基于 TREC 真实邮件数据集的朴素贝叶斯分类教学案例构建与应用效果实证研究

曹寒问，陈锦文，车金星，张毓华
江西水利电力大学 理学院，江西 南昌 330099

DOI:10.61369/ASDS.2026010007

摘　　要： 人工智能技术的迅猛发展对概率统计学科教学提出了前所未有的革新要求。当前传统课程面临理论与应用脱节、教学案例缺乏应用、课堂缺乏实践，难以培养学生将概率模型转化为解决实际问题的能力。针对这一挑战，本研究基于 TREC Public Corpus 包含的75，419封真实邮件数据集，系统构建朴素贝叶斯分类教学案例。借助 Python 工具链实现邮件解析、文本预处理、特征工程至概率决策的全流程教学转化，使学生能够动态修改邮件内容并实时观测朴素贝叶斯后验概率变化。结合人工智能技术，学生得以了解理论知识的具体实践应用场景。

实证研究结果表明：模型在22，607封测试邮件中整体准确率较高，其垃圾邮件识别和正常邮件的识别精确度均表现优异。特征重要性分析揭示“pill”在垃圾邮件中出现概率显著高于正常邮件，而“per”和“desjardin”等商业词汇构成关键判别模式。教学实验中，学生通过添加“meeting”等工作词汇，成功降低测试邮件的垃圾概率，直观验证先验分布与似然概率的协同决策机制。并通过对223名学生分为案例教学组和传统教学组开展对比研究，独立样本 t 检验结果显示，案例教学组的期末成绩显著优于传统教学组，两组差异达到统计显著水平，平均成绩提升5.30分，及格率提高16.8个百分点。

该案例将条件概率、全概率公式等抽象理论转化为可操作的实践载体，显著提升学生构建概率模型解决复杂问题的能力，突破传统教学中公式记忆和机械演算的认知局限，实现理论向应用层面的跃迁。

关 键 词： 概率统计；朴素贝叶斯；条件概率；理论实践融合

Empirical Study on the Construction and Application Effectiveness of a Naive Bayes Classification Teaching Case Based on the TREC Authentic Email Dataset

Cao Hanwen, Chen Jinwen, Che Jinxing, Zhang Yuhua
School of Science, Jiangxi University of Water Resources and Electric Power, Nanchang, Jiangxi 330099

Abstract: The rapid advancement of artificial intelligence technology has imposed unprecedented demands for innovation in probability and statistics education. Current traditional curricula face challenges such as disconnect between theory and application, lack of practical teaching cases, and insufficient hands-on classroom activities, making it difficult to cultivate students' ability to translate probability models into solutions for real-world problems. To address this challenge, this study systematically constructs a Naive Bayes classification teaching case based on the TREC Public Corpus dataset comprising 75,419 authentic emails. Utilizing Python toolchains, the entire teaching process—from email parsing and text preprocessing to feature engineering and probabilistic decision-making is fully implemented. This enables students to dynamically modify email content and observe real-time changes in the Naive Bayes posterior probability. By integrating artificial intelligence technology, students gain insight into the concrete practical applications of theoretical knowledge.

Empirical research findings indicate that the model demonstrated high overall accuracy across 22,607 test emails, with outstanding precision in identifying both spam and legitimate messages. Feature importance analysis revealed that “pill” appears significantly more frequently in spam emails than in legitimate ones, while commercial terms like ‘per’ and “desjardin” form key discriminative patterns. In the teaching experiment, students successfully reduced spam probability by adding work-related vocabulary like “meeting,” intuitively validating the collaborative decision-making mechanism

基金项目：江西省教育厅高等学校教学改革研究省级重点课题（AI 赋能《概率论与数理统计》的个性化教学方式探索，编号：JXJG-24-18-5）。

作者简介：
曹寒问，江西水利电力大学理学院，硕士，副教授，研究方向：数据分析，22099166@qq.com；
陈锦文，江西水利电力大学理学院，本科生，专业：应用统计学，1283671376@qq.com；
车金星，江西水利电力大学理学院，博士，教授，研究方向：人工智能与水电能源统计，jinxingche@163.com；
张毓华，江西水利电力大学理学院，博士，副教授，研究方向：能源经济统计分析，1984zhangyuhua@163.com。

between prior distribution and likelihood probability. A comparative study involving 223 students divided into a case-based teaching group and a traditional teaching group was conducted. Independent samples t-test results showed that the case-based teaching group achieved significantly higher final scores than the traditional teaching group, with the difference reaching statistical significance. The average score improved by 5.30 points, and the pass rate increased by 16.8 percentage points.

This case study transforms abstract theories such as conditional probability and the law of total probability into practical tools, significantly enhancing students' ability to construct probability models for solving complex problems. It breaks through the cognitive limitations of traditional teaching methods focused on formula memorization and mechanical calculations, achieving a leap from theory to application.

Keywords : probability and statistics; naive bayes; conditional probability; theory-practice integration

引言

全球 AI 融合应用正加速推进社会变革，据市场分析预测，2025年人工智能市场规模将突破4.8万亿美元，技术渗透率预计超过68%。在此背景下，概率统计学科需应对技术发展带来的教学革新需求。根据教育部《高等学校人工智能创新行动计划》的战略部署，基础数学课程与人工智能技术的深度整合已成为高等教育改革的关键路径。尽管产业界对数据分析人才的需求呈现指数级增长，现行概率统计课程却暴露出教学内容与技术前沿的严重脱节问题——这种脱节既反映在教学案例的时效性不足，更体现在数学理论与算法实现之间的认知壁垒，最终制约学生将数学概念转化为实际问题解决能力的发展[1]。

概率统计课程教学案例体系目前存在三个问题：内容更新滞后、理论实践断层和真实训练缺失。教育部大学数学课程群虚拟教研室2025年调研数据显示，全国72.3% 高校的概率统计课程仍在沿用工业化时期的经典案例，这些案例已难以匹配现代数据分析需求；在贝叶斯公式教学中，普遍采用的简化版疾病检测模型既弱化了先验概率设定的复杂性，又模糊了似然函数估计的技术难点，致使真实决策过程无法完整呈现；更值得关注的是，2023年全国12所高校课程评估报告表明，仅有29.6% 的课程作业涉及真实数据集分析，这种设计缺陷直接导致教学陷入“公式记忆 - 机械演算”的恶性循环。显然，现有教学范式既无法有效培养学生基于真实数据构建概率模型的能力，也难以满足人工智能时代对创新人才核心素养的培养要求，迫切需要开发整合真实数据集与现代技术场景的新型教学载体[2]。

本研究基于 TREC Public Corpus 真实邮件数据集开发朴素贝叶斯分类教学案例，该数据集包含75，419封真实邮件，案例设计以传统统计学中的条件概率理论为根基，将朴素贝叶斯算法转化为可授课的教学载体，通过结构化解析邮件文本内容、结合文本清洗技术与词干归并处理，构建适合课堂演示的特征工程流程。教学方法重点阐释词频统计与传统概率估计的内在联系，引导学生掌握机器学习算法对经典统计方法的拓展逻辑。研究借助 Python 教学工具演示贝叶斯决策过程，建立传统统计理论与智能算法实践的教学桥梁。

一、朴素贝叶斯分类方法与数学原理

(一) 垃圾邮件分类问题描述

文本分类作为机器学习领域的经典任务，核心目标在于将文档自动归类至预定义类别体系。伴随数字化信息交流的蓬勃发展，电子邮件已成为关键沟通载体，其重要性不言而喻。但垃圾邮件的泛滥严重破坏着这一渠道的高效与安全，引发信息过载、隐私泄露及网络欺诈等多重风险。为应对此挑战，机器学习技术被引入电子邮件过滤领域，其中朴素贝叶斯分类算法凭借简洁高效的优势脱颖而出。该算法以贝叶斯定理为理论基础，通过数据集统计规律学习实现分类决策。“朴素”命名源于特征独立性假设，此假设显著降低计算复杂度，并为复杂分类问题构建基础框架。在垃圾邮件过滤场景中，基于邮件文本特征的深度分析，该算法能够精准识别垃圾邮件与正常邮件，大幅提升用户信息管理效率[3]。

(二) 朴素贝叶斯算法原理

朴素贝叶斯分类算法的理论根基深植于概率论体系，其核心机制在于最大化后验概率[4]。朴素贝叶斯分类器的构建旨在精准计算每个类别 C_i 在给定特征向量 X 下的后验概率 $P(C_i|X)$，依据贝叶斯定理，该概率可严谨推导为：

$$P(C_i|X)=\frac{P(X|C_i)\cdot P(C_i)}{P(X)}$$

其中，$P(C_i)$ 是通过训练数据集中类别 C_i 样本出现的相对频率估计得出的先验概率；$P(X|C_i)$ 是在类别 C_i 确定的条件下，特征向量 X 出现的条件概率，即似然概率；$P(X)$ 则是特征向量 X 在整

个样本空间中出现的边缘概率。在实际运算中，鉴于$P(X)$对于所有类别而言保持恒定，因此在比较不同类别对应的后验概率时，可将其视为常数项予以简化处理。朴素贝叶斯算法通过精确计算每个类别的后验概率，并依据最大后验概率准则确定样本类别，从而实现高效精准的分类决策。

多项式朴素贝叶斯垃圾邮件分类器的工作流程可以清晰地总结为图1。

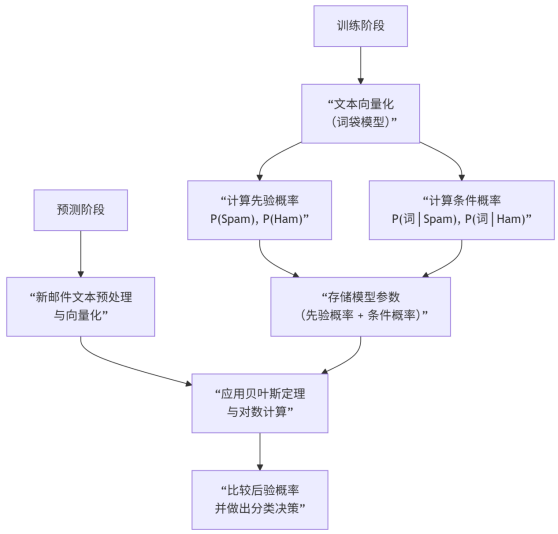

图1：贝叶斯垃圾邮件分类器工作流程图

Figure 1: Workflow Diagram of the Bayesian Spam Classifier

二、教学案例设计

（一）数据准备及预处理

研究采用 TREC Public Corpus 提供的75，419封真实邮件作为数据集。邮件解析过程借助 Python 邮件模块处理原始文件[5]，有效应对多部分邮件结构与多样编码格式的解析需求。数据预处理涵盖三个核心环节：文本清洗环节移除 HTML 标签、URL 链接、邮箱地址及非字母字符；停用词过滤环节基于 NLTK 英语停用词表剔除“the”“and”等无意义词汇[6]；词干提取环节采用 PorterStemmer 算法将词形变体归并为统一词根。标签数据通过独立索引文件加载实现邮箱精确标记，其中垃圾邮件占总量的66.5%，一共50，126封。据此，我们通过带入先验概率计算公式计算各类别邮件在总体样本中的频率，得到了朴素贝叶斯分类器所需的先验概率：

垃圾邮件（Spam）的先验概率 P(Spam) = 50，126 / 75，419 ≈ 66.5%。

正常邮件（Ham）的先验概率 P(Ham) = 25，293 / 75，419 ≈ 33.5%。

（二）模型训练及结果展示

模型训练基于多项式朴素贝叶斯分类器，核心任务聚焦两类关键概率参数估计。先验概率计算获得：P(Spam)=66.5%，P(Ham)=33.5%，该结果准确反映数据集真实分布。条件概率参数基于特征词频统计进行估计。特征词概率分布呈现显著差异，作出表1，其数据显示：商业推广词汇“pill”在垃圾邮件中的概率达0.015，“price”在垃圾邮件中概率为0.0089；相比之下，工作相关词汇“offic”在正常邮件中出现概率更高。特征重要性分析进一步揭示“anatrim”、“ciali”等词汇具有显著判别力，此类差异构成分类决策的数学基础[7]。

表1：最重要的垃圾邮件特征词

Table 1: Most Significant Spam Feature Words

排名	特征词	垃圾概率	正常概率	重要性
1	pill	0.015147	0.000135	0.015012
2	per	0.01213	0.000636	0.011494
3	desjardin	0.010682	0.000001	0.010681
4	price	0.00892	0.001744	0.007176
5	item	0.007833	0.000746	0.007087
6	save	0.007351	0.000803	0.006547
7	product	0.006641	0.001262	0.005379
8	votr	0.005221	0.000003	0.005219
9	viagra	0.004886	0.000014	0.004872
10	onlin	0.005584	0.000798	0.004786
11	transact	0.004852	0.000176	0.004676
12	vou	0.004579	0.00001	0.004569
13	anatrim	0.004123	0.000001	0.004123
14	retail	0.003755	0.000117	0.003638
15	ciali	0.003478	0.000005	0.003473
16	men	0.003541	0.000169	0.003372
17	buy	0.003652	0.000409	0.003242
18	money	0.003602	0.000429	0.003173
19	qualiti	0.003251	0.00025	0.003001
20	adob	0.002984	0.00007	0.002914

对邮件中的关键词特征重要性和特征概率对比作图2。

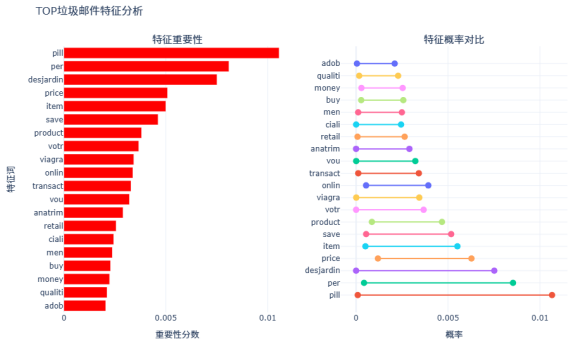

图2：TOP 垃圾邮件特征词分析

Figure 2 : Analysis of TOP Spam Keywords

图2特征重要性分析结果呈现词汇判别能力分布规律，前三关键特征词——“pill”（药物类）、“per”（促销类）、“desjardin”（品牌类）的重要性值均超过0.01。这些词汇在垃圾邮件中的出现概率较正常邮件高出数十倍至百倍。商业推广词汇明显构成垃圾邮件的核心标识，而工作相关词汇则与正常邮件显著关联。随后

对高频词的频率做统计并且可视化作出图3。

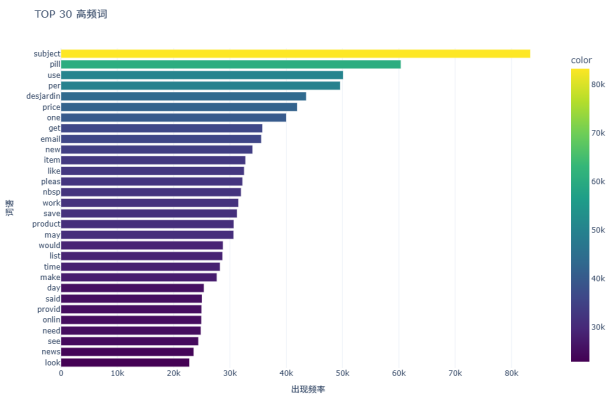

图3：邮件内容 TOP30 高频词分布

Figure 3: Distribution of the TOP 30 High-Frequency Words in Email Content

高频词统计图3呈现邮件文本的基础构成特征。主题词“subject”以超8万次的绝对频次居首位，体现邮件基础结构的普遍性。需关注的是，高判别力特征词“pill”同时位列高频榜前五，其出现频率达常规词汇“news”的7倍。该双重特征印证特征选择策略的有效性——高频商业词汇天然携带强判别力，应作为过滤规则优化的核心目标。

课堂交互演示环节，输入测试邮件后模型以100% 置信度判定其为垃圾邮件。决策分析表明“offer”、“discount”等词汇引发高条件概率响应，此类词汇在特征重要性排名中均居前20位，直观揭示特征词对分类决策的作用机制。

测试阶段采用22，607封邮件构成测试集，正常邮件7，566封，垃圾邮件15，041封。作出混淆矩阵如图4显示：正常邮件正确识别率达83.5%，7，515封正确分类，51封误判为垃圾邮件；垃圾邮件成功拦截13，554封，漏判1，487封。垃圾邮件识别存在显著漏判现象，此特性表明模型优先保障正常邮件的准确传递（符合”避免误拦重要邮件”的实际需求），同时需重点识别垃圾邮件的隐蔽变体。

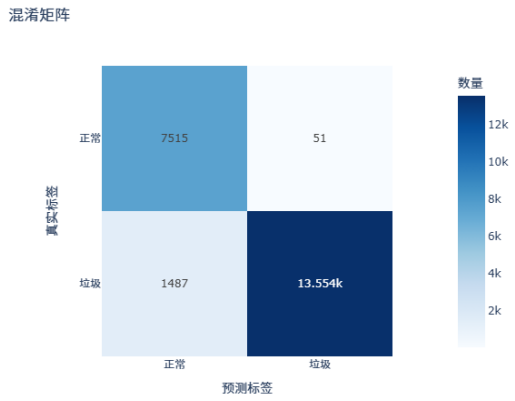

图4：朴素贝叶斯分类器混淆矩阵

Figure 4: Confusion Matrix of the Naive Bayes Classifier

整理出模型的关键性能指标作出表2。

表2：分类性能指标

Table 2 Classification Performance Metrics

类别	精确率	召回率	F1分数	样本量
正常邮件	83.5%	99.3%	0.91	7，566
垃圾邮件	99.6%	90.1%	0.95	15，041

垃圾邮件识别精确率达99.6%，召回率为90.1%；正常邮件召回率高达99.3%，但精确率仅83.5%。该差异反映模型对垃圾邮件的判定策略较为保守，而对正常邮件的识别相对宽松，分类结果体现朴素贝叶斯模型在垃圾邮件过滤中的特性：垃圾邮件判定精确度极高达到99.6%，但正常邮件识别精确率为83.5%。这种非对称性能源于实际需求平衡——系统选择允许部分垃圾邮件进入收件箱，以近乎100% 的保障重要邮件送达。此设计契合邮件服务的核心原则：优先确保通信可靠性，其次提升过滤效率。

三、教学效果实证分析

（一）实验设计与研究方法

本研究选取2022，2023，2024三个连续学年度的应用统计学专业教学班级作为研究对象，所有班级均由同一教师授课，统一学习概率论课程，确保教学条件的一致性。教学改革组包含2023级（传统教学结合基础练习，n=85）与2024级（案例教学法，n=59）两个班级，共计144名学生；传统教学组为2022级学生群体（n=79），采用纯理论讲授模式。在教学干预实施过程中，教学改革组强调将概率统计理论知识与真实世界问题解决相结合，通过朴素贝叶斯分类器在垃圾邮件识别中的具体应用，将条件概率、贝叶斯公式等抽象数学概念转化为可操作的实践工具；而传统教学组则维持以公式推导与例题讲解为主的理论传授模式[8]。

研究采用独立样本 t 检验探究教学模式（教学改革 vs 传统教学）对期末考试成绩影响的统计学显著性，检验的原假设设定为两种教学模式下的学生平均成绩无显著差异。以评估教学改革的实际成效。

（二）实验结果分析

对数据做描述性统计结果作出表3，教学改革组学生的平均成绩为63.69分，传统教学组学生的平均成绩为58.39，两组间存在5.30分的均值差异。

表3：不同教学模式学生成绩描述性统计对比

Table 3: Descriptive Statistics Comparison of Student Performance Across Different Teaching Modes

组别	样本量	均值	标准差	最小值	25% 分位	中位数	75% 分位	最大值
传统教学组	79	58.39	16.66	26.0	47.0	60.0	70.0	98.0
教学改革组	144	63.69	16.11	14.0	51.0	64.0	75.0	93.0

为了直观展示实验分析过程，对数据进行可视化作出图5。

图5两组学生成绩分布比较直观展示了教学改革组与传统教学组的成绩分布特征。从箱线图可见，教学改革组的中位数位置更高，四分位距分布更为集中；小提琴图则进一步揭示了教学改革组在中等分数区间的密度更高，表明该教学方法对中等水平学生的提升效果尤为明显。均值比较与及格率分析结果清晰显示，教

学改革组在平均成绩与及格率两个维度均显著优于传统教学组。均值比较图中的置信区间表明两组差异的估计精度，而及格率比较则凸显了教学改革在降低不及格率方面的实际成效，这一发现具有重要的教学实践意义。

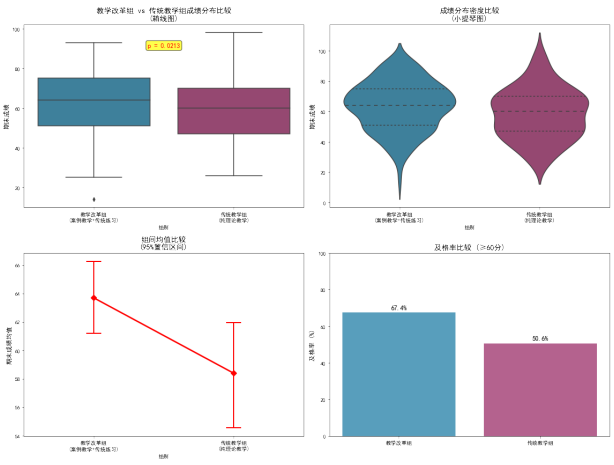

图5：教学效果检验可视化

Figure 5: Visualization of Teaching Effectiveness Evaluation

为验证组间差异的统计学显著性，研究进行了独立样本 t 检验。分析结果整理成表4，发现教学改革组与传统教学组的成绩差异达到统计显著水平 t=2.319，因此我们拒绝原假设（H_0），接受备择假设，即教学改革组与传统教学组的学生平均成绩存在显著差异。且 p=0.021<0.05结果显著 [9]。

表4: 教学改革效果推断统计检验结果

Table 4: Statistical Test Results for Inferring the Effects of Teaching Reform

指标项	t 值	自由度	p 值	均值差异	95% CI	Cohen’s d
数值	2.319	221	0.0213	5.30	0.78 - 9.81	0.325

四、结论与展望

本研究基于 TREC Public Corpus 真实邮件数据集构建融合概率统计理论与机器学习实践的朴素贝叶斯分类教学案例。通过对75，419封邮件的系统化文本特征解析，完成从条件概率理论至贝叶斯决策算法的教学创新，证实传统概率论知识在现代人工智能应用中的核心价值。案例教学表明：

理论实践融合：贝叶斯公式、全概率公式等核心概念通过邮件分类任务具象化，学生能直观理解条件概率估计与先验分布的协同决策机制 [10]；

算法可解释性：特征重要性分析，将抽象的概率计算转化为可观测的语言特征规律；

教学有效性：测试集高度的精确率证明，基于真实数据集的案例设计显著提升学生解决复杂问题的能力，突破传统教学中只有公式记忆，机械演算的认知局限。将理论层面提高到了应用层面 [11]。

通过持续迭代教学案例库与实验工具链，推动概率统计课程实现从理论传授到”数据驱动 – 算法实现 – 决策优化”的能力培养范式转型，为人工智能时代培养兼具数学素养与工程能力的创新人才 [12]。

参考文献

[1] 肖睿，王峰，黄文彬 . 人工智能赋能教育的发展态势与未来路径 [J]. 现代教育技术 ,2023,33(1):12-20.

[2] 陈希儒，刘乐平 . 新时代统计学教育改革的方向与路径 [J]. 统计研究 ,2022,39(4):145-156.

[3] 周志华 . 机器学习 [M]. 北京：清华大学出版社 ,2021:123-135.

[4] 李航 . 统计学习方法 [M]. 第2版 . 北京：清华大学出版社 ,2021:58-72.

[5] 张良均，王靖，刘名军 .Python 数据挖掘与机器学习实战 [M]. 北京：人民邮电出版社 ,2022:89-105.

[6] 宗成庆 . 统计自然语言处理 [M]. 第2版 . 北京：清华大学出版社 ,2021:67-82.

[7] 黄文彬，徐健 . 基于特征重要性的文本分类模型可解释性研究 [J]. 计算机研究与发展 ,2023,60(3):567-578.

[8] 温忠麟，刘红云 . 教育实证研究中的统计分析方法 [M]. 北京：北京师范大学出版社 ,2020:156-170.

[9] 张厚粲，徐建平 . 现代心理与教育统计学 [M]. 第5版 . 北京：北京师范大学出版社 ,2021:245-260.

[10] 王陆，刘菁 . 人工智能时代案例教学法的创新路径研究 [J]. 电化教育研究 ,2023,44(2):78-85.

[11] 任友群，李锋 . 面向人工智能时代的中小学计算思维培养 [J]. 中国电化教育 ,2022(5):1-8.

[12] 祝智庭，魏非 . 教育数字化转型的现实路径与发展趋势 [J]. 华东师范大学学报 (教育科学版),2023,41(1):1-15.

