微专业背景下《机器学习》课程实验大纲设计

刘永明 赵转哲 刘志博 鲁月林 倪佳佳
安徽工程大学人工智能学院，安徽 芜湖 241000

[摘要]近年来，随着人工智能技术的快速发展，人工智能工程师和数据科学家的需求量也越来越大。因此，作为人工智能技术核心技术的机器学习微专业课程和实验也变得越来越重要。但传统的机器学习课程往往存在教学内容过于抽象、理论与实践脱节等问题，学生的学习效果和应用能力不够理想。因此，本论文旨在设计一个机器学习课程的实验大纲，以提高学生对机器学习的理解与应用。

[关键词]机器学习课程设计；实验课程；实验大纲设计
DOI：10.33142/fme.v6i1.14978 中图分类号：G642 文献标识码：A

Design of Experimental Syllabus for Machine Learning Course under the Background of Micro Major

LIU Yongming, ZHAO Zhuanzhe, LIU Zhibo, LU Yuelin, NI Jiajia
School of Artificial Intelligence, Anhui Polytechnic University, Wuhu, Anhui, 241000, China

Abstract: In recent years, with the rapid development of artificial intelligence technology, the demand for artificial intelligence engineers and data scientists has also been increasing. Therefore, as the core technology of artificial intelligence, machine learning micro professional courses and experiments have become increasingly important. However, traditional machine learning courses often have problems such as overly abstract teaching content and a disconnect between theory and practice, resulting in less than ideal learning outcomes and application abilities for students. Therefore, this paper aims to design an experimental outline for a machine learning course to enhance students' understanding and application of machine learning.

Keywords: machine learning course design; experimental courses; experimental outline design

1 概述

机器学习已成为当前具有代表性的人工智能技术研究课题；综合涉及人工智能、数学、机械科学、电子学、控制论、计算机技术等多个学科领域[1]。机器学习主要是指通过系统或者知识识别；对机械学习能力进行科学化提升；进而获得新技能、新知识。如果不通过系统学习或未能掌握合适的学习方法可能难以掌握全新的问题分析、解决方法；所以机器学习要做到不断创新、不断发展；它正迎合了当前人工智能发展领域的快速发展要求[2]。机器学习的研究目的；是希望计算机具有像人类一样从现实世界获取知识的能力；同时；建立学习的计算理论；构造各种学习系统并应用到各个领域中去[3]。

1.1 机器学习中存在的问题

（1）数学理论与实践的脱节：由于机器学习涉及很多数学理论和算法，如线性代数、概率论、最优化方法等，因此在整体授课过程中比较枯燥，调动学生情绪比较困难[4]。并且有些教学内容可能过于理论化，在教学中缺乏与实际应用的关联，导致学生难以将理论知识应用到实际问题解决中。

（2）缺乏实际案例和项目：机器学习是一门实践驱动的学科，但在大学课堂中，可能缺乏实际案例和项目的引入。学生只通过理论的学习，难以真正理解机器学习在实际中的应用场景和解决方法。

（3）缺乏实践操作和编程能力培养：机器学习主要致力于计算机复杂计算方面，包括数据处理、图像视觉处理、与语音识别等领域，其核心都是机器学习基础算法[5-7]。机器学习算法通常需要使用编程语言进行实现和应用，如 Python、Matlab 等。然而，在大学课堂中，可能缺乏足够的实践操作和编程能力的培养，使学生难以掌握实际问题解决的技能。

1.2 实验课程的重要性

实验课程主要培养学生独立进行机器学习应用的实际工作能力，在认识问题、分析问题和解决问题等方面受到较全面的训练。在实验课上，学生应当是实验的执行者，教师作为引导者的主要任务是确定实验目标、制定实验计划与进行实验考核[8]。通过对本课程的学习使学生系统地掌握机器学习应用的相关知识，对机器学习过程中的常见问题及解决方法有一定了解。为今后从事人工智能系统开发与设计打下基础。

1.3 实验教学目标

（1）针对设计任务，通过文献检索和其他途径来收集和整理相关资料。使用学术数据库、图书馆资源、专业

Copyright © 2025 by authors and Viser Technology Pte. Ltd.

期刊等渠道，进行关键词检索并筛选出与课题相关的文献。通过阅读和分析这些文献，结合工程原理和现有研究，可以对设计任务进行合理的分析。

（2）根据设计任务的需求，综合运用基础理论知识和文献资料，提出合理的总体设计方案。这包括设计各组成模块，并对设计方案进行优选比较和论证其可行性。通过综合考虑各种因素，例如成本、性能、可靠性等，来选择最合适的设计方案。

（3）运用设计资料手册和使用数据进行估算和处理进行机器学习。可以借助机器学习算法，使用设计资料手册提供的数据对模型进行训练和测试，在设计过程中运用机器学习算法进行数据分析、模型评估以及结果呈现。此外，可以利用仿真工具软件对机器学习结果进行分析，以确保结果的合理性和可靠性。

（4）能够将设计内容以口头和文稿形式准确表达并回应指令。

2 实验内容

本实验总学时数为 14 学时，具体实验内容包括：线性模型、决策树模型、贝叶斯模型支持向量机模型、聚类分析、主成分分析、EM 算法。

本实验大纲以 EM 算法为例，进行实验内容设计。EM 算法是进行极大似然估计的一种有效方法，主要应用于以下两种非完全数据参数估计：第一，观测数据不完全；第二，似然函数不是解析的，或者似然函数的表达式过于复杂而导致极大似然函数的传统估计方法失效[9]。

2.1 实验目的

熟悉和掌握随机数的产生方法、掌握 EM 算法。

2.2 实验原理

问题描述：假设有两枚硬币分别标记为 A 和 B，做 5 轮随机抛掷实验，每轮抛掷 10 次，实验时每轮投掷的硬币标记 A 或 B 的信息丢失，只记录投掷硬币出现正反面的结果，实验结果为：

第一轮（来自硬币 A 或 B 的信息丢失）：5 次正面、5 次数反面。

第二轮（来自硬币 A 或 B 的信息丢失）：9 次正面、1 次数反面。

第三轮（来自硬币 A 或 B 的信息丢失）：8 次正面、2 次数反面。

第四轮（来自硬币 A 或 B 的信息丢失）：4 次正面、6 次数反面。

第五轮（来自硬币 A 或 B 的信息丢失）：7 次正面、3 次数反面。

估计两枚硬币抛出正面的概率 θ_A和 θ_B

该问题包含一个隐变量 $z=(z_1, z_2, z_3, z_4, z_5)$，代表每一轮所使用的硬币标记，要想估计两枚硬币抛出正面的概率 θ_A和 θ_B，我们需要知道每一轮抛掷所使用的硬币是 A 还是 B，这样才能估计 θ_A和 θ_B的值，但是估计隐变量 z 我们又需要知道 θ_A和 θ_B的值，才能用极大似然估计法去估计出 z。

其解决方法就是先随机初始化 θ_A和 θ_B，然后估 z，然后基 z 按照最大似然概率估计新的 θ_A和 θ_B，循环至收敛。

EM 算法目标：EM 算法解决的问题是极大化含有隐变量 z 的观察数据(不完全数据)X 关于参数 θ 的对数似然，即极大化工 $L(\theta)=\log P(X|\theta)=\log\sum_z\Sigma P(X,z|\theta)$。

EM 算法的核心思想非常简单，分为两步：Expection-Step 和 Maximization-Step。E-Step 主要通过观察数据和现有模型来估计参数，然后用这个估计的参数值来计算似然函数的期望值;而 M-Step 是寻找似然函数最大化时对应的参数。由于算法会保证在每次迭代之后似然函数都会增加，所以函数最终会收敛。

设隐含变量 z，X 为可直接观测数据，θ 为模型参数向量。我们无法直接得知 θ 估计的似然函数 L(θ | X)。可直接得知参数 θ 给定情况下 X 和 z 的联合概率分布 p(X,z | θ)及参数向量和可观测数据给定情况下 Z 取值状态的条件概率分布 p(z | X, θ)。

（1）离散问题 EM 算法：给定数据集，假设样本间相互独立，我们想要拟合模型 p(x; θ)到数据的参数。根据分布我们可以得到如下似然函数：

$$L(\theta)=\sum_{i=1}^{n}logp(x_i;\theta)=\sum_{i=1}^{n}log\sum_z p(x_i,z;\theta) \quad (1)$$

对于每个样本 i，我们用 $Q_i(z)$表示样本 i 隐含变量 z 的某种分布，且 $Q_i(z)$满足条件($\sum_z^Z Q_i(z)=1$, $Q_i(z)\geq 0$)。

我们将上面的式子做以下变化：

$$\begin{aligned}\sum_i^n logp(x_i;\theta)&=\sum_i^n log\sum_z p(x_i,z;\theta)\\&=\sum_i^n log\sum_z^Z Q_i(z)\frac{p(x_i,z;\theta)}{Q_i(z)}\\&\geq\sum_i^n\sum_z^Z Q_i(z)log\frac{p(x_i,z;\theta)}{Q_i(z)}\end{aligned} \quad (2)$$

上面式子中，第一步是求和每个样本的所有可能的类别 Z 的联合概率密度函数，但是这一步直接求导非常困难，所以将其分母都乘以函数 $Q_i(z)$，转换到第二步。从第二步到第三步是利用 Jensen 不等式。

我们来简单证明下：

Jensen 不等式给出：如果 f 是凹函数，X 是随机变量，则 E[f(X)]≤f(E[X])，当 f 严格是凹函数时，则 E[f(X)]<f(E[X])，凸函数反之。当 X=E[X]时，即为常数时等式成立。

我们把第一步中的 log 函数看成一个整体，由于 log(x)的二阶导数小于 0，所以原函数为凹函数。我们把

Copyright © 2025 by authors and Viser Technology Pte. Ltd.

$Q_i(z)$看成一个概率 p_z，把 $\frac{p(x_i,z;\theta)}{Q_i(z)}$ 看成 z 的函数 g(z)。根据期望公式有：

$$E(z)=p_z g(z)=\sum_z^Z Q_i(z)[\frac{p(x_i,z;\theta)}{Q_i(z)}] \tag{3}$$

根据 Jensen 不等式的性质：$f(\sum_z^Z Q_i(z)[\frac{p(x_i,z;\theta)}{Q_i(z)}])=f(E[z])\geq E[f(z)]=\sum_z^Z Q_i(z)f([\frac{p(x_i,z;\theta)}{Q_i(z)}])$ 证明结束。

由此可得对数似然的下界函数：$B(\theta)=\sum_i^n\sum_z^Z Q_i(z)log\frac{p(x_i,z;\theta)}{Q_i(z)}$。

令 $Q_i(z)=p(z|x_i,\theta_t)$，下界函数可转化为：$B(\theta,\theta_t)=\sum_i^n\sum_z^Z Q_i(z)log\frac{p(x_i,z;\theta)}{Q_i(z)}=\sum_i^n\sum_z^Z Q_i(z)logp(x_i,z;\theta)-\sum_i^n\sum_z^Z Q_i(z)logQ_i(z)$

略去下界函数 B(θ， θ_t)中与待求参数向量 θ 无关的项得到如下 Q 函数：

$$Q(\theta,\theta_t)=\sum_i^n\sum_z^Z Q_i(z)logp(x_i,z;\theta) \tag{4}$$

（$Q(\theta,\theta_t)$ 隐含变量 z 的函数 L(θ |x_i, z)在概率分布 $p(z|x_i,\theta_t)$ 下的数学期望）

由于 B(θ， θ_t)≤ InL(θ |X)，故可通过迭代选取不同下界函数 B(θ， θ_t)最大值的方式逐步逼近对数似然 InL(θ |X)的最大值。

目前有两个问题：

（1）什么时候下界 B(θ， θ_t)与 L(θ)相等?

（2）为什么一定会收敛?

首先第一个问题，当 X =E [X]时，即为常数时等式成立：

$$\frac{p(x_i,z;\theta)}{Q_i(z)}=c \tag{5}$$

做如下变换：$\sum_z p(x_i,z;\theta)=\sum_z Q_i(z)c$

其中 $\sum_z Q_i(z)=1$，所以可以推导出：

$$\sum_z p(x_i,z;\theta)=c \tag{6}$$

因此得到了：

$$\begin{aligned}Q_i(z)&=\frac{p(x_i,z;\theta)}{\sum_z p(x_i,z;\theta)}\\&=\frac{p(x_i,z;\theta)}{p(x_i;\theta)}\\&=p(z|x_i;\theta)\end{aligned} \tag{7}$$

至此，我们推出了在固定参数下，使下界拉升的 Q(z)的计算公式就是后验概率，同时解决了 Q(z)如何选择的问题。这就是我们刚刚说的 EM 算法中的 E-Step，目的是建立 L(θ)的下界接下来得到 M-Step 目的是在给定 Q(z)后调整 θ，从而极大化似然函数 L(θ)的下界 J(z, Q)。

对于第二个问题，为什么一定会收敛?

这边简单说一下，因为每次 θ 更新时(每次迭代时)，都可以得到更大的似然函数，也就是说极大似然函数时单调递增，那么我们最终就会得到极大似然估计的最大值。

但是要注意，迭代一定会收敛，但不一定会收敛到真实的参数值，因为可能会陷入局部最优。所以 EM 算法的结果很受初始值的影响。

EM 的应用有很多，比如混合高斯模型、聚类、HMM 等等。

（2）连续问题 EM 算法

假设联合概率 p(X, Z | θ)可直接得知，故似然函数可转化为：

$$L(\theta|X)=\int p(X,Z|\theta)dZ \tag{8}$$

对数似然为：

$$\ln L(\theta|X)=\ln\int p(X,Z|\theta)dZ \tag{9}$$

确定上式最大值即可求得参数向量的最大似然估计，由于 Z 的存在，最大值无法直接求得。

对数似然函数变形为：

$$\ln L(\theta|X)=\ln\int\frac{p(X,Z|\theta)}{p(Z)}p(Z)dZ \tag{10}$$

其中 p(Z)为隐含数据 Z 的某一分布。由于对数函数为上凸函数，故成立如下不等式：Jensen 不等式。

$$\ln L(\theta|X)=\ln\int\frac{p(X,Z|\theta)}{p(Z)}p(Z)dZ\geq\int\ln\frac{p(X,Z|\theta)}{p(Z)}p(Z)dZ \tag{11}$$

由此可得对数似然的下界函数：

$$B(\theta)=\int\ln\left[\frac{p(X,Z|\theta)}{p(Z)}\right]p(Z)dZ \tag{12}$$

令 $p(Z)=p(Z|X,\theta_t)$，下界函数可转化为：

$$\begin{aligned}B(\theta,\theta_t)&=\int\ln\left[\frac{p(X,Z|\theta)}{p(Z)}\right]p(Z)dZ=\int\ln\left[\frac{p(X,Z|\theta)}{p(Z|X,\theta_t)}\right]p(Z|X,\theta_t)dZ\\&=\int\ln p(X,Z|\theta)p(Z|X,\theta_t)dZ-\int\ln p(Z|X,\theta_t)p(Z|X,\theta_t)dZ\end{aligned} \tag{13}$$

略去下界函数 B(θ， θ_t)中与待求参数向量 θ 无关的项得到如下 Q 函数：

$$Q(\theta,\theta_t)=\int\ln p(X,Z|\theta)p(Z|X,\theta_t)dZ \tag{14}$$

（$Q(\theta,\theta_t)$ 隐含变量 Z 的函数 L(θ |X,Z)在概率分布 p(Z|X， θ_t)下的数学期望）

由于 B(θ， θ_t)≤ InL(θ |X)，故可通过迭代选取不同下界函数 B(θ， θ_t)最大值的方式逐步逼近对数似然 InL(θ |X)的最大值。

EM 算法的基本步骤如下：

（1）设置初始参数 θ_0和迭代停止条件；

（2）E 步(期望步)：根据可直接观测数据 X 和当前参数向量取值 θ_t：计算 Q(θ， θ_t)；

Copyright © 2025 by authors and Viser Technology Pte. Ltd.

$$Q(\theta,\theta_t)=\int \ln p(X,Z\mid\theta)p(Z\mid X,\theta_t)dZ \quad (15)$$

（3）M步(最大化步)：最大化Q(θ，θt)并根据Q(θ，θt)最大值更新参数θt的取值。

$$\theta_{t+1}=\arg_{\beta}\max Q(\theta,\theta_t) \quad (16)$$

（4）判断是否满足迭代停止条件，若满足则停止迭代，否则令 t=t+1 并返回步骤（2）。

2.3 实验条件

（1）Matlab2024

（2）CPU：4 核心（AMD/INTEL）

（3）内存：4G

（4）硬盘空间占用：6G 最少安装占用

2.4 实验内容

（1）产生两个均值不同的正态分布的随机数

（2）根据产生的随机数实现 EM 算法。

2.5 实验步骤

（1）编写随机数产生程序。

（2）编写 EM 算法程序。

3 实验课程安排

实验安排如表 1 所示：

表 1　实验课程安排

序号	教学内容	推荐学时/天	教学方式
1	设计动员及任务布置	0.5	讲授
2	熟悉题目、查资料、拟定初步方案	2	现场指导
3	确定方案、机器学习模型训练	5	现场指导
4	撰写设计说明书。	2	现场指导
5	答辩，完善设计说明书	0.5	现场指导

4 实验课程考核

4.1 课程考核环节

课程考核检验课程目标达成情况。考核环节包括实查阅文献、方案设计、模型训练与结果呈现和答辩成绩。

4.2 达成课程目标的途径与措施

课程目标达成度评价包括课程分目标达成度刘永明，男（1986.1-），河南商丘人，博士，讲师，硕士生导师，研究方向机器学习、智能优化设计、可靠性工程。评价，以及课程总目标达成度评价，具体计算方法如下：

课程分目标达成度

$$=\frac{\text{总评成绩中支撑该课程项目相关考核环节平均得分之和}}{\text{总评成绩中支撑该课程目标相关考核环节目标总分}}$$

$$\text{课程总目标达成度}=\frac{\text{该课程学生总评成绩平均值}}{\text{该课程总评成绩总分}(100\text{ 分})}$$

5 结语

机器学习是当代专用人工智能的技术核心，是人工智能领域的一个热门话题，并受到了广泛关注[10]。尽管机器学习在许多领域都取得了显著的成功，但我们也要认识到机器学习仍然面临许多挑战和限制。因此，设计一个机器学习课程的实验大纲，以提高学生对机器学习的理解与应用是很有必要的。并且面对当前的实验课问题，教师也要紧跟实验课改革步伐，及时改进实验课教学方法，促进机器学习实验课的发展，为我国培养更多的应用型人才。

基金项目：2023 年安徽省省级质量工程项目（2023sdxx043，2023xjzlts040），2023 年度新时代育人省级研究生教育教学改革研究项目，安徽工程大学校级质量工程项目（2022jyxm02，2022yszy03，2023yz1005）。

[参考文献]

[1]张幸幸,朱振峰,赵亚威,等.机器学习中原型学习研究进展[J].软件学报,2022,33(10):3732-3753.

[2]林晓农.试论人工智能与机器学习技术在智慧城市中的应用[J].信息系统工程,2020(1):2.

[3]李健宏.人工智能中的机器学习研究及其应用[J].江西科技师范学院学报,2004(5):3.

[4]韦南,殷丽华,宁洪,等.本科“机器学习”课程教学改革初探[J].网络与信息安全学报,2022,8(4):182-189.

[5]暴琳,陈熙维,魏海峰,等.人工智能课程群联合宽广数据资源拓展自动化品牌专业新工科建设[J].高教学刊,2022,8(26):21-24.

[6]苏向东,刘娜.《机器学习》课程项目实践方案研究[J].网络安全技术与应用,2022(8):97-99.

[7]李阳.机器学习课程的设计与教学模式实践[J].电子技术,2022,51(8):170-171.

[8]郭杰.人工智能实验课教学改革分析[J].无线互联科技,2019,16(15):87-88.

[9]孙大飞,陈志国,刘文举.基于 EM 算法的极大似然参数估计探讨[J].河南大学学报:自然科学版,2002,32(4):7.

[10]刘凯,胡祥恩,王培.机器也需教育?论通用人工智能与教育学的革新[J].开放教育研究,2018,24(1):10-15.

作者简介：刘永明（1986.1—），男，河南商丘人，博士，讲师，硕士生导师，研究方向机器学习、智能优化设计、可靠性工程。

Copyright © 2025 by authors and Viser Technology Pte. Ltd.