被业界公认的最好的开源大语言模型，因此可以拿来直接用于医学诊断。

错误 (×)

解析：通用大模型不能直接用于高风险专业领域，需专业微调和验证。

二、选择题

1. AIGC与搜索引擎技术的主要区别是什么？

A. 搜索引擎能获取已经存在的内容，而AIGC不能

B. AIGC能生成新的内容，而搜索引擎只能获取已经存在的内容

C. 搜索引擎能更快找到信息，而AIGC不能

D. AIGC能找到需要的图像，而搜索引擎不能

答案：B

解析：AIGC能生成新内容，搜索引擎只能检索已有内容。
2. LLM模型的特点是什么？

A. 预训练数据无须人工标注

B. 参数规模小

C. 在海量的文本数据上进行训练

D. 参数规模通常达到数十亿到数千亿的级别

答案：A, C, D

解析：LLM参数规模大，预训练用自监督学习，无需人工标注。
3. 训练AGI模型的数据集应该包括哪些内容？

A. 文本数据 B. 语音数据 C. 视频数据 D. 压力传感器数据

答案：A, B, C, D

解析：AGI需多模态数据，涵盖文本、语音、视频、传感器等。
4. 预训练的概念最早是由谁提出来的？

A. 谷歌 B. 辛顿 C. BERT团队 D. OpenAI

答案：B. 辛顿

解析：辛顿2006年提出逐层预训练思想。
5. 在研究AGI的技术突破时，为什么需要同步研究AGI伦理？

A. 确保AGI的设计符合道德规范

B. 提高AGI的智商水平

C. 降低AGI的训练成本

D. 加快AGI的研发速度

答案：A

解析：AGI伦理研究是保障人类社会安全的关键。
6. 当使用大语言模型平台时，用户需要注意什么问题？

A. 模型的大小 B. 模型的运行速度 C. 隐私和数据保护 D. 模型的训练成本

答案：C. 隐私和数据保护

解析：用户输入的数据可能被用于分析或训练，存在隐私风险。
7. GAI已经能在哪些环境中胜任人类的工作？

A. 智能助教 B. 智能教学评估 C. 智能客服 D. 影视创作

答案：A, B, C, D

解析：GAI已在这些领域有实际应用。
8. 在2017年斯坦福大学的对比实验中，哪一组人员成功通过了真假新闻鉴别的测试？

解析：事实核查员表现最好，但没有一组完全“成功通过”。
9. BERT模型采用了哪种方式进行无监督学习？

A. 梯度下降 B. 随机掩码 C. 梯度上升 D. 逐层向后训练

答案：B. 随机掩码

解析：BERT采用掩码语言模型（MLM）进行自监督学习。

三、简答题

- 简述GPT和ChatGPT的区别。
 - GPT是基础大语言模型，是文本生成的“引擎”；ChatGPT是在GPT基础上通过对话数据微调和人类反馈强化学习优化的聊天机器人产品。GPT是底层技术，ChatGPT是面向用户的成品，更擅长多轮对话和指令理解。
- 简述预训练-微调的两种不同的方式。
 - 特征提取模式：冻结预训练模型参数，仅用其输出特征训练下游小模型，速度快，计算开销小。
 - 微调模式：解冻部分或全部参数，在新任务数据上以较小学习率继续训练，性能更好但计算开销大。
- 简述强人工智能和弱人工智能的区别及强人工智能面临的挑战。
 - 弱人工智能（Narrow AI）：专注于特定任务，无法跨领域迁移，目前所有AI均属此类。
 - 强人工智能（AGI）：具备通用智能，能跨领域解决复杂问题，目前尚未实现。
 - 挑战：技术难题（理解、推理、意识）、能源消耗、伦理与安全风险（对齐问题）、意识与哲学问题等。

第十三章：AI前沿技术与应用

一、判断题

- 提示词微调(Prompt Tuning)主要更新的是 embedding 参数。

正确 (✓)

解析：Prompt Tuning只训练输入层新增的软提示向量（embedding参数），其余参数冻结。
- 扩散模型的名字来源于经济学中的市场扩散理论。

错误 (✗)

解析：扩散模型的理论基础来自物理学中的热力学和统计物理。
- 扩散模型的前向过程是对原始图像不断加入噪声生成模糊图像。

正确 (✓)

解析：前向过程是逐步向原始图像添加高斯噪声，最终变为纯噪声图像。
- RLHF是使用强化学习的方式直接优化带有反馈的语言模型，实现了与人类价值观的对齐。

正确 (✓)

解析：RLHF通过人类反馈训练奖励模型，再用强化学习微调语言模型以对齐人类价值观。
- 文生图技术就是通过用户输入的提示词，从图库中搜索出一张最接近的图像给用户。

错误 (✗)

解析：文生图是生成全新图像，不是检索已有图片。

二、选择题

- 关于预训练大语言模型，以下说法正确的是哪些？
 - A. 它们已经具备了所有的专业知识
 - B. 它们经常会出现答非所问的情况
 - C. 它们主要是在公开的语料库中训练，缺乏专业知识训练
 - D. 它们不需要进行微调就能完美适应各种应用场景

答案：B, C
- 将个人数据提供给AI平台进行个性化模型训练会存在哪些问题？
 - A. AI平台无法保证数据的安全性
 - B. 个性化模型训练对个人用户没有实际价值
 - C. 现有的AI平台训练不出符合我们需要的模型
 - D. 个性化模型训练对个人用户来讲费用太昂贵

答案：A, D
- 要发挥私人AI助手的优势，必须满足哪些条件？
 - A. 将大语言模型部署在云端服务器上
 - B. 将大语言模型部署在本地的个人电脑上
 - C. 积累足够多的个人数据集
 - D. 重新设计一个全新的语言模型

答案：B, C
- 大语言模型的局限性体现在以下哪些方面？
 - A. 需要大量新样本才能进行推理
 - B. 对视觉是“盲目”的
 - C. 无法生成没有见过的内容
 - D. 经常出现AI幻觉

答案：B, D
- 图像和视频编码器在多模态大型语言模型中的作用是什么？
 - A. 类似于人类的耳朵，接收和预处理声学信号

- B. 类似于人类的大脑，进行理解和推理
- C. 类似于人类的眼睛，接收和预处理光学信号
- D. 类似于生成器，输出除文本之外的其他模态

答案：C

6. 用扩散模型生成图像，在模型训练和图像生成过程中会有哪些步骤？

- A. 在原始图像中添加随机噪声
- B. 通过判别器来辨别图像生成的效果
- C. 通过训练模型U-Net的参数，满足前向噪声和反向预测噪声的分布残差最小化
- D. 从噪声数据中恢复原始数据分布

答案：A, C, D

7. 以下哪个应用采用的是扩散模型技术来生成图像或视频？

- A. GPT-2
- B. DALL-E3
- C. Stable Diffusion
- D. Sora

答案：B, C, D

8. AARON是哪位艺术家发明的？

- A. 约翰·沃特斯
- B. 杰克逊·波洛克
- C. 哈罗德·科恩
- D. 安德鲁·怀斯

答案：C

9. DALL-E3的技术架构包含哪些组件？

- A. GPT-3 (Transformer)
- B. VAE(自编码器)
- C. CLIP(视觉语言预训练模型)
- D. Diffusion (扩散模型)

答案：B, C, D

10. CLIP模型的作用和目标有哪些？

- A. 将现实世界的实体编码为计算机算法可运算的数据格式
- B. 实现文本与图像的匹配
- C. 生成高质量的图像
- D. 通过训练获得文本编码器和图像编码器的最优参数

答案：A, B, D

三、简答题

1. 简述AI绘画的3个主要发展阶段。

- 基于规则的阶段（单模态）：如AARON，依靠规则和程序控制作画。
- 基于网络的阶段（单模态）：如GAN、DeepDream，从噪声或图像生成新图像。
- 文生图阶段（多模态）：如DALL-E、Stable Diffusion，结合文本提示和扩散模型生成高质量图像。

2. 什么是隐空间？

- 隐空间（Latent Space）是数据的低维抽象表示。编码器将高维复杂实体（如图像、文本）压缩为稠密数值向量，这些向量构成的空间即为隐空间，语义相近的实体在隐空间中距离也近。

3. 列举微调大语言模型的5种方法。

- i. 提示词微调（Prompt Tuning）：冻结模型，仅训练输入的软提示向量。
- ii. 前缀微调（Prefix Tuning）：每层前加可训练前缀向量。
- iii. 适配器微调（Adapter Tuning）：在每层插入小型可训练适配器模块。
- iv. LoRA（Low-Rank Adaptation）：用低秩分解减少微调参数量。
- v. QLoRA：LoRA的优化版，结合量化技术，适合资源有限环境下微调大