

lec01-绪论.md

lec01-作业-绪论

课程：李飞飞《谷歌云首席科学家李飞飞：一堂人工智能公开课》

一、简答题

1. Terry Winograd 看来要实现人工智能需要的三个要素（10 分）

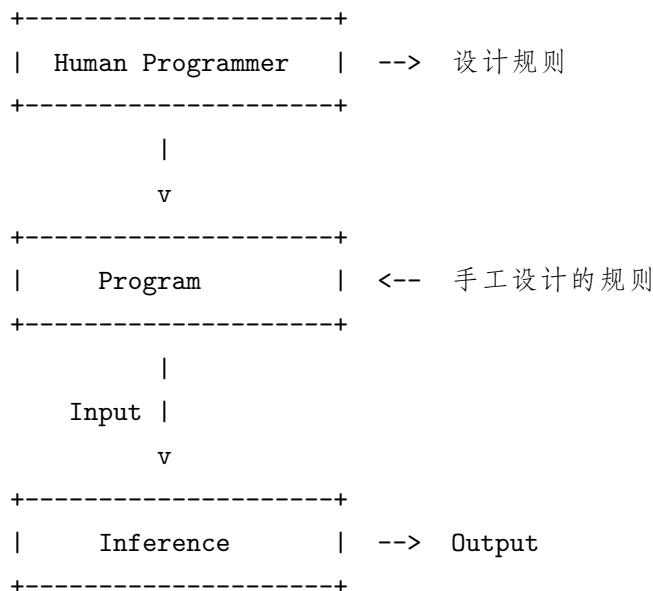
根据李飞飞在公开课中的介绍，Terry Winograd 认为要实现人工智能需要三个要素：

1. **Syntax**（语法/结构）：对世界结构的理解。在自然语言处理领域叫”语法”，在计算机视觉领域叫”三维结构”。
2. **Semantics**（语义/含义）：对含义的理解。语言有语言的含义，视觉有物体、动作等视觉的含义。
3. **Inference**（推理）：统计推导、统计推理的过程。当语法和语义解决之后，智能机器或人的主要任务就是进行推理。

2. 传统编程和机器学习的差别（10 分）

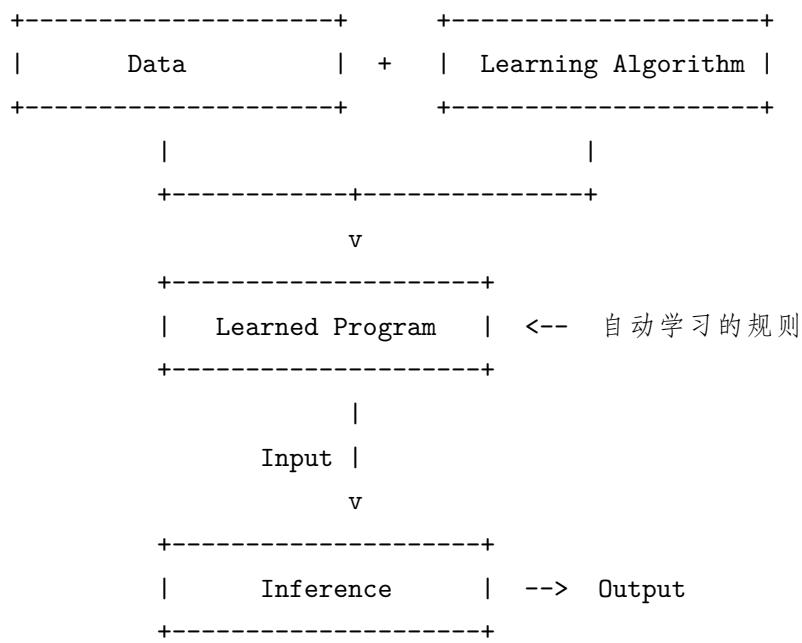
李飞飞在公开课中清晰地对比了传统编程 (Traditional Programming) 和机器学习 (Machine Learning) 的本质区别：

传统编程 (Traditional Programming)



特点：由人类程序员设计程序规则，然后将输入放入程序进行推理得到输出。

机器学习 (Machine Learning)



特点: 移除了人类程序员这一层, 改用学习算法 (Learning Algorithm), 直接从数据中学习出程序, 再用学习到的程序对输入进行推理得到输出。

核心差异对比

方面	传统编程	机器学习
规则来源	人工手动设计	从数据中自动学习
可扩展性	难以覆盖所有情况	数据越多效果越好
适应性	迁移到新领域需重新设计	具有更好的泛化能力
灵活性	规则固定, 难以处理模糊情况	能够处理复杂、模糊的场景

3. “看图说话”工作用到的两个模型 (10 分)

李飞飞在公开课后半部分介绍了团队将 vision 和 language、syntax 和 semantics 融合起来做“看图说话”(Image Captioning) 的工作, 这项工作用到了两个模型:

1. 卷积神经网络 (Convolutional Neural Network, CNN)

- 用于图片/图像的特征提取和理解
- 处理视觉信息, 提取图像中的物体、场景等特征

2. 循环神经网络 (Recurrent Neural Network, 具体是 LSTM)

- LSTM 全称: Long Short-Term Memory (长短期记忆网络)
- 用于语言的生成
- 基于 CNN 提取的视觉特征, 生成描述图片内容的句子

工作流程示意:

图片 → CNN(图像模型) → 图像特征 → LSTM(语言模型) → 生成句子

示例效果:

- 输入一张图片，计算机能够输出描述，如：
 - “A girl in a pink dress is jumping in the air”
 - “A construction worker in orange vest is working on a road”
 - “A man in black shirt is playing guitar”

4. 什么是人工智能？人工智能研究的基本内容有哪些？(20 分)

人工智能的定义

人工智能 (**Artificial Intelligence, AI**) 是计算机科学的一个分支，旨在研究和开发能够模拟、延伸和扩展人类智能的理论、方法、技术和应用系统。其核心目标是让机器能够执行通常需要人类智能才能完成的任务。

根据李飞飞在公开课中引用的图灵的观点：“*To build intelligent machines, it may be best to provide the machine with the best sense organs that money can buy and to teach it to understand and speak English.*” 一个有智能的机器需要两个重要元素：

- 感知 (**Sensing**): 如视觉系统
- 语言理解与推断 (**Understanding and Speaking Language**)

人工智能研究的基本内容

研究领域	主要内容
机器学习	让计算机从数据中学习规律，包括监督学习、无监督学习、强化学习等
深度学习	使用多层神经网络进行特征学习，如 CNN、RNN、Transformer 等
计算机视觉	让机器理解图像和视频，包括物体识别、场景理解、三维重建等
自然语言处理	让机器理解和生成人类语言，包括语法分析、语义理解、机器翻译等
知识表示与推理	如何表示知识并进行逻辑推理
机器人学	感知-行动系统，与物理世界交互
专家系统	模拟人类专家的决策能力
多模态融合	结合视觉、语言等多种信息进行综合理解（如李飞飞介绍的“看图说话”）

人工智能发展的三大要素（根据李飞飞公开课）

深度学习/神经网络取得突破性进展的三个关键因素：

1. 算法 (**Algorithm**): 神经网络算法的不断优化和成熟
2. 数据 (**Data**): 互联网带来的海量数据（如 ImageNet 数据集）
3. 计算能力 (**Computation/Hardware**): GPU 等硬件的发展

5. 简述人工智能三大主要学派及其特点和成就 (30 分)

人工智能发展至今，形成了三大主要学派：符号主义、联结主义和行为主义。

一、符号主义 (Symbolicism)

又称: 逻辑主义、心理学派、计算机学派

核心思想:

- 认为人类认知的基本单元是**符号**
- 认知过程是**符号操作**的过程
- 通过形式化的**符号、知识、规则和算法**来模拟人类智能

主要特点:

特点	说明
基于逻辑和规则	用逻辑规则和符号表示知识
显式知识表示	知识可以清晰表示, 便于解释和修改
推理过程明确	具有强可解释性
自上而下	采用演绎式推理, 从一般到特殊

主要成就:

- **专家系统**: 如 MYCIN (医疗诊断)、XCON (计算机配置)
- **知识图谱**: 结构化知识表示
- **定理证明**: 自动推理系统
- **SHRDLU 系统**: Terry Winograd 的博士论文, 70 年代初的积木世界问答系统

局限性: 知识获取困难、难以处理不确定性和常识推理

二、联结主义 (Connectionism)

又称: 仿生学派、生理学派

核心思想:

- 认为思维基元是**神经元**
- 通过模拟人脑神经网络来实现智能
- 智能通过大量简单处理单元的**连接和学习**产生

主要特点:

特点	说明
模拟神经网络	模仿生物神经系统的结构和功能
模式识别	擅长从数据中发现规律
分布式表示	智能分布在网络连接中
自下而上	采用归纳式推理, 从特殊到一般
强泛化能力	能对未见过的数据做出预测

发展历程（根据李飞飞公开课）：

- **1950 年代**: Rosenblatt 提出感知机 (Perceptron)
- **1960 年代**: Hubel & Wiesel 发现视觉皮层的层级结构
- **1980 年**: Fukushima 的 Neocognitron
- **1990 年代**: Hinton、LeCun 发展反向传播和卷积神经网络
- **2012 年**: AlexNet 在 ImageNet 上取得突破性成果

主要成就:

- **深度学习**: 图像识别错误率从 28% 降至 3.6%（接近或超越人类水平）
- **语音识别**: Siri、语音助手
- **机器翻译**: 神经机器翻译
- **看图说话**: CNN + LSTM 的多模态融合

局限性: 黑箱特性、可解释性差、需要大量训练数据

三、行为主义 (Behaviorism)

又称: 进化主义、控制论学派

核心思想:

- 智能取决于**感知和行动**
- 不需要复杂的知识表示和推理
- 智能通过与环境的**交互**产生

主要特点:

特点	说明
感知-行动	直接对环境刺激做出反应
无需内部模型	不依赖对世界的完整表征
智能涌现	复杂行为从简单规则中涌现
强调环境交互	智能是系统对环境的适应

主要成就:

- **强化学习**: AlphaGo、游戏 AI
- **机器人控制**: 移动机器人、自主导航
- **进化算法**: 遗传算法、进化策略
- **群体智能**: 蚁群算法、粒子群优化

局限性: 在复杂认知任务上泛化能力有限

三大学派对比总结

学派	核心观点	代表技术	优势	局限
符号主义	智能 = 符号运算	专家系统、知识图谱	可解释性强	知识获取难
联结主义	智能 = 神经网络	深度学习、CNN、RNN	模式识别强	黑箱问题
行为主义	智能 = 环境交互	强化学习、机器人	实时适应	认知受限

发展趋势: 现代 AI 往往融合多个学派的思想。例如 AlphaGo 结合了符号主义（蒙特卡洛树搜索）和联结主义（深度神经网络）以及行为主义（强化学习）的方法。

6. 人工智能技术在 2023 年杭州亚运会上的应用举例（20 分）

2023 年杭州亚运会是亚运会历史上首次提出“智能”办赛理念的综合性国际体育赛事，广泛运用了人工智能技术，打造了一届“智能亚运”。

一、数字火炬手与数字点火

应用描述：

- 超过 1 亿用户通过线上平台成为“数字火炬手”，参与虚拟火炬传递
- 运用 AI 捏脸和 AI 动作捕捉技术，实现“一人一面、自由换装”
- 提供数万亿种数字形象组合的可能性
- 开幕式上，由 1 亿多数字火炬手汇聚成的巨大数字人与现实火炬手共同点燃主火炬塔
- 每位数字火炬手都收到基于区块链技术的独特数字证书

技术融合：AI 数字人 + 云服务 + 区块链

二、云计算与智能管理平台

应用描述：

- 亚运会核心系统 100%“上云”，取代传统数据中心
- 首次采用云转播技术，确保赛事内容实时传输与分发
- 多个竞赛场馆配备数字化“智慧大脑”：
 - 杭州奥体中心体育场：“赛事在线”指挥平台
 - 数智气象平台
 - 智能医疗急救保障系统
- “亚运钉”平台：全球首个大型体育赛事一体化数字办赛平台，服务数十万工作人员和志愿者

三、AI 机器人应用

应用描述：

- **机器狗捡标枪：**田径赛场上首次使用机器狗捡拾标枪，保障安全并节省人力
- **全息 AI 机器人：**全球首款无介质全息 AI 机器人在场馆亮相，通过语音、手势与观众多维互动
- **AI 乒乓球机器人：**市民可与“AI 许昕”机器人对弈，体验世界冠军球技

四、5G/5G-Advanced 与智能观赛

应用描述：

- 全球首条 **5G-Advanced 万兆网络示范路线**
- 支持 **4K/8K 超高清赛事直播和裸眼 3D 应用**
- “**智能亚运一站通**”: 数字观赛服务平台，集成购票、出行、观赛、住宿等 28 项服务
- **AI+AR 观赛体验**: 观众可与亚运吉祥物 AR 互动、体验 AR 导航
- **AR 智能巴士**: 车载大屏幕展示历史文化与现实场景叠加

五、数字孪生与智能场馆

应用描述:

- **电网数字孪生**: 应用于泛亚运区域输配电网，提高电网运行安全性
- **地下管廊监控**: 亚运村地下综合管廊的实时监控
- **智能草坪养护**: 温州体育中心体育场草坪下安装传感器，实时采集数据
- **马术智能系统**: 桐庐亚运马术中心对马匹健康状况进行数字化管理

六、亚运元宇宙

应用描述:

- 杭州亚组委与中国移动联合发布“**亚运元宇宙**”平台
- 用户可在虚拟世界中:
 - 了解亚运知识
 - 游览城市
 - 参与虚拟竞技
 - 见证火炬传递

应用总结

应用领域	具体应用	涉及技术
开幕式	数字火炬手、数字点火	AI 数字人、区块链、云计算
赛事管理	云转播、智慧大脑、亚运钉	云计算、大数据、AI
赛场服务	机器狗、全息机器人	机器人技术、AI
观赛体验	裸眼 3D、AR 互动、元宇宙	5G、AR/VR、AI
场馆运维	数字孪生、智能监控	物联网、数字孪生

意义: 杭州亚运会展示了人工智能技术在大型国际体育赛事中的全方位应用，实现了科技与体育的深度融合，为未来智能体育赛事树立了标杆。

答案根据李飞飞《人工智能公开课》视频内容及相关资料整理

lec02-机器学习 1.md

lec02-作业-机器学习 1

题量: 15 | 满分: 120 | 作答时间: 2025-10-09 10:01 至 2025-10-23 23:59

一. 单选题 (共 4 题, 8 分)

1. (单选题, 2 分) 机器学习是一门多领域交叉学科, 涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。它主要的目标是让计算机系统 ()。

- A. 更快地执行指令
- B. 具有智能的外观
- C. 自动地从数据中学习模式并进行预测或决策
- D. 模仿人类的所有行为

我的答案: C: 自动地从数据中学习模式并进行预测或决策

2. (单选题, 2 分) 逻辑回归的整体经验风险函数包括经验风险 (交叉熵损失函数) 和结构风险 (L_2 正则), 正则化过度 (通常是正则化因子 λ 值太大) 会如何?

- A. 过拟合
- B. 欠拟合
- C. 没有影响
- D. 求解困难

我的答案: B: 欠拟合

3. (单选题, 2 分) 逻辑斯蒂回归模型 (logistic regression) 属于机器学习中的 ()

- A. 监督学习
- B. 无监督学习
- C. 半监督学习
- D. 以上都不是

我的答案: A: 监督学习

4. (单选题, 2 分) 在简单线性回归模型中, 目标是最小化以下哪一项?

- A. 预测值与实际值之间的乘积
- B. 预测值与实际值之间的均方误差
- C. 预测值与实际值之间的绝对误差
- D. 预测值与实际值之间的余弦相似度

我的答案: B: 预测值与实际值之间的均方误差

二. 多选题 (共 3 题, 12 分)

5. (多选题, 4 分) 数据通常由多个特征来描述, 以下是关于特征的说法中, 正确的是 ()。

- A. 特征是数据中可用于描述和区分不同样本的属性。不同的特征可以从不同的角度反映样本的特点。
- B. 所有特征对模型的贡献都是同等重要的。特征一旦确定, 就不能再进行调整和优化。
- C. 特征可以是数值型、类别型、文本型等不同类型。对于不同类型的特征, 需要采用不同的处理方法。
- D. 特征之间可能存在相关性。一些特征可能是冗余的, 去除冗余特征可以提高模型的训练效率和性能。

我的答案: ACD: 特征是数据中可用于描述和区分不同样本的属性。不同的特征可以从不同的角度反映样本的特点。; 特征可以是数值型、类别型、文本型等不同类型。对于不同类型的特征, 需要采用不同的处理方法。; 特征之间可能存在相关性。一些特征可能是冗余的, 去除冗余特征可以提高模型的训练效率和性能。

6. (多选题, 4 分) 常用来解决过拟合的方法有 ()

- A. 数据集扩增
- B. 正则化
- C. 选取合适的停止训练标准
- D. 对数据进行 K 折交叉验证

我的答案: ABCD: 数据集扩增; 正则化; 选取合适的停止训练标准; 对数据进行 K 折交叉验证

7. (多选题, 4 分) 梯度下降算法可能面临的问题有 ()。

- A. 梯度下降只适合于求解简单的线性回归问题。
- B. 学习率选择不当导致无法收敛 (损失函数值变得稳定, 模型能较好地拟合数据)。
- C. 批量梯度下降算法的计算量大, 尤其是对于大规模数据集。
- D. 对初始值非常敏感, 容易陷入局部最小值。

我的答案: BCD: 学习率选择不当导致无法收敛 (损失函数值变得稳定, 模型能较好地拟合数据)。; 批量梯度下降算法的计算量大, 尤其是对于大规模数据集。; 对初始值非常敏感, 容易陷入局部最小值。

三. 简答题 (共 5 题, 50 分)

8. (简答题, 10 分) 我们知道, 机器学习课程中需要我们掌握相当数量的模型和算法等, 为什么机器学习不叫“人类学习”? 请从你的角度阐释一下机器学习的定义, 以及其和人类学习的差别、联系。

我的答案:

机器学习是一种让计算机从数据中自动学习规律、改进性能的技术。它不需要人工写出具体规则, 而是通过算法和大量数据, 让计算机自己“摸索”出解决问题的方法。比如垃圾邮件分类器会从大量邮件中总结出哪些特征代表垃圾邮件, 从而在新邮件上做出判断。

之所以叫“机器学习”而不是“人类学习”, 是因为两者的本质不同。人类学习依靠感知、思考和理

解，通过经验形成知识和创造力；而机器学习只是通过数学模型和算法，不断调整参数来拟合数据规律。人类能理解、举一反三，机器只能优化、预测。

不过，两者也有联系。机器学习很多灵感来自人类学习，比如神经网络模仿了人脑结构，强化学习模仿了人类的奖惩机制。可以说，机器学习是用数学和计算去实现一种“类人”的经验总结方式，它虽然不会真正理解世界，但能在特定任务上表现出类似学习的能力。

9. (简答题, 10 分) 机器学习的三要素通常是指数据、模型和算法。请就你的认识，阐述“模型”一词的含义。

我的答案：

在机器学习中，“模型”是指通过对训练数据进行学习而得到的一种数学结构或算法表示。从功能角度来看，模型的作用是对新的输入数据进行预测或分类等任务。它从已有的数据中学习到数据的内在规律和模式，以便在面对未知数据时能够做出合理的推断。

从形式角度来讲，模型可以有多种表现形式。例如，线性回归模型是一个数学方程，通过学习数据中的特征与目标值之间的线性关系进行预测；神经网络模型则是由多个神经元组成的复杂结构，通过对数据进行多层的非线性变换来学习数据的特征表示。

模型的建立通常需要经过数据收集、数据预处理、选择合适的算法和模型架构、训练模型、评估模型性能等步骤。在训练过程中，模型不断调整自身的参数，以最小化损失函数，从而更好地拟合训练数据。而一个好的模型应该具有良好的泛化能力，即不仅在训练数据上表现良好，在新的、未见过的数据上也能有准确的预测或分类结果。

10. (简答题, 10 分) 请举例说明监督机器学习和无监督机器学习之间的区别。

我的答案：

监督学习是指从给定的有标注的训练数据集中学习出一个函数（模型参数），当新的数据到来时可以根据这个函数预测结果，常见任务包括分类与回归。

无监督学习是指对没有标注的训练数据集，需要根据样板间的统计规律对样本集进行分析，常见任务如聚类等。

11. (简答题, 10 分) 在机器学习中，什么是 k 折交叉验证？k 折交叉验证的作用是什么？

我的答案：

k 折交叉验证是指将原始数据集划分为 k 个子集，将其中一个子集作为验证集，其余 $k-1$ 个子集作为训练集，如此训练和验证一轮称为一次交叉验证。交叉验证重复 k 次，每个子集都做一次验证集，得到 k 个模型，加权平均 k 个模型的结果作为评估整体模型的依据。

k 折交叉验证的作用：在数据量小的情况下，通过 k 折交叉验证可以避免模型的过拟合现象。

12. (简答题, 10 分) 线性回归采用了均方误差 (Mean Squared Error) 作为损失函数。为什么线性回归不能采用平均误差 (Mean Error) 作为损失函数呢？请给出你的说明。提示：可结合求解、图示等多方面进行阐述。

我的答案：

在线性回归中，我们希望找到一组参数，使预测值最接近真实值。设样本点为 (x_i, y_i) ，模型预测值为 $\hat{y}_i = wx_i + b$ ，误差为 $e_i = y_i - \hat{y}_i$ 。

如果使用 **平均误差** (Mean Error, ME) 作为损失函数：

$$L = \frac{1}{n} \sum (y_i - \hat{y}_i) = \frac{1}{n} \sum (y_i - wx_i - b)$$

对参数 w 求偏导：

$$\frac{\partial L}{\partial w} = -\frac{1}{n} \sum x_i$$

对 b 求偏导：

$$\frac{\partial L}{\partial b} = -\frac{1}{n} \sum 1 = -1$$

可见，这两个导数与模型的预测误差无关，是常数或简单线性项，意味着无论模型预测得多差，梯度方向几乎不变或为零，无法正确指引参数优化。这种情况下，梯度下降法无法找到合适的 w 和 b ，模型根本学不到规律。

而若使用 **均方误差** (Mean Squared Error, MSE)：

$$L = \frac{1}{n} \sum (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum (y_i - wx_i - b)^2$$

对参数 w 求偏导：

$$\frac{\partial L}{\partial w} = -\frac{2}{n} \sum x_i (y_i - wx_i - b)$$

对 b 求偏导：

$$\frac{\partial L}{\partial b} = -\frac{2}{n} \sum (y_i - wx_i - b)$$

可以看到，导数与误差成比例，梯度方向会指向减少误差的方向，因此可以通过梯度下降有效更新参数。

从图形上看，平均误差对应的损失函数 $L(w, b)$ 关于 w 和 b 几乎是一条近似平面，正负误差会抵消，可能出现大范围的平坦区或多个局部极值点，优化过程不稳定；而均方误差的损失函数是一个光滑的抛物面形状，只有一个全局最小值，参数更新时梯度方向明确，优化容易收敛。

因此，线性回归不能采用平均误差作为损失函数，因为平均误差既不能准确衡量预测偏差，也无法提供有意义的梯度信息；而均方误差既能反映整体误差大小，又能保证损失函数连续可导、优化过程平稳，最终得到最优参数。

四. 其它 (共 3 题, 50 分)

13. (其它, 15 分) 请选择你认为合适的回归模型 (线性回归、多项式回归、岭回归等) 来预测中国 2024 年的出生人口和死亡人口。请简单展示你的数据、算法模型、最终结果，并附上相应代码 (能正常运行得出答案的代码)。

附：参考数据（见学习通“资料”-“课程资料-高海东”-“lec02”-“china_population.txt”），参考代码（见学习通“资料”-“课程资料-高海东”-“lec02”-“calc_regression.py”）。

注：亦可自行在网上搜集更全的数据来做训练（但需在答案中备注数据来源）。

我的答案：

一、实验目的 使用多种回归模型 (线性回归、多项式回归、岭回归等) 预测中国 2024 年的出生人口和死亡人口，并对比不同模型的性能，选择最优模型。

二、数据展示 2.1 数据来源

使用 1997-2023 年中国人口数据 (china_population.txt)，包含：

- 年份 (year): 1997-2023 年
- 出生人口 (birth): 单位为万人
- 死亡人口 (death): 单位为万人

2.2 数据概览

数据共 27 条记录，展示了中国近 27 年的人口变化趋势：

- 出生人口呈现明显的下降趋势，从 1997 年的 2038 万降至 2023 年的 902 万
- 死亡人口呈现上升趋势，从 1997 年的 801 万升至 2023 年的 1110 万
- 自 2020 年起，死亡人口已超过出生人口，出现人口负增长

三、算法模型 3.1 模型选择

本次实验对比了以下 5 种回归模型：

1. 线性回归 (Linear Regression)
 - 最简单的回归模型，假设因变量与自变量呈线性关系
 - 公式： $y = \beta_0 + \beta_1 x$
2. 多项式回归 - 二次 (Polynomial Regression, degree=2)
 - 可以捕捉数据的非线性趋势

- 公式: $y = \beta_0 + \beta_1 x + \beta_2 x^2$
3. 多项式回归 - 三次 (Polynomial Regression, degree=3)
- 更高阶的多项式, 可以拟合更复杂的曲线
 - 公式: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$
4. 岭回归 (Ridge Regression, $\alpha = 1.0$)
- 带 L2 正则化的线性回归, 可以防止过拟合
 - 添加惩罚项: $\min(||y - X\beta||^2 + \alpha||\beta||^2)$
5. 多项式岭回归 (Polynomial Ridge Regression, degree=2, $\alpha = 10.0$)
- 结合多项式特征和岭回归的优势

3.2 模型评估指标

- RMSE (Root Mean Square Error):** 均方根误差, 越小越好
- R² Score:** 决定系数, 范围 [0,1], 越接近 1 表示拟合效果越好

四、实验结果 4.1 出生人口预测模型对比

模型	训练集 RMSE	测试集 RMSE	训练集 R ²	测试集 R ²
线性回归	189.10	97.49	0.5547	0.6781
多项式回归 (度 2)	159.39	196.53	0.6836	-0.3081
多项式回归 (度 3)	65.11	65.87	0.9472	0.8531
岭回归	189.10	97.47	0.5547	0.6783
多项式岭回归 (度 2)	188.89	97.77	0.5557	0.6763

最佳模型：多项式回归 (度 3)

- 训练集 $R^2 = 0.9472$ (拟合度非常好)
- 测试集 $R^2 = 0.8531$ (泛化能力强)
- 测试集 RMSE = 65.87 (误差较小)

4.2 死亡人口预测模型对比

模型	训练集 RMSE	测试集 RMSE	训练集 R ²	测试集 R ²
线性回归	21.40	21.79	0.9410	0.8846
多项式回归 (度 2)	21.12	21.45	0.9425	0.8881
多项式回归 (度 3)	21.02	22.45	0.9431	0.8775
岭回归	21.40	21.77	0.9410	0.8848
多项式岭回归 (度 2)	21.42	21.82	0.9409	0.8843

最佳模型：多项式回归 (度 2)

- 训练集 $R^2 = 0.9425$ (拟合度优秀)
- 测试集 $R^2 = 0.8881$ (泛化能力优秀)
- 测试集 RMSE = 21.45 (误差很小)

4.3 2024 年预测结果

使用最佳模型对 2024 年进行预测：

- 预测年份: 2024
- 预测出生人口: 571.85 万人
- 预测死亡人口: 1062.53 万人
- 预测人口自然增长: -490.68 万人

14. (其它, 15 分) 根据参考资料复旦大学邱锡鹏老师的《神经网络与深度学习》(见学习通“资料”-“课内外资料”)一书中的“附录 B.3 矩阵微积分”, 或《矩阵求导基础(简略)》(见学习通“资料”-“课程资料-党源杰”-“lec02”), 或知乎网站 (<https://zhuanlan.zhihu.com/p/262751195>), 请推导多元线性岭回归 (Ridge Regression) 的直接求解公式 (从矩阵求导的角度来推导岭回归的最小化问题)。

我的答案：

目标：最小化岭回归目标函数，给出闭式解。设 $X \in \mathbb{R}^{N \times D}$, $Y \in \mathbb{R}^{N \times K}$ ($K = 1$ 时是普通一元输出), 参数 $W \in \mathbb{R}^{D \times K}$ 。损失定义为：

$$J(W) = \|Y - XW\|_F^2 + \lambda\|W\|_F^2, \quad \text{其中 } \lambda \geq 0$$

用矩阵求导 (trace 技巧) 推导：

把 Frobenius 范数写成迹：

$$\begin{aligned} \|Y - XW\|_F^2 &= \text{tr}[(Y - XW)^T(Y - XW)] \\ &= \text{tr}(Y^T Y) - 2\text{tr}(Y^T XW) + \text{tr}(W^T X^T XW) \end{aligned}$$

因此

$$J(W) = \text{tr}(Y^T Y) - 2\text{tr}(Y^T XW) + \text{tr}(W^T X^T XW) + \lambda\text{tr}(W^T W)$$

对 W 求梯度, 利用恒等式:

$\frac{\partial}{\partial W} \text{tr}(A^T W) = A$, $\frac{\partial}{\partial W} \text{tr}(W^T A W) = (A + A^T)W$, 当 A 对称时为 $2AW$ 。这里 $X^T X$ 与 I 都是对称的。

$$\nabla_W J = -2X^T Y + 2X^T XW + 2\lambda W$$

令梯度为 0 (必要条件):

$$(X^T X + \lambda I_D) W = X^T Y$$

若 $\lambda > 0$, 则 $X^T X + \lambda I_D$ 正定可逆, 解为闭式:

$$W^* = (X^T X + \lambda I_D)^{-1} X^T Y$$

($K = 1$ 时退化为 $w^* = (X^T X + \lambda I)^{-1} X^T y_0$)

15. (其它, 20 分) 【附加题】在智能制造中, 工件的表面质量和几何形状是质量控制的重要指标。利用激光扫描或其他传感器获得的 3D 点云数据, 可以反映工件表面的实际状态。为检测平整度和定位校准, 常需要根据采集到的点云数据拟合出一个平面方程。

假设某工件表面经过激光扫描获得了下列 10 个 3D 点 (单位: 毫米): P1=(10.2, 20.1, 30.5) P2=(12.0, 21.3, 31.0) P3=(11.5, 19.8, 30.8) P4=(13.1, 22.0, 31.5) P5=(12.3, 20.5, 31.2) P6=(11.8, 21.0, 30.9) P7=(12.7, 20.8, 31.3) P8=(13.0, 22.2, 31.7) P9=(12.5, 21.5, 31.1) P10=(12.9, 21.8, 31.4)

请计算出从这些点拟合出来的平面方程。注意: 可以手动计算, 亦可使用代码; 但必须张贴详细的过程 (手动计算过程或代码截图等)。

我的答案:

- SVD 奇异值分解法: $z = 0.333142x + 0.075349y + 25.485803$
- 最小二乘法: $z = 0.323603x + 0.082250y + 25.456571$

lec08-机器学习 2.md

lec08-作业-机器学习 2

题量: 24 | 满分: 120 | 作答时间: 2025-10-24 10:44 至 2025-11-07 23:59

一. 单选题 (共 10 题, 20 分)

1. (单选题, 2 分) 支持向量机的主要目标是什么?

- A. 最大化分类错误
- B. 最小化支持向量的数量
- C. 寻找最优分类超平面
- D. 增加模型的复杂度

我的答案: C

2. (单选题, 2 分) 在支持向量机中, 核函数的主要作用是什么?

- A. 一次性计算好数据点之间的距离, 避免重复计算。
- B. 将数据映射到更高维的空间, 使数据变得线性可分。
- C. 优化支持向量机的超参数, 使模型更加简洁。
- D. 引入更多的参数, 从而提高模型的精度。

我的答案: B

3. (单选题, 2 分) 假设你在做股市预测。你想预测某家上市公司会在未来一个月内宣布破产 (例如通过对之前面临破产风险的类似公司的数据进行训练)。你会把这个事情当作一个 () 问题来建模?

- A. 回归
- B. 聚类
- C. 降维
- D. 分类

我的答案: D

4. (单选题, 2 分) 决策树算法中, 哪个指标常用于选择最佳分裂特征?

- A. 均方误差
- B. 支持向量
- C. 熵或信息增益
- D. 欧几里得距离

我的答案: C

5. (单选题, 2 分) K-Means 聚类的核心目标是什么?

- A. 最大化簇间距离
- B. 最小化分类错误率

- C. 最大化簇内数据点的距离
- D. 最小化簇内数据点的方差

我的答案: D

6. (单选题, 2 分) 以下哪种算法是无监督学习算法? ()

- A. 支持向量机
- B. 决策树
- C. K-means 算法
- D. 以上都是

我的答案: C

7. (单选题, 2 分) K 近邻 (KNN) 和 K 均值 (K-means) 都是经典的机器学习算法, 以下说法正确的是?

- A. 作为一种监督学习算法, KNN 自然优于属于无监督算法的 K-means。
- B. KNN 不需要进行算法的训练。故相比 K-means, KNN 有明显的优越性。
- C. K-means 算法经过了精心的设计和严格的推导, 比 KNN 更有可解释性。
- D. KNN 和 K-means 的一个影响算法效果的因素都包括: K 值的选择。

我的答案: D

8. (单选题, 2 分) 为了 (), 强化学习算法经常使用折扣因子 γ 。

- A. 平衡探索和利用
- B. 控制学习速率
- C. 平衡短期奖励和长期奖励
- D. 控制动作选择的随机性

我的答案: C

9. (单选题, 2 分) 在 Q 学习 (Q-learning) 算法中, Q 值更新的公式是什么?

- A. $Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha[R + \gamma Q(s', a')]$
- B. $Q(s, a) \leftarrow Q(s, a) + \alpha[R + \gamma Q(s', a') - Q(s, a)]$
 $Q(s, a) \leftarrow Q(s, a) + \alpha[R + \gamma \max_{a'} Q(s', a') - Q(s, a)]$
- C. $Q(s, a) \leftarrow Q(s, a) + \alpha[R + \gamma \max_{a'} Q(s', a') - Q(s, a)]$
- D. $Q(s, a) \leftarrow Q(s, a) + \alpha[R + \gamma \max_{a'} Q(s', a) - Q(s, a)]$

我的答案: C

10. (单选题, 2 分) 对于一个二分类问题而言, 假设识别结果的混淆矩阵如下表所示, 表中正例 (Positives) 表示识别对的; 负例 (Negatives) 表示识别错的, 那么正确率 (Accuracy) 为 ()。

		预测值	
		Positive	Negatives
实际值	Positive	TP=50	FN=10
	Negatives	FP=20	TN=20

- A. $\frac{5}{7}$
- B. $\frac{5}{6}$
- C. $\frac{10}{13}$
- D. $\frac{7}{10}$

我的答案: D

二. 多选题 (共 5 题, 20 分)

11. (多选题, 4 分) 支持向量机中可以使用哪些核函数?

- A. 线性核
- B. 多项式核
- C. RBF (径向基函数) 核
- D. Sigmoid 核

我的答案: ABCD

12. (多选题, 4 分) 以下哪些描述是关于支持向量机的?

- A. 通过最大化间隔来进行分类
- B. 支持向量机只能用于线性可分的数据
- C. 核技巧可以将低维数据映射到高维空间
- D. 支持向量是最靠近决策边界 (超平面) 的数据点。

我的答案: ACD

13. (多选题, 4 分) K-Means 聚类算法中的 K 值对结果有什么影响?

- A. K 值决定了簇的数量
- B. K 值越大, 聚类效果一定越好

- C. K 值可以通过肘部法则 (Elbow Method) 选择
- D. K 值必须与数据集中的类数量相等

我的答案: AC

14. (多选题, 4 分) 以下哪些算法可以用于降维?

- A. K-Means
- B. PCA
- C. t-SNE
- D. 决策树

我的答案: BC

15. (多选题, 4 分) 全局最优是指在所有可能的解中, 找到一个解使得目标函数值达到最大或最小。这是一个决策与所有可能的决策相比是最优的解决方案。下列模型中, 哪些模型不能保证找到全局最优解。()

- A. 使用最小二乘求解的线性回归
- B. 使用梯度下降求解的逻辑回归
- C. 使用信息增益求解的决策树
- D. 采用随机初始化的 K-means 算法

我的答案: BCD

三. 简答题 (共 2 题, 10 分)

16. (简答题, 5 分) 比较 SVM 和决策树这两种分类算法的优缺点, 以及它们适用的数据场景。

我的答案: 支持向量机 (SVM) 适合高维数据, 尤其是在数据线性可分或通过核函数可以映射到高维线性可分的场景。SVM 具有良好的泛化能力, 但在处理大型数据集时, 计算复杂度较高。另一方面, 决策树结构简单, 容易理解和解释, 适用于处理类别型和数值型数据, 但容易过拟合, 尤其是在数据不纯或数据量较少时。决策树还对噪声数据比较敏感。两者各有优缺点, SVM 更适合高维和复杂数据, 决策树适合需要解释性的场景。

17. (简答题, 5 分) 逻辑回归和线性回归有什么区别?

我的答案: 逻辑回归解决的是分类问题, 而线性回归既可以解决分类问题又可以解决回归问题。逻辑回归可以看成是广义线性回归的一种特殊形式, 其特殊之处在于其目标的取值服从二元分布。

四. 计算题 (共 5 题, 50 分)

18. (计算题, 5 分) 损失函数在深度学习、机器学习中扮演了重要作用。

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

1) MSE 损失 , 其中 y_i 、 \hat{y}_i 、 n 分别表示真实值、预测值和样本个数。假设真实值 \mathbf{y} 分别为 3, -1, 2, 4, 1; 预测值 $\hat{\mathbf{y}}$ 分别为 2.5, -0.9, 1.8, 3.9, 0.8。计算它们的 MSE 损失值。(1 分)

$$\mathcal{L}(\mathbf{y}, \mathbf{p}) = - \sum_{i=1}^C y_i \ln(p_i)$$

2) 交叉熵损失 , 其中 \mathbf{y} 、 \mathbf{p} 、 C 分别表示真实值向量 (通常是 one-hot 的, 即向量中只有一个值为 1, 其他全为 0)、预测值向量 (概率向量, 所有值在 0~1 之间, 向量和为 1) 和类别个数。假设真实值向量是 $\mathbf{y} = [0, 0, 1]$; 预测值向量 $\mathbf{p} = [0.3, 0.3, 0.4]$ 。计算它们的交叉熵损失值。(1 分)

3) 在分类问题中, 待分类的类别数量会对交叉熵数值有重要影响。试分别计算: 在分类类别数目为 2、10 时, 交叉熵损失的数值基线分别是多少? 提示: 对于分类来说, 随机猜测 (从待分类的类别中随机选一个作为输出结果) 可以作为算法的一个基线。换句话说, 我们训练出来的算法不应差于随机猜测, 而随机猜测下的交叉熵损失数值可以作为算法训练的一个基线。(3 分)

我的答案:

1. 0.07
2. 0.9163
3. C=2: 0.6931; C=10: 2.3026

19. (计算题, 5 分) 给定如下的数据点 (每个点第一维对应输入特征 x, 第二维对应输出目标 y):

$$(1, 2), (2, 4), (3, 6), (4, 8)$$

请根据这些数据点, 使用最小二乘法手动计算线性回归模型 $y = wx + b$ 中的参数 w 和 b 。

我的答案:

$$w = 2 \quad b = 0$$

20. (计算题, 10 分) 假设有一个关于是否购买电脑 (“是”或“否”) 的数据集, 包含三个属性: 年龄 (“青年”、“中年”、“老年”)、收入 (“高”、“中”、“低”) 和信用等级 (“好”、“中”、“差”)。数据集共有 15 个样本, 具体分布如下:

年龄	收入	信用等级	是否购买电脑
青年	高	好	是
青年	高	中	是
青年	中	好	是

年龄	收入	信用等级	是否购买电脑
青年	中	中	否
青年	低	好	否
青年	低	中	否
中年	高	好	是
中年	高	中	是
中年	中	好	是
中年	中	中	否
中年	低	好	否
中年	低	中	否
老年	高	好	否
老年	高	中	否
老年	中	好	否

请计算在根节点处，年龄、收入和信用等级这三个属性的信息增益，以确定哪个属性最适合作为根节点的分裂属性。（提示：计算信息增益需要先计算信息熵）

我的答案：

年龄: 0.1710 收入: 0.2800 信用等级: 0.0348 最适合作为根节点的分裂属性: 收入 (信息增益最大)

21. (计算题, 10 分) 假设你有以下 5 个二维数据点：

$$(2, 3), (3, 3), (6, 8), (8, 8), (9, 10)$$

使用 K-Means 聚类算法将这些点分为 2 个簇。

- 初始聚类中心为 (2,3) 和 (6,8)。
- 进行一次迭代后
 - 计算簇所包含的点。(4 分)
 - 并更新新的聚类中心。(4 分)

根据你的计算，此处的 K-means 算法需要几次迭代就能收敛？(2 分)

我的答案：

1. C1: (2,3),(3,3); C2: (6,8),(8,8),(9,10)
2. (2.5,3.0); (7.667,8.667)
3. 2 次迭代

22. (计算题, 20 分) 你是工厂的质量主管，你有一些工件在两次测试中的测试结果。基于这些测试数据，你打算构建一个分类模型，来决定接受还是放弃工件。example.py 给出了一个简单的基于逻辑回归来做分类的例子，但是它的准确率不够高（未及 0.5）。请你基于逻辑回归，或支持向量机，或决策树，来训练一个在测试集上准确率超过 0.7 的模型。请编写代码实现（将代码及运行结果截图答在下面）。

我的答案：

```

1  import numpy as np
2  from pathlib import Path
3  from sklearn.preprocessing import StandardScaler
4  from sklearn.pipeline import Pipeline
5  from sklearn.linear_model import LogisticRegression
6  from sklearn.metrics import accuracy_score
7  from sklearn.svm import SVC
8  from sklearn.tree import DecisionTreeClassifier
9
10 root = Path(__file__).resolve().parent.parent
11 train_path = root / "../data_train.txt"
12 test_path = root / "../data_test.txt"
13
14 def load_data(path):
15     data = np.loadtxt(path, delimiter=",")
16     X = data[:, :2]
17     y = data[:, 2].astype(int)
18     return X, y
19
20 X_train, y_train = load_data(train_path)
21 X_test, y_test = load_data(test_path)
22
23 candidates = {
24     "logreg": Pipeline([
25         ("scaler", StandardScaler()),
26         ("clf", LogisticRegression(max_iter=1000))
27     ]),
28     "svm": Pipeline([
29         ("scaler", StandardScaler()),
30         ("clf", SVC(kernel="rbf", C=1.0, gamma="scale"))
31     ]),
32     "tree": DecisionTreeClassifier(max_depth=3, random_state=42),
33 }
34
35 best_name = None
36 best_model = None
37 best_acc = -1.0
38
39 for name, model in candidates.items():
40     model.fit(X_train, y_train)
41     preds = model.predict(X_test)
42     acc = accuracy_score(y_test, preds)
43     print(f"{name}: {acc:.3f}")
44     if acc > best_acc:
45         best_acc = acc
46         best_model = model
47         best_name = name
48
49 print(f"Best model: {best_name} (accuracy={best_acc:.3f})")
50
51 preds = best_model.predict(X_test)
52 np.savetxt(Path(__file__).with_suffix('.preds.txt'), preds, fmt="%d")
53

```

logreg: 0.458
svm: 0.708
tree: 0.667
Best model: svm
taw@TAWdeMacBook

五. 计算题 (共 2 题, 20 分)

23. (计算题, 5 分) 假定对维数为 2 的数据 X 进行主成分分析。经过均值处理之后, 再计算得到的协方差矩阵为

$$\begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$$

请: 1) 计算第一主成分 v_1 、第二主成分 v_2 ; 2) 第一主成分所占的比例。(要求: 给出详细的计算过程)。

我的答案:

$$v1 = [2/\sqrt{5}, 1/\sqrt{5}] \quad v2 = [-1/\sqrt{5}, 2/\sqrt{5}] \quad \text{比例} = 0.8571$$

计算过程

^ ⊙ ≡

协方差矩阵

$$\Sigma = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$$

的特征值由特征方程

$$\det(\Sigma - \lambda I) = 0$$

确定：

$$\begin{vmatrix} 5 - \lambda & 2 \\ 2 & 2 - \lambda \end{vmatrix} = (5 - \lambda)(2 - \lambda) - 4 = \lambda^2 - 7\lambda + 6 = 0.$$

解得特征值 $\lambda_1 = 6, \lambda_2 = 1$ 。

对 $\lambda_1 = 6$, 解线性方程 $(\Sigma - 6I)v = 0$:

$$\begin{bmatrix} -1 & 2 \\ 2 & -4 \end{bmatrix} v = 0 \Rightarrow v \propto (2, 1).$$

单位化得

$$v_1 = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

对 $\lambda_2 = 1$, 解 $(\Sigma - I)v = 0$:

$$\begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix} v = 0 \Rightarrow v \propto (-1, 2).$$

单位化得

$$v_2 = \frac{1}{\sqrt{5}} \begin{bmatrix} -1 \\ 2 \end{bmatrix}.$$

第一主成分方差占比为

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{6}{6 + 1} = \frac{6}{7} \approx 0.8571.$$

24. (计算题, 15 分) 假设有一个简单的 MDP (马尔可夫决策过程), 包含两个状态 S_1 和 S_2 , 以及两个动作 A_1 和 A_2 。状态转移和奖励如下所示:

$S_1 \rightarrow A_1 \rightarrow S_2$ (奖励+2, 转移概率0.5)

$S_1 \rightarrow A_2 \rightarrow S_1$ (奖励-1, 转移概率1.0)

$S_2 \rightarrow A_1 \rightarrow S_1$ (奖励+3, 转移概率0.7)

$S_2 \rightarrow A_2 \rightarrow S_2$ (奖励-1, 转移概率1.0)

折扣因子 γ 为 0.9。试根据价值迭代算法计算 S_1 和 S_2 价值函数值 $V_\pi(S_1)$ 和 $V_\pi(S_2)$ 。

- 求出 $V_\pi(S_1)$ 和 $V_\pi(S_2)$ 的递推关系式。(5 分)

$$V_\pi(s) \leftarrow \max_a \sum_{s'} P(s'|s, a)[R(s, a, s') + \gamma V_\pi(s')]$$

提示:

- 根据递推关系式迭代求出 $V_\pi(S_1)$ 和 $V_\pi(S_2)$ 。提示: 1) 可编写代码 (将其截图放在下面), 通过迭代来进行求解 (结果保留 3 位有效数字); 2) 或通过分析递推表达式求解 (结果可写成分数, 或保留 3 位有效数字的小数)。(10 分)

我的答案:

根据价值迭代公式: $V(s) \leftarrow \max_a \sum P(s'|s, a)[R(s, a, s') + V(s')]$

对于 S : 选择 A : $Q(S, A) = 0.5 \times [2 + 0.9 \times V(S)]$ 选择 A : $Q(S, A) = 1.0 \times [-1 + 0.9 \times V(S)]$ $V(S) = \max\{0.5 \times [2 + 0.9V(S)], 1.0 \times [-1 + 0.9V(S)]\} = \max\{1 + 0.45V(S), -1 + 0.9V(S)\}$

对于 S : 选择 A : $Q(S, A) = 0.7 \times [3 + 0.9 \times V(S)]$ 选择 A : $Q(S, A) = 1.0 \times [-1 + 0.9 \times V(S)]$ $V(S) = \max\{0.7 \times [3 + 0.9V(S)], 1.0 \times [-1 + 0.9V(S)]\} = \max\{2.1 + 0.63V(S), -1 + 0.9V(S)\}$

$$\begin{aligned} V(S1) &= \max(0.5*(2+0.9*V(S2)), -1+0.9*V(S1)) \\ V(S2) &= \max(0.7*(3+0.9*V(S1)), -1+0.9*V(S2)) \\ V(S1) &\approx 2.715, V(S2) \approx 3.810 \end{aligned}$$

```
1 gamma = 0.9
2
3 # Initialize values
4 V1, V2 = 0.0, 0.0
5
6 print(f"k=0: V1={V1:.4f}, V2={V2:.4f}")
7
8 for k in range(1, 21):
9     q1_a1 = 0.5 * (2 + gamma * V2)
10    q1_a2 = -1 + gamma * V1
11    q2_a1 = 0.7 * (3 + gamma * V1)
12    q2_a2 = -1 + gamma * V2
13    V1_new = max(q1_a1, q1_a2)
14    V2_new = max(q2_a1, q2_a2)
15    V1, V2 = V1_new, V2_new
16    print(f"k={k}: V1={V1:.4f}, V2={V2:.4f}")
17
18 print("Final:", V1, V2)
19
```

```
k=0: V1=0.0000, V2=0.0000
k=1: V1=1.0000, V2=2.1000
k=2: V1=1.9450, V2=2.7300
k=3: V1=2.2285, V2=3.3253
k=4: V1=2.4964, V2=3.5040
k=5: V1=2.5768, V2=3.6727
k=6: V1=2.6527, V2=3.7234
k=7: V1=2.6755, V2=3.7712
k=8: V1=2.6970, V2=3.7856
k=9: V1=2.7035, V2=3.7991
k=10: V1=2.7096, V2=3.8032
k=11: V1=2.7114, V2=3.8071
k=12: V1=2.7132, V2=3.8082
k=13: V1=2.7137, V2=3.8093
k=14: V1=2.7142, V2=3.8096
k=15: V1=2.7143, V2=3.8099
k=16: V1=2.7145, V2=3.8100
k=17: V1=2.7145, V2=3.8101
k=18: V1=2.7146, V2=3.8101
k=19: V1=2.7146, V2=3.8102
k=20: V1=2.7146, V2=3.8102
Final: 2.714575683122194 3.8101756374928475
taw@TAWdeMacBook-Pro lec-08-homework-machineLearning %
```

lec16-深度学习.md

lec16-作业-深度学习 1

题量: 22 | 满分: 120 | 作答时间: 2025-11-28 10:18 至 2025-12-12 23:59

一. 单选题 (共 10 题, 20 分)

1. (单选题, 2 分) 从你的角度来看, 怎么理解“人工智能就是新电力”这种说法?

- A. 人工智能为我们的家用或公用设备供电。
- B. 人工智能是由电力驱动, 并在计算机上运行, 能做很多的事。
- C. 人工智能和电力革命一样, 改变了很多的行业。
- D. 人工智能能让电网更加智能, 从而提供新的电能。

我的答案: C

2. (单选题, 2 分) 反向传播算法中的“梯度下降”是用来:

- A. 训练神经网络的机构。
- B. 计算损失函数的值。
- C. 更新网络权重。
- D. 计算激活函数的值。

我的答案: C

3. (单选题, 2 分) 深度学习 (Deep Learning) 中“深度”这个概念是指?

- A. 计算机理解问题的深度。
- B. 神经网络能够深入地剖析问题本身。
- C. 神经网络的层数很多。
- D. 计算机的求解更加精确。

我的答案: C

4. (单选题, 2 分) 深度学习经常用到大量的全连接层。而全连接层的计算核心即为矩阵相乘。假设三个稠密矩阵 A 、 B 和 C , 其尺寸分别为 $l \times m$ 、 $m \times n$ 和 $n \times p$, 且满足 $l < m < n < p$ 。则下列说法正确的是?

- A. $(AB)C$ 的计算效率低于 $A(BC)$ 。
- B. $(AB)C$ 的计算效率等于 $A(BC)$ 。
- C. $(AB)C$ 的计算效率高于 $A(BC)$ 。
- D. $(AB)C$ 和 $A(BC)$ 的计算效率无法比较。

我的答案: C

5. (单选题, 2 分) CNN 的卷积层为什么能够捕捉到局部特征?

- A. 因为卷积层使用了高斯核。

- B. 因为卷积层使用了 dropout。
- C. 因为卷积层的权重是局部连接的。
- D. 因为卷积层有多个输出通道。

我的答案: C

6. (单选题, 2 分) 空洞卷积 (Dilated Convolution) 是一种在卷积神经网络中使用的卷积操作, 它通过在卷积核的每个元素之间插入空洞 (即插入零) 来扩大卷积核的感受野, 同时保持计算复杂度不变。这种方法尤其在语音、图像和时序数据等任务中非常有用, 能够帮助模型捕捉更大范围的上下文信息。

空洞率为 1: 即传统卷积, 卷积核的每个元素直接相邻。空洞率为 2: 卷积核的每个元素之间有一个零, 卷积核的感受野会增加一倍。空洞率为 3: 卷积核的每个元素之间有两个零, 感受野进一步增大。

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \xrightarrow{\text{空洞率为2}} \begin{bmatrix} 1 & 0 & 2 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 \\ 4 & 0 & 5 & 0 & 6 \\ 0 & 0 & 0 & 0 & 0 \\ 7 & 0 & 8 & 0 & 9 \end{bmatrix}$$

例如:

题目: 在一个卷积神经网络中, 输入特征图的尺寸为 64×64 , 通道数为 3。使用一个空洞卷积层, 卷积核大小为 3×3 , 空洞率为 2, 步长为 1, 补边为 1。请问经过该空洞卷积层后输出特征图的尺寸是多少?

- A. 64
- B. 62
- C. 32
- D. 31

我的答案: B

7. (单选题, 2 分) 人工神经网络的基本组成单元是 0

- A. 感知机
- B. 卷积核
- C. 神经元
- D. 池化层

我的答案: C

8. (单选题, 2 分) 如果一个神经网络只有输入层和输出层, 没有隐藏层, 那么它本质上是一个?

- A. 感知机
- B. 线性回归或逻辑回归模型
- C. 自编码器

- D. 卷积神经网络

我的答案: B

9. (单选题, 2 分) 关于全连接层, 以下描述正确的是?

- A. 层内神经元之间两两连接
- B. 前一层的每一个神经元与后一层的每一个神经元都相互连接
- C. 只有相邻的神经元之间存在连接
- D. 层与层之间没有连接

我的答案: B

10. (单选题, 2 分) 池化层 (如最大池化) 的主要目的是?

- A. 增加特征的维度
- B. 提取图像的色彩信息
- C. 进行下采样, 保持特征不变性并减少参数
- D. 引入非线性特性

我的答案: C

二. 多选题 (共 8 题, 32 分)

11. (多选题, 4 分) 卷积神经网络广泛应用于计算机视觉等领域。其中的卷积计算主要的特性有 ()?

- A. 局部连接
- B. 等变表达
- C. 共享参数
- D. 位置编码

我的答案: ABC

12. (多选题, 4 分) 面对一个图像分类任务, 你用卷积、池化、ReLU、全连接算子自己设计了一个图像分类网络。然而, 当你训练这个网络时, 却发现 loss 始终降不下来。那么, 根据你的所学, 从模型结构的角度来看你可以尝试下列哪些操作来帮助你?

- A. 检查数据: 如数据和标签是否存在错误的对应等;
- B. 数据处理: 对输入数据做归一化处理等;
- C. 网络模型: 使用归一化层 (如 BN, 即 Batch Normalization)、更换合适的权重 (weight) 初始化方法、尝试带残差 (Residual) 结构的网络等;
- D. 学习率: 尝试不同的学习率和选择合适的优化器等;

我的答案: ABCD

13. (多选题, 4 分) 激活函数在深度学习中扮演非常重要的作用, 下列关于它的说法正确的是 ()

- A. 两个相邻全连接层中间如果不加激活函数, 这几乎没有意义。

- B. 激活函数应该是连续并可导（允许少数点上不可导）的，以便通过反向传播算法计算梯度
- C. 激活函数解决了梯度消失的问题。
- D. sigmoid 激活函数导数的最大值为 0.25，在输入小于-5 或大于 5 时，导数已经小于 0.01。

我的答案: ABD

14. (多选题, 4 分) 由于卷积核一般比原始图像小，所以卷积之后的图像尺寸往往变小。有时候我们需要将卷积后的图像还原成原始图像的尺寸，即实现图像从小分辨率到大分辨率的映射，这种操作就叫做上采样 (Upsampling)。而反卷积正是一种上采样方法。题目：反卷积一般多用于下列哪些深度学习任务 ()。

- A. 像素级别的图像语义分割
- B. 矩形框级别的图像目标检测
- C. 基于自编码器 (AE) 的图像生成
- D. 图像全图级别的分类

我的答案: AC

15. (多选题, 4 分) 一个典型的卷积神经网络 (CNN) 通常包含以下哪些层？

- A. 卷积层
- B. 池化层 (汇聚层)
- C. 全连接层
- D. 循环层

我的答案: ABC

16. (多选题, 4 分) 关于卷积核，以下描述正确的有？

- A. 它在输入数据上滑动，计算局部区域的加权和
- B. 不同的卷积核可以提取不同的特征（如边缘、纹理）
- C. 它的权重参数是在训练过程中学习得到的
- D. 它的尺寸（如 3x3）和滑动步长是超参数

我的答案: ABCD

17. (多选题, 4 分) 关于损失函数，以下说法正确的是：

- A. 用于衡量模型预测与真实标签之间的差距
- B. 损失函数的值越小，模型表现一定越好
- C. 交叉熵损失适用于分类任务
- D. 均方误差损失适用于回归任务

我的答案: ACD

18. (多选题, 4 分) 以下哪些属于前向传播过程中的步骤？

- A. 计算加权输入
- B. 应用激活函数
- C. 计算梯度

- D. 输出预测结果

我的答案: ABD

三. 简答题 (共 3 题, 48 分)

19. (简答题, 18 分) 在 <http://playground.tensorflow.org/> 上练习螺旋形数据集的分类, 训练和测试集的划分比例为 50%: 50%。请给出:

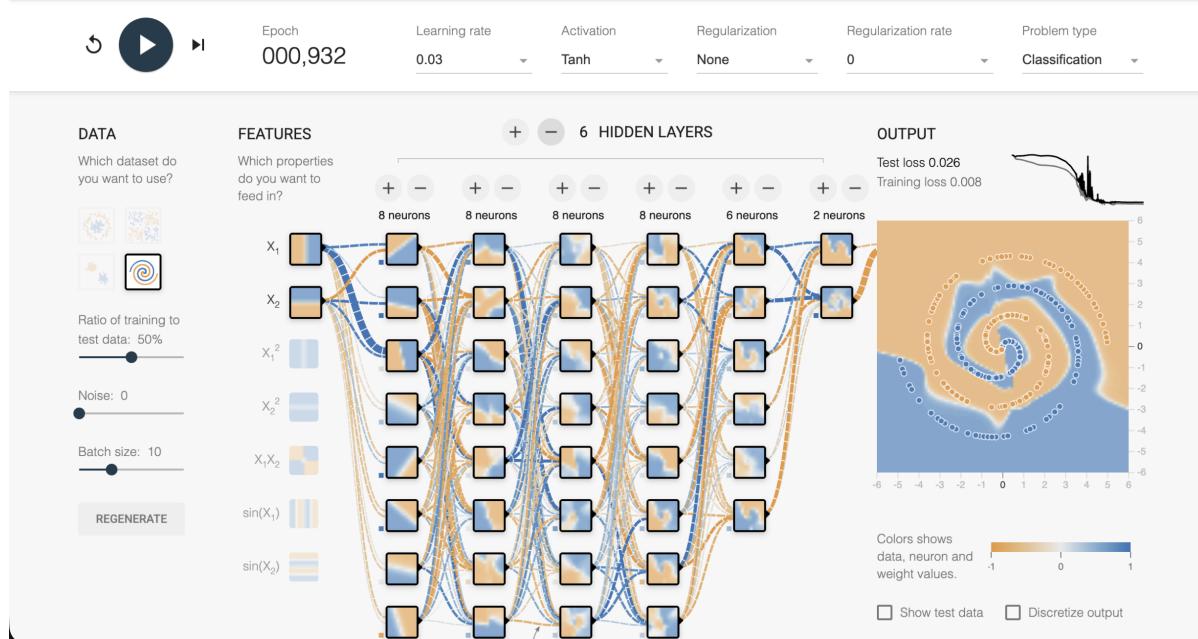
1) 你所设计的能实现正确分类的神经网络中: 隐藏层的层数、每层的神经元个数 (6 分); 2) 你所选择的激活函数, 并分析各种不同激活函数的特点 (6 分); 3) 给出最终训练好的结果图 (包含学习率、激活函数、Batch Size 参数的屏幕结果) (6 分)。

我的答案:

1) 隐藏层 6 层, 神经元个数: 8、8、8、8、6、2

2) 我选择 Tanh。Tanh 的输出范围在 -1 到 1 之间, 其零中心化的特性有助于加快收敛速度, 且能生成非常平滑的非线性决策边界, 最适合拟合螺旋形状。ReLU 虽然解决深层网络的梯度消失问题, 但在本实验有限的神经元下容易产生锯齿状边界; Sigmoid 容易引发梯度消失导致深层参数无法更新; 而 Linear 线性激活函数无法处理非线性问题, 会导致分类失败。

3)



20. (简答题, 10 分) 若对车牌图像中的每个英文字符和数字字符进行联合识别, 训练集中每张图片的大小为 28*28 像素。假设采用含有 2 个隐层 (隐层神经元个数分别为 300 和 100) 的 BP 神经网络, 隐藏层和输出层都带偏置。请问:

1) 要训练的神经网络参数量为多少? (5 分) 注意: 输出层的神经元数目请根据提示 中的说明进行设置。2) 并选择输出层的激活函数及训练的代价函数。(5 分) 注意: 勿将代价函数和损失函数混淆。

提示： 输入为各个像素点的灰度值，输出采用 one-hot 编码； 车牌中的英文字符为大写英文字符，且没有英文字母 O 和 I； 每层默认有偏置。

我的答案：

- 1) 269,034
- 2) 输出层选择 Softmax 激活函数；训练代价函数选择多分类交叉熵代价函数 (Categorical Cross-Entropy Cost)。

21. (简答题, 20 分) 假设用卷积神经网络 (1 个卷积层、1 个池化层和 1 个全连接层) 实现图像分类 (有 3 个类别)，输入为 $30 \times 30 \times 3$ 的彩色图像 (其中 3 为通道数)，其卷积层采用 10 个卷积核 (滤波器) 和 ReLU 激活函数，卷积核的大小为 5×5 ，步长为 1，用 1 层零填充，无偏置；池化层采用 2×2 的滤波器，步长为 2 的最大池化，无填充；全连接层无偏置。请问：

1) 卷积层输出为多少个大小为多少的特征图，有多少个训练参数，共多少个连接？(6 分) 2) 池化层输出大小为多少的特征图，共有多少个神经元？(4 分) 3) 全连接层有多少个神经元，多少个训练参数？(4 分) 4) 请总结卷积和池化的作用及好处。(6 分)

我的答案：

- 1) 输出 10 个大小为 28×28 的特征图。共有 750 个训练参数。共有 588,000 个连接。
- 2) 池化层输出大小为 14×14 的特征图。共有 1960 个神经元。
- 3) 全连接层有 3 个神经元 (指输出层节点)。共有 5880 个训练参数。
- 4) 卷积层的主要作用是利用局部感知和权值共享机制从输入中提取边缘、纹理等关键特征，这种方式在保证有效捕捉图像局部信息的同时，极大地减少了模型所需的训练参数和计算复杂度。

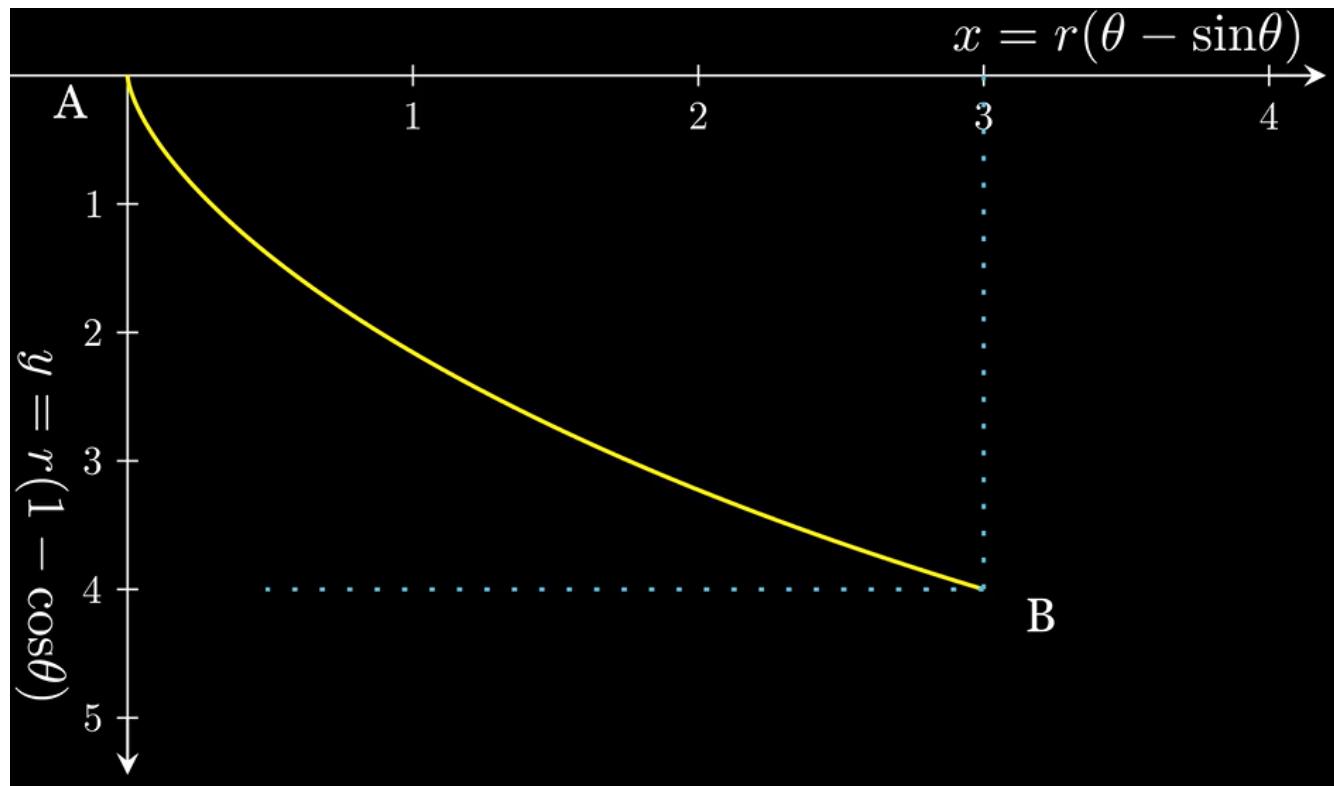
池化层的主要作用是对特征图进行下采样 (降维)，在保留最显著特征信息的前提下缩小数据规模，从而进一步减少后续层的计算量和参数数量，并增强模型对图像微小位移或形变的鲁棒性，有助于防止过拟合。

四. 其它 (共 1 题, 20 分)

22. (其它, 20 分) 用 BP 网络逼近“最速下降线”这样一个非线性函数段。此函数段端点分别为 A(0,0) 和

$$\begin{cases} x = r(\theta - \sin \theta) \\ y = r(1 - \cos \theta) \end{cases}$$

B(3,4)，且其参数方程形式如下：



请： 1) 根据万能逼近定理，设计 BP 网络的具体结构，并画出其结构图。(5 分) 2) 选择隐藏层和输出层的激活函数，并给出训练的代价函数。(5 分) 3) 根据端点联立方程，计算出 r 和 θ 的最大值（对应端点 B 处取得， $\theta \in [0, 2\pi]$ ），从而可构建训练数据。(5 分) 提示：可采用牛顿法或牛顿-拉夫逊方法求解。4) 计算 1) 所设计网络的参数量。(5 分) 5) 给出训练样本的示例，并说明欲使 1) 所设计的 BP 网络能较好地推广，即具有较好的泛化能力，一般情况下至少需要多少个训练样本？(2 分) 6) 本题的任务中，欲取得较好的效果，你采用了多少的训练样本？请构造测试样本，并绘制平均测试损失值和训练样本数量的关系图。(3 分)

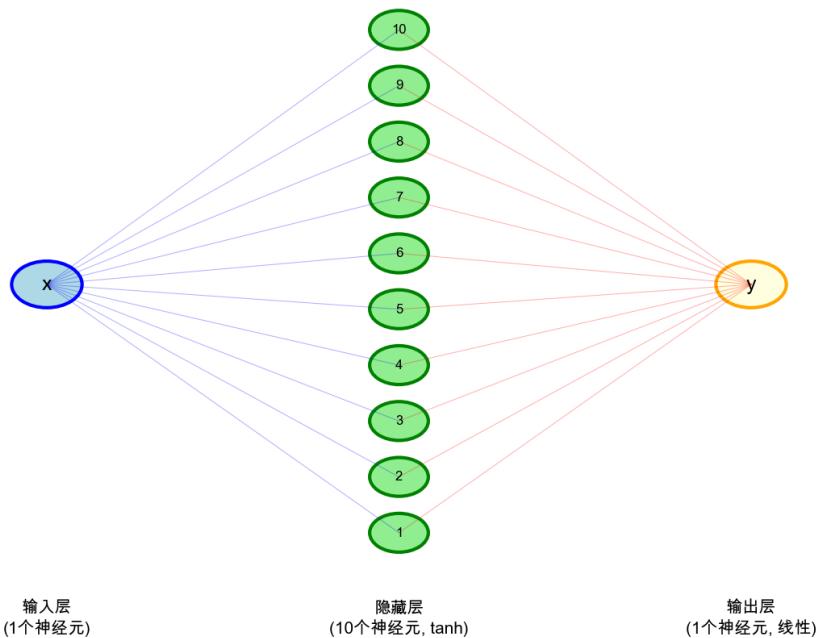
我的答案：

BP 神经网络逼近最速下降线

1. BP 网络结构设计

根据万能逼近定理，采用单隐层前馈网络即可逼近任意连续函数。网络结构为 1-10-1：输入层 1 个神经元输入 x 坐标，隐藏层 10 个神经元，输出层 1 个神经元输出 y 坐标。

BP神经网络结构图 (1-10-1)



2. 激活函数和代价函数

隐藏层用 \tanh 激活函数，输出范围 $[-1,1]$ ，满足万能逼近定理的非线性要求。输出层用线性激活，因为是回归任务需要输出连续值。代价函数用均方误差 $MSE: L = (1/N) * \sum(y_i - \hat{y}_i)^2$

3. 求解 r 和 $_max$

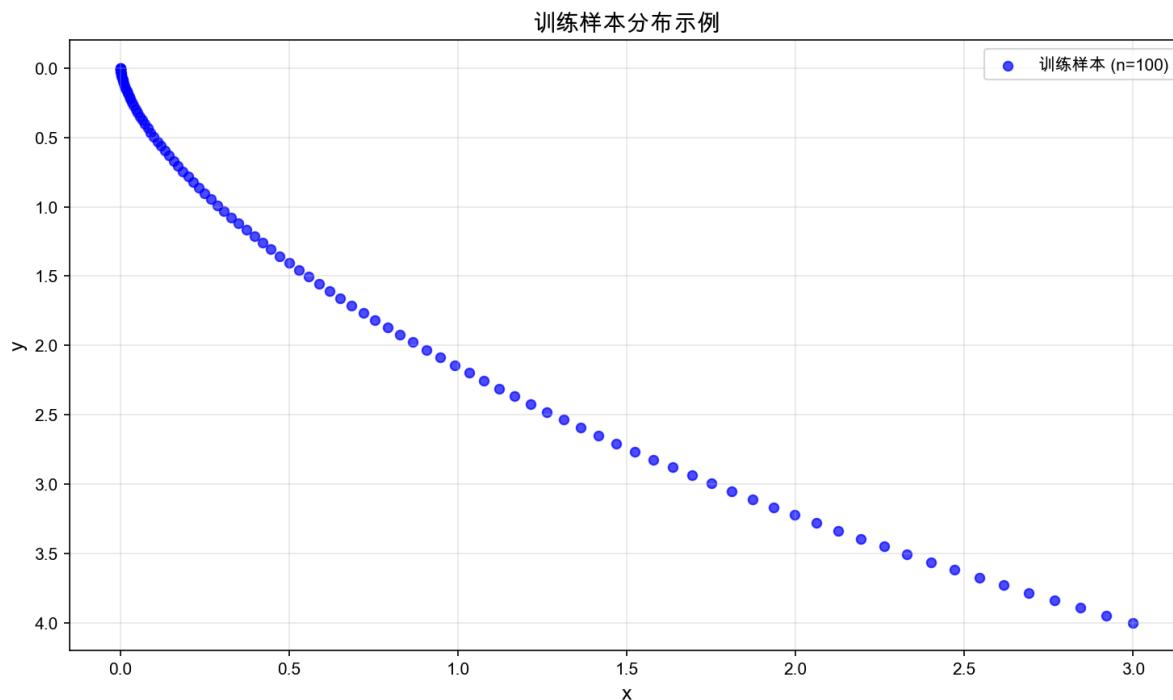
联立端点 $B(3,4)$ 处的方程: $x = r(-\sin\theta) = 3$ $y = r(1 - \cos\theta) = 4$ 用牛顿-拉夫逊法迭代求解，得到: $r = 1.074935$ $_max = 2.782735$ rad (约 159.45°) 验证: 代入验算 $x=3.000000$, $y=4.000000$, 正确。

4. 网络参数量

隐藏层权重 $W1: 1 \times 10 = 10$ 个
隐藏层偏置 $b1: 10$ 个
输出层权重 $W2: 10 \times 1 = 10$ 个
输出层偏置 $b2: 1$ 个
总共: $10 + 10 + 10 + 1 = 31$ 个参数

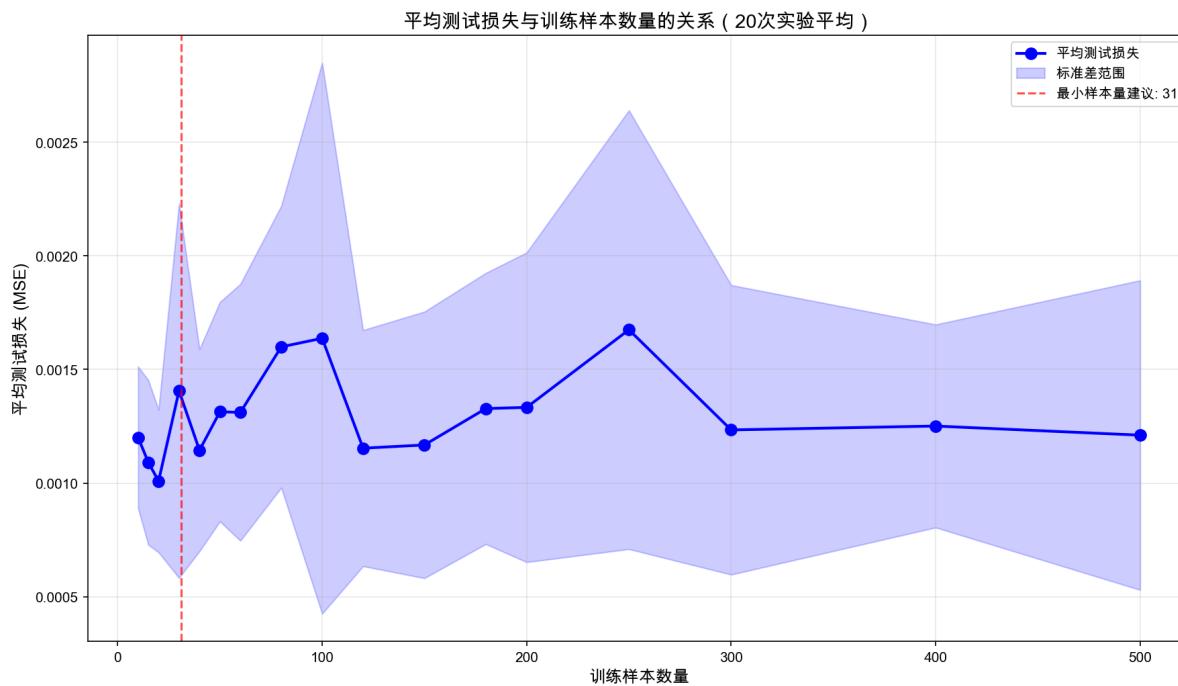
5. 训练样本与泛化能力

训练样本示例（均匀采样生成）: $(0.0000, 0.0000), (0.0001, 0.0051), (0.0002, 0.0091) \dots$ 为保证泛化能力，训练样本数应满足 $N \geq \text{参数量} + 1$ ，即至少需要 32 个样本。

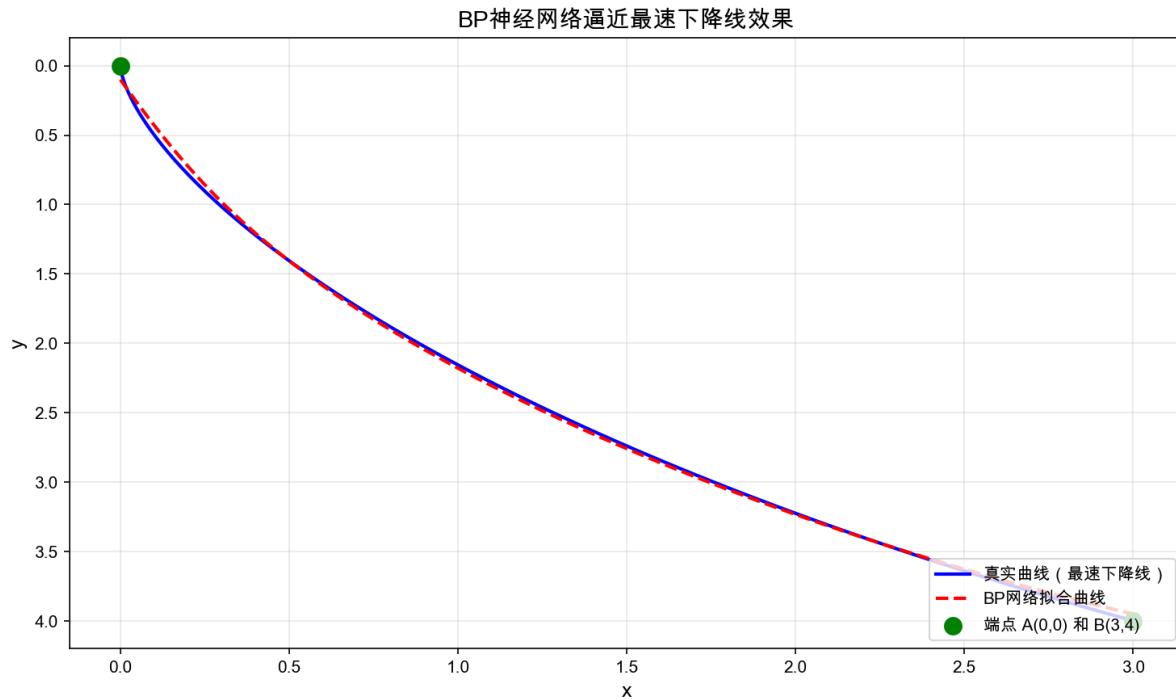


6. 测试损失与样本数量关系

实验中采用 100 个训练样本。测试时对不同样本数量各做 20 次实验取平均。



拟合效果



lec19-不确定推理.md

lec19-作业-不确定推理

题量: 2 | 满分: 100

一. 简答题 (共 2 题, 100 分)

1. (简答题, 50 分) 模糊推理习题

设 $\mathbf{U} = \mathbf{V} = \{1, 2, 3, 4, 5\}$, 且有如下推理规则 \mathbf{R}_1 :

IF x is 冷 THEN y is 高

其中, “冷”与“高”分别是 \mathbf{U} 与 \mathbf{V} 上的模糊集合, 设:

$$\text{冷} = \frac{0.9}{1} + \frac{0.7}{2} + \frac{0.4}{3}$$

$$\text{高} = \frac{0.3}{2} + \frac{0.7}{3} + \frac{0.9}{4} + \frac{1}{5}$$

已知事实为 x is 相当冷, “相当冷”的模糊集合为:

$$\text{相当冷} = \frac{0.85}{1} + \frac{0.59}{2} + \frac{0.25}{3}$$

要求: 请用推理规则和已知事实求出模糊结论 (要求使用 Mamdani 推理)。

我的答案:

1. 向量表示

将论域 $U = V = \{1, 2, 3, 4, 5\}$ 上的模糊集合表示为向量:

- $A(\text{冷}) = (0.9, 0.7, 0.4, 0, 0)$
- $B(\text{高}) = (0, 0.3, 0.7, 0.9, 1)$
- $A'(\text{相当冷}) = (0.85, 0.59, 0.25, 0, 0)$

2. 建立模糊关系矩阵 R

根据 Mamdani 推理规则 (取小运算 $R(i, j) = \min(A(i), B(j))$), 矩阵 R 为:

$$R = \begin{bmatrix} 0 & 0.3 & 0.7 & 0.9 & 0.9 \\ 0 & 0.3 & 0.7 & 0.7 & 0.7 \\ 0 & 0.3 & 0.4 & 0.4 & 0.4 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

(注: 第一行由 $\min(0.9, B)$ 得到, 第二行由 $\min(0.7, B)$ 得到, 以此类推)

3. 合成推理

根据最大-最小合成规则 ($B' = A' \circ R$) 进行计算:

$$B'(y) = \bigvee_{x \in U} (A'(x) \wedge R(x, y))$$

逐个计算 y 的隶属度:

- $y=1$: $\max(\min(0.85, 0), \min(0.59, 0), \min(0.25, 0), \dots) = 0$
- $y=2$: $\max(\min(0.85, 0.3), \min(0.59, 0.3), \min(0.25, 0.3)) = \max(0.3, 0.3, 0.25) = 0.3$
- $y=3$: $\max(\min(0.85, 0.7), \min(0.59, 0.7), \min(0.25, 0.4)) = \max(0.7, 0.59, 0.25) = 0.7$
- $y=4$: $\max(\min(0.85, 0.9), \min(0.59, 0.7), \min(0.25, 0.4)) = \max(0.85, 0.59, 0.25) = 0.85$
- $y=5$: $\max(\min(0.85, 0.9), \min(0.59, 0.7), \min(0.25, 0.4)) = \max(0.85, 0.59, 0.25) = 0.85$

4. 结论

模糊结论为:

$$B' = \frac{0.3}{2} + \frac{0.7}{3} + \frac{0.85}{4} + \frac{0.85}{5}$$

2. (简答题, 50 分) 证据理论习题

已识别框架 $U = \{\text{目标 A, 背景 B, 诱饵 C}\}$ 。对此识别框架的红外传感器和毫米波传感器, 假设已获得三个测量周期的后验可信度分配数据 (即基本概率分配函数) 如表所示。 (备注: “不明” 即全集 {目标 A, 背景 B, 诱饵 C})

周期	红外: 目标 A	红外: 背景 B	红外: 诱饵 C	毫米波: 目标 A	毫米波: 背景 B	毫米波: 诱饵 C	毫米波: 不明
第一周期	0.6	0.2	0.2	0.4	0.2	0.3	0.1
第二周期	0.6	0.25	0.15	0.3	0.1	0.2	0.2
第三周期	0.7	0.15	0.15	0.3	0.25	0.3	0.15

要求: 基于 D-S 证据理论, 计算传感器系统第一个周期对目标 A、背景 B、诱饵 C 的融合可信度函数 $M(A)$ 、 $M(B)$ 和 $M(C)$ 。

我的答案:

1. 设定变量

- 红外传感器概率分配为 m_1
- 毫米波传感器概率分配为 m_2

第一周期数据：

- $m_1(A) = 0.6, m_1(B) = 0.2, m_1(C) = 0.2$
- $m_2(A) = 0.4, m_2(B) = 0.2, m_2(C) = 0.3, m_2(U) = 0.1$

2. 计算冲突系数 K

$$K = \sum_{X \cap Y = \emptyset} m_1(X)m_2(Y)$$

$$\begin{aligned} K &= m_1(A)m_2(B) + m_1(A)m_2(C) \\ &\quad + m_1(B)m_2(A) + m_1(B)m_2(C) \\ &\quad + m_1(C)m_2(A) + m_1(C)m_2(B) \\ &= (0.6 \times 0.2) + (0.6 \times 0.3) \\ &\quad + (0.2 \times 0.4) + (0.2 \times 0.3) \\ &\quad + (0.2 \times 0.4) + (0.2 \times 0.2) \\ &= 0.12 + 0.18 + 0.08 + 0.06 + 0.08 + 0.04 \\ &= 0.56 \end{aligned}$$

3. 计算归一化系数

$$1 - K = 1 - 0.56 = 0.44$$

4. 计算融合可信度 M

利用 Dempster 合成规则： $M(Z) = \frac{1}{1-K} \sum_{X \cap Y = Z} m_1(X)m_2(Y)$

计算 $M(A)$:

$$\begin{aligned} M(A) &= \frac{m_1(A)m_2(A) + m_1(A)m_2(U)}{1 - K} \\ &= \frac{0.24 + 0.06}{0.44} = \frac{0.30}{0.44} \approx 0.6818 \end{aligned}$$

计算 $M(B)$:

$$\begin{aligned} M(B) &= \frac{m_1(B)m_2(B) + m_1(B)m_2(U)}{1 - K} \\ &= \frac{0.04 + 0.02}{0.44} = \frac{0.06}{0.44} \approx 0.1364 \end{aligned}$$

计算 $M(C)$:

$$\begin{aligned}M(C) &= \frac{m_1(C)m_2(C) + m_1(C)m_2(U)}{1 - K} \\&= \frac{0.06 + 0.02}{0.44} = \frac{0.08}{0.44} \approx 0.1818\end{aligned}$$

lec22-深度学习和智能计算.md

lec22-作业-深度学习和智能计算

题量: 16 | 满分: 120 | 作答时间: 01-05 11:03 至 01-19 23:55

一. 单选题 (共 8 题, 32 分)

1. (单选题, 4 分) 启发信息的分类中, 哪一种启发信息用于更准确、更精炼地描述状态?

- A. 过程性启发信息
- B. 控制性启发信息
- C. 陈述性启发信息
- D. 估价函数

我的答案: C: 陈述性启发信息

2. (单选题, 4 分) 在启发式图搜索策略中, A* 算法的主要难点是什么?

- A. 状态空间的表示
- B. 启发函数的确定
- C. 节点的扩展顺序
- D. 目标状态的定义

我的答案: B: 启发函数的确定

3. (单选题, 4 分) 在启发信息的分类中, 控制性启发信息主要用于什么?

- A. 描述状态
- B. 生成后继节点
- C. 构造操作算子
- D. 表示控制策略的知识

我的答案: D: 表示控制策略的知识

4. (单选题, 4 分) 以下哪种算法是遗传算法的改进算法之一?

- A. 粒子群优化算法
- B. 双倍体遗传算法
- C. 蚁群算法
- D. 神经网络

我的答案: B: 双倍体遗传算法

5. (单选题, 4 分) 遗传算法的选择概率是如何计算的?

- A. 根据适应度计算出来的
- B. 事先设计好的, 取决于排序
- C. 随机选择

- D. 根据个体的年龄

我的答案: A: 根据适应度计算出来的

6. (单选题, 4 分) 遗传算法的主要不足之一是什么?

- A. 计算速度慢
- B. 早熟/停滞: 局部收敛
- C. 适应性差
- D. 需要大量的初始数据

我的答案: B: 早熟/停滞: 局部收敛

7. (单选题, 4 分) 群智能算法的产生背景主要是为了克服什么问题?

- A. 数据处理能力不足
- B. 传统算法的局限性
- C. 计算资源的浪费
- D. 人工智能的复杂性

我的答案: B: 传统算法的局限性

8. (单选题, 4 分) 以下关于 RNN、LSTM 和 Transformer 网络的描述中, 哪一项是正确的?

- A. RNN 和 LSTM 都无法解决长期依赖问题, 因此需要使用 Transformer 来处理。
- B. LSTM 通过引入门机制 (如输入门、遗忘门和输出门) 改善了 RNN 在长期依赖问题上的表现。
- C. Transformer 网络是完全基于循环结构设计的, 用于提高序列数据的处理能力。
- D. 与 LSTM 和 RNN 相比, Transformer 更适合处理小规模的时间序列数据。

我的答案: B:LSTM 通过引入门机制 (如输入门、遗忘门和输出门) 改善了 RNN 在长期依赖问题上的表现。

二. 多选题 (共 4 题, 20 分)

9. (多选题, 5 分) 以下哪些是启发式图搜索策略的特点?

- A. 重排开放表
- B. 选择最有希望的节点扩展
- C. 采用盲目搜索策略
- D. 使用启发信息简化搜索过程

我的答案: ABD: 重排开放表; 选择最有希望的节点扩展; 使用启发信息简化搜索过程

10. (多选题, 5 分) 在 A* 搜索算法中, 以下哪些因素会影响其性能?

- A. 启发函数的选择
- B. 状态空间的大小
- C. 估价函数的准确性

- D. 搜索策略的实现方式

我的答案: ACD: 启发函数的选择; 估价函数的准确性; 搜索策略的实现方式

11. (多选题, 5 分) 在遗传算法中, 以下哪些因素可能导致早熟或停滞现象?

- A. 选择压力过大
- B. 变异率过低
- C. 种群多样性不足
- D. 适应度函数设计不合理

我的答案: ABCD: 选择压力过大; 变异率过低; 种群多样性不足; 适应度函数设计不合理

12. (多选题, 5 分) 以下哪些是遗传算法的基本步骤?

- A. 初始化种群
- B. 选择
- C. 变异
- D. 归纳

我的答案: ABC: 初始化种群; 选择; 变异

三. 简答题 (共 3 题, 48 分)

13. (简答题, 16 分) 对 10 个城市的 TSP 问题采用部分匹配交叉, 如果父代 1 和父代 2 如下所示, 其中 | 为交叉位, 请给出交叉后的子代 1 和子代 2。父代 1: 9 0 3 5 | 8 1 2 | 7 4 6 父代 2: 0 8 3 9 | 5 7 1 | 4 2 6

我的答案:

子代 1: 9 0 3 8 | 5 7 1 | 2 4 6 子代 2: 0 5 3 9 | 8 1 2 | 4 7 6

14. (简答题, 16 分) 遗传算法求解非线性函数最小值用遗传算法求解函数 $f(x) = \sin(2x) + 0.1x^2$ 的最小值, 定义域为 $x \in [0, \pi]$ 。请回答以下三个问题:

(1) 设计一个初始种群包含 3 个个体, 并用 5 位二进制编码表示。(2) 计算每个个体的目标函数值和适应度值 (适应度值定义为 $fitness(x) = 1/(f(x) + 1)$)。(3) 假设在交叉操作中选择了两个个体, 使用单点交叉 (交叉点为第 3 位) 生成下一代。

我的答案:

(1) 初始种群设计

设计种群包含 3 个个体, 采用 5 位二进制编码, 随机生成的初始种群如下:

- 个体 1: 01010
- 个体 2: 10101
- 个体 3: 11110

(2) 计算目标函数值与适应度值

解码公式为 $x = 0 + (\text{十进制值}/31) \times 3.1416$

计算结果汇总如下：

个体 1 (01010)：

- 十进制值 = 10
- 实数值 $x = 1.0134$
- 目标函数值 $f(x) = 1.001$
- 适应度值 = 0.4998

个体 2 (10101)：

- 十进制值 = 21
- 实数值 $x = 2.1282$
- 目标函数值 $f(x) = -0.449$
- 适应度值 = **1.8149 (最优)**

个体 3 (11110)：

- 十进制值 = 30
- 实数值 $x = 3.0402$
- 目标函数值 $f(x) = 0.726$
- 适应度值 = 0.5794

(3) 单点交叉操作

选择个体 1 和个体 2 作为父代，在第 3 位后进行单点交叉。

交叉前父代：

- 父代 1: 010 | 10
- 父代 2: 101 | 01

交叉操作：交换交叉位后的基因片段。

交叉后子代：

- 子代 1: 01001 (由 010 + 01 组成)
- 子代 2: 10110 (由 101 + 10 组成)

15. (简答题, 16 分) 设有如图所示的博弈树（方框表示 MAX 节点，圆框表示 MIN 节点），其中最下面的数字为假设的估值，请对该博弈树利用 $\alpha - \beta$ 剪枝技术剪去不必要的分枝，并回答以下问题：

(1) A 节点的最终值是？(2) 产生了几次剪枝？哪些节点被剪掉了？分别是什么类型的剪枝？

我的答案：

(1) -19

(2) 共产生了 3 次剪枝。

- 剪掉节点 I: 属于 Alpha 剪枝。
- 剪掉节点 X: 属于 Alpha 剪枝。
- 剪掉节点 ZC: 属于 Beta 剪枝。

四. 计算题 (共 1 题, 20 分)

16. (计算题, 20 分) 如上图所示, ViT (Vision Transformer) 是 Google 研究团队在 2020 年发表的开创性论文《An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale》中提出的模型。它首次成功地将原本在自然语言处理领域大放异彩的 Transformer 架构直接、纯粹地应用于图像分类任务, 并取得了与当时最先进的卷积神经网络相媲美甚至更优的性能, 从而彻底改变了计算机视觉的研究范式。

以 ViT-Base/14 模型为例, 用于 ImageNet 分类任务时, 输入图像大小为 $224 \times 224 \times 3$, 其关键配置为:

1) Patch Size (上图左方分块图像每块的大小): 14×14 2) 隐藏维度 (embedding dimension): $d = 768$ 3) Transformer Encoder 层数: $L = 12$ 4) Attention 头数: $h = 12$ 5) FFN (上图右方的 MLP) 中间层维度: 3072 6) 分类类别数: 1000

假设条件:

1) Patch Embedding (上图左方的图像分块) 使用一个 Conv2D 实现: kernel size 为 14×14 , stride 为 14, 输入通道数为 3, 输出通道数 (卷积个数) 为 768 2) 所有线性层均包含偏置 3) LayerNorm 包 (上图右方的 Norm) 含可学习参数 4) 每个 Transformer Encoder Block 包含: 多头自注意力 (上图右方的 Multi-Head Attention)、两层前馈网络 (上图右方的 MLP)、两个 LayerNorm (上图右方的 Norm) 5) 忽略 Dropout、激活函数带来的参数 6) ViT 使用一个可学习的 CLS Token 和位置嵌入 7) 分类头是一个线性层

问题:

请在给出详细计算过程的情况下, 回答以下问题:

(1) Patch Embedding 层的参数量是多少? (3 分) (2) Position Embedding 和 CLS Token 的参数量是多少? (4 分) (3) 单个 Transformer Encoder Block 中各部分的参数量分别是多少? 包括: Multi-Head Self-Attention、MLP (FFN)、LayerNorm, 并给出单个 Block 的总参数量 (6 分) (4) 12 个 Transformer Encoder Blocks 的参数量总和是多少? (2 分) (5) 分类头 (MLP Head) 的参数量是多少? (2 分) (6) ViT-Base/14 整个网络的总参数量是多少? (3 分)

我的答案:

(1) Patch Embedding 层的参数量

使用 Conv2D 实现, kernel=14 × 14, 输入通道 =3, 输出通道 =768

- 权重: $14 \times 14 \times 3 \times 768 = 451,584$
- 偏置: 768
- 答: 452,352

(2) Position Embedding 和 CLS Token 的参数量

- Patch 数量: $(224/14) \times (224/14) = 16 \times 16 = 256$
- 序列长度: $256 + 1(CLSS) = 257$
- CLS Token: 768
- Position Embedding: $257 \times 768 = 197,376$
- 答: **768 + 197,376 = 198,144**

(3) 单个 Transformer Encoder Block 参数量

LayerNorm (×2): 每个 LN 有 α 和 β , 各 768 维 $2 \times (768 + 768) = 3,072$

Multi-Head Self-Attention: W_Q, W_K, W_V, W_O 各为 768×768

- 权重: $4 \times 768 \times 768 = 2,359,296$
- 偏置: $4 \times 768 = 3,072$
- 小计: 2,362,368

MLP (FFN):

- 第一层 $768 \rightarrow 3072$: $768 \times 3072 + 3072 = 2,362,368$
- 第二层 $3072 \rightarrow 768$: $3072 \times 768 + 768 = 2,360,064$
- 小计: 4,722,432

答: $3,072 + 2,362,368 + 4,722,432 = \mathbf{7,087,872}$

(4) 12 个 Transformer Encoder Blocks 的参数量总和

答: $12 \times 7,087,872 = \mathbf{85,054,464}$

(5) 分类头 (MLP Head) 的参数量

线性层 $768 \rightarrow 1000$

- 权重: $768 \times 1000 = 768,000$
- 偏置: 1000
- 答: **769,000**

(6) ViT-Base/14 整个网络的总参数量

- Patch Embedding: 452,352
- Pos Embed + CLS: 198,144
- $12 \times$ Encoder Block: 85,054,464
- 分类头: 769,000

分步求和:

- Embedding 部分: $452,352 + 198,144 = 650,496$
- Backbone + Head: $85,054,464 + 769,000 = 85,823,464$

答: $650,496 + 85,823,464 = \mathbf{86,473,960}$ (约 86.47M)