

机器学习算法教学实践与探索： 以线性回归为例

文 乐, 陈有华*

华南农业大学经济管理学院, 广东 广州

收稿日期: 2023年8月8日; 录用日期: 2023年9月13日; 发布日期: 2023年9月22日

摘 要

随着当今人工智能技术的迅速发展和广泛应用,机器学习作为其重要组成部分彰显出关键性作用。然而,机器学习领域的知识涵盖范围广泛、更新快速且难度较大,给教师和学生带来了巨大挑战,特别是针对非计算机专业的经管类本科生,缺乏针对性的教学经验。本文以线性回归为例,探讨了一种融合理论与实践的机器学习算法教学方法。通过简洁的理论讲解算法原理,然后引导学生运用实际数据进行操作,涵盖数据预处理、模型训练和评估等步骤,这种方法有助于学生从被动地接受知识转变为主动学习并参与实践,深化对算法的理解,也有助于机器学习课程的设计与改进。

关键词

机器学习, 线性回归, 人工智能, 教学模式

Teaching Practice and Exploration of Machine Learning Algorithms: A Case Study of Linear Regression

Le Wen, Youhua Chen*

College of Economic & Management, South China Agricultural University, Guangzhou Guangdong

Received: Aug. 8th, 2023; accepted: Sep. 13th, 2023; published: Sep. 22nd, 2023

Abstract

With the rapid development and widespread application of today's artificial intelligence technol-

*通讯作者。

ogy, machine learning plays a crucial role as a significant component. However, the field of machine learning encompasses a wide range of knowledge, updates swiftly, and presents substantial challenges for both educators and students. This challenge is particularly pronounced for non-computer science majors, such as business and management undergraduate students, who lack tailored teaching experience. This paper takes linear regression as an example to explore a blended approach to teaching machine learning algorithms that integrates theory and practice. Through concise theoretical explanations of algorithm principles, students are guided to apply practical data and cover steps like data preprocessing, model training, and evaluation. This method facilitates the transformation of students from passive knowledge recipients to active participants in practical applications, thereby deepening their understanding of algorithms. It also holds potential for the design and enhancement of machine learning courses.

Keywords

Machine Learning, Linear Regression, Artificial Intelligence, Teaching Approach

Copyright © 2023 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

机器学习(machine learning)是一套从大数据中提取知识的算法,也被称为预测分析或统计学习。通俗而言,机器学习就是专门研究计算机如何尽可能地模拟或实现人类的学习行为,以获取新的知识或技能,从而不断改善系统自身的性能[1][2]。机器学习是人工智能的基础和核心,如今已经渗透到了人们日常生活的各个方面,包括各大电影、购物、新闻等应用软件的个性化推荐,到医疗诊断和社交网络上寻找好友等[3]。

值得注意的是,尽管机器学习在各领域取得了显著的成就,推动着人工智能的发展,然而,在面向经管类本科生的教学中却面临着诸多挑战[4][5]。许多机器学习课程往往注重软件工具的应用,而缺乏对算法背后原理的深入讲解。这种现象可能导致学生仅仅掌握了表面的操作方法,而对于算法的理解和推导却知之甚少。随着机器学习的不断发展,教育者需要更好地平衡理论和实践,使学生能够真正理解算法的内在原理,从而在未来的应用中能够更灵活地运用这些知识[6]。

此外,教育部积极响应国家科技创新战略,于2018年4月发布了《高等学校人工智能创新行动计划》,旨在推动我国高校人工智能领域的人才培养。2020年11月,国家教育部进一步发布了《新文科建设宣言》。在这一宣言中,强调了将现代信息技术融入文科教育,以提供学生全面跨学科的学习机会,培养更具综合素养的人才。

在上述背景下,本文旨在探讨面向经管类专业本科生的机器学习算法的教学方法,具体以线性回归作为案例进行详细展示。线性回归是机器学习领域中最基础、最常用的算法之一,在解决许多实际问题时具有重要的作用[7]。深入剖析机器学习算法的原理和应用,有利于将理论与实践相结合,为学生提供跨学科学习的机会,培养其创新思维、问题解决和综合分析能力。此外,探索机器学习算法的教学也有利于满足教育部提出的人工智能培养目标和新文科建设的要求,也是推动高等教育实现优质发展的必然选择。

2. 文献回顾

在近年来,机器学习作为一门跨学科领域,在全球范围内引起了广泛的关注。机器学习涵盖了多

个学科领域, 如数学、统计学、计算机科学等[5], 它通过算法从海量数据中提取知识, 以模拟人类学习的过程[1]。然而, 正因为机器学习的多学科性和高难度性, 使其在教学实践中面临一些挑战[4]。

针对机器学习的教学模式, 目前已有一些成果, 但也存在着一些问题。具体来说, 一方面, 部分教学模式强调工具的使用, 忽视了算法背后的理论原理, 导致学生在具体操作中熟练掌握了一些技能, 但缺乏对于算法本质的深刻理解。换言之, 学生可能通过软件工具完成一些任务, 但却不太理解算法的内在机制[8]。这种现象导致了学生在理论和实践之间存在脱节, 使得他们难以真正掌握算法的本质和应用范围[6]。另一方面, 一些教学模式则偏向于理论, 过多地强调数学推导和模型的数学基础, 却缺乏实际案例和实践操作的引导。这导致学生容易丧失学习兴趣, 并难以将所学知识应用于复杂的现实场景[9]。

总的来看, 现有教学模式未能充分满足机器学习教育的核心要求: 既要使学生掌握实际应用的操作技能, 又要培养他们对算法背后原理的深入理解[3]。在机器学习领域, 轻视理论和只关注实际应用, 或者反之, 都会限制学生在日后工作中的创新能力和问题解决能力。寻求一种能够平衡理论与实践的教学方法, 成为了迫切的需求。因此, 本文拟通过以线性回归为例的机器学习算法教学实践, 探讨一种融合理论与实践的教学方法。通过引导学生从理论到实际操作, 从而更全面地掌握机器学习算法的核心概念和应用能力。

3. 教学设计

3.1. 算法概述

线性回归是机器学习的一类基础的算法。回归的英文单词是 **Regression**, 顾名思义, 即指返回、回到某种状态。在统计与计量经济学中, 回归一般是指回归到数据的中间值[10]。具体来说, 回归就是用一条曲线对数据点进行拟合, 该曲线被称为最佳拟合曲线。回归问题的求解过程就是通过训练分类器采用最优化算法来寻找最佳拟合参数集[2]。当拟合线为直线时, 就是经典的线性回归模型。

线性回归一般用来做连续值的预测, 即预测的结果是一个连续值。举例来说, 银行了解到贷款人的年龄和工资等信息, 然后根据这些信息给出相应的贷款额度, 在此处, 贷款额度就是一个连续性的变量。线性回归的基本思想是根据样本特征(年龄、工资)来预测结果(贷款金额)。值得注意的是, 线性回归模型不仅需要提供样本的特征向量, 还需要提供样本的实际结果(即标记, **label**), 因此, 线性回归模型也属于机器学习的一种监督学习。

线性回归的基本流程如图 1 所示, 大致分为六个步骤。首先对数据进行基本的预处理, 为接下来的建模做好准备, 然后定义模型损失函数, 并基于样本和特征大小, 可采用不同方法对参数进行求解。估计出初步结果后, 需要对模型拟合效果进行检验和调试, 最后做出经济预测。

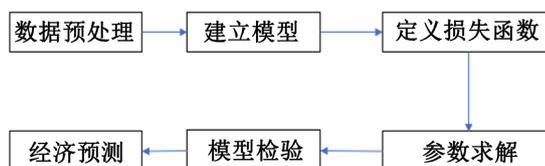


Figure 1. Flowchart of the linear regression algorithm

图 1. 线性回归算法流程图

3.2. 基本原理和算法步骤

线性回归是机器学习中最基本和常用的回归算法之一, 用于建立输入特征与连续输出之间的线性关系。其基本原理是通过找到最佳的线性函数来拟合数据, 使得预测值与实际观测值之间的误差最小化[7]。

下面详细介绍线性回归的基本原理和算法步骤。

1) 模型建立。线性回归假设输出与输入特征之间存在线性关系, 即输出可以通过输入的线性组合来表示。现假设有输入特征向量 $x = [1, x_1, x_2, \dots, x_m]$, 权重向量 $w = [w_0, w_1, w_2, \dots, w_m]$, 以及实际值 y 。线性回归模型的预测值 \hat{y} 可以表示为内积形式:

$$\hat{y} = w \cdot x^T \text{ 或 } \hat{y} = x \cdot w^T$$

其中, w 表示权重向量, x 表示输入特征向量, T 表示向量转置。线性回归的目标是通过调整权重向量 w , 使得预测值 \hat{y} 尽可能接近实际观测值 y 。

2) 定义损失函数。损失函数在机器学习中扮演着至关重要的角色, 它是衡量模型预测值与实际观测值之间差异的指标。在线性回归或其他模型训练中, 设计合适的损失函数可以帮助优化算法找到最佳的模型参数, 使模型能够更好地拟合数据。在线性回归中, 可以从似然函数出发来推导损失函数, 从而找到使模型最适应数据的参数。考虑简单的线性回归模型, 其中 x 是输入特征, y 是实际观测值。假设数据点之间是独立同分布的, 给定参数 w (权重), 观测值 y 的似然函数可以表示为高斯分布(正态分布)的概率密度函数:

$$P(y|x, w) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-wx)^2}{2\sigma^2}\right)$$

其中, w 是权重, σ 是噪声标准差。线性回归的目标是找到最适合数据的模型参数 w , 使得似然函数最大化。为使预测值成为真实值的可能性最大, 进行最大似然估计, 得到似然函数如下:

$$L(w) = \log\left(\prod_{i=1}^n P(y_i|x_i, w)\right)$$

$$L(w) = \frac{n}{2} \log(2\pi\sigma^2) - \sum_{i=1}^n \frac{1}{2\sigma^2} (y_i - (wx_i))^2$$

从上述推导可以发现, 最后得到的对数似然函数正是线性回归的损失函数的一种组合, 即一个与样本数量(n)和噪声标准差(σ)有关的常数项, 再加上一个均方误差项(Mean Squared Error, MSE)。通常, 为了进行参数优化, 忽略与参数无关的项, 从而得到标准的线性回归损失函数, 具体如下:

$$\text{MSE}(w) = \frac{1}{n} \sum_{i=1}^n (y_i - wx_i)^2$$

通过最小化该损失函数, 能够找到最佳的模型参数 w , 使得模型在训练数据上能够最好地拟合。上述推导过程将似然函数与损失函数联系在一起, 有助于理解为什么在线性回归中使用均方误差作为损失函数。

3) 参数估计。估计模型参数可以采用两种方法: 一是建立正规方程组, 通过求解一个封闭形式的方程直接找到最优的参数。二是利用梯度下降方法, 梯度下降是一种迭代优化算法, 通过逐步更新参数来逼近最优解。选择使用正规方程还是梯度下降取决于问题的性质和数据的规模。正规方程在数据较小且特征数不太大时适用, 它必须有解析解; 而梯度下降适用于大规模数据和高维特征空间, 在机器学习中更为常用。下面详细介绍梯度下降方法。第一步, 初始化参数, 通常可以随机选择一个初始值 w 。第二步, 计算损失函数梯度: 在每次迭代中, 计算损失函数关于参数的梯度, 即损失函数的偏导数。对于均方误差(MSE)损失函数, 梯度的计算公式如下:

$$\frac{\partial \text{MSE}}{\partial w_i} = -\frac{1}{n} \sum_{j=1}^n 2x_{j,i} (y_j - w \cdot x_j)$$

在迭代过程中, 需要注意控制学习的步长, 即学习率 α 。梯度下降的具体更新规则如下:

$$w_{i+1} = w_i - \alpha \frac{\partial MSE}{\partial w_i}$$

第三步, 进行迭代, 不断重复第二步, 直到达到预定的迭代次数或损失函数收敛到满意的程度。在迭代之前, 需要定义合适的停止条件, 如损失函数达到某个阈值, 参数变化很小等。如果满足停止条件, 则优化过程结束。一般当损失函数值变化已不大时, 可以停止迭代。

需要注意的是, 学习率 α 控制参数更新的步长, 然而过小的学习率会导致收敛速度慢, 而过大的学习率可能使优化过程不稳定。通常需要根据经验调整学习率。其次, 对特征进行标准化或缩放可以帮助梯度下降更快地收敛, 避免不同特征值范围带来的问题。此外, 批量梯度下降使用所有样本来计算梯度, 而随机梯度下降每次只使用一个样本。随机梯度下降在某些情况下可能更快, 但批量梯度下降可能更稳定。

3.3. 案例应用与 Python 实践

在前面的讨论中, 已经了解了线性回归是如何工作的, 以及如何使用梯度下降来优化模型参数, 使其最优地拟合数据。现在使用 Python 及其自带的波士顿房价数据集进行房价预测。这个案例将有助于将学到的知识转化为实际的操作。在这案例中, 将学会加载数据集, 准备数据, 构建线性回归模型, 使用梯度下降来优化模型参数, 并最终预测波士顿地区房价。

波士顿房价数据集是一个经典的数据集, 其中包含了影响波士顿地区住房价格的各种因素。具体而言, 数据集的特征包括犯罪率、房产税率、学生 - 教师比例等, 目标变量是房屋的中位数价格。我们的任务是通过构建线性回归模型, 使用数据集中的特征来预测房屋的中位数价格。通过对数据的学习, 希望找到一个最佳的线性关系, 将输入特征与房价之间的关系建模。以下是说明线性回归应用于波士顿房价数据集的过程:

第一步, 加载数据。利用 Python 工具包加载波士顿房价数据集, 获得特征和目标变量。为便于阅读 Python 代码, 添加了必要的注释, 具体如下:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import SGDRegressor
from sklearn.metrics import mean_squared_error
# 加载波士顿房价数据集
from sklearn.datasets import load_boston
boston = load_boston()
data = pd.DataFrame(boston.data, columns=boston.feature_names)
data['PRICE'] = boston.target
# 准备数据
X = data.drop('PRICE', axis=1) # 使用了数据集中除房价外的所有特征
y = data['PRICE']
```

第二步, 数据划分。将数据划分为训练集和测试集, 以使用训练集来训练模型, 用测试集来评估模型的性能。代码如下:

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)
```

第三步, 创建模型, 创建一个随机梯度下降回归模型。代码如下:

```
# 创建随机梯度下降回归模型
model = SGDRegressor(max_iter=10000, alpha=0.01,
learning_rate='constant', eta0=0.01, random_state=42)
```

第四步, 训练与测试模型, 并计算均方误差以评估模型。代码如下:

```
# 在训练集上训练模型
model.fit(X_train, y_train)
# 在测试集上进行预测
y_pred = model.predict(X_test)
# 计算均方误差
mse = mean_squared_error(y_test, y_pred)
print(f"Mean Squared Error: {mse}")
# 输出模型的参数
print("Model Coefficients (Weights):")
print("Intercept (Bias):", model.intercept_)
print("Coefficients (Weights):", model.coef_)
```

第五步, 可视化模型结果。将模型的预测结果与真实的房价进行对比, 通过可视化展示模型的预测效果。同时, 可以分析模型的权重参数, 了解各个特征对房价的影响程度, 从而得出关于波士顿房价的一些洞察。代码如下:

```
# 可视化预测结果
plt.scatter(y_test, y_pred)
plt.xlabel("True Prices")
plt.ylabel("Predicted Prices")
plt.title("True Prices vs Predicted Prices")
plt.show()
```

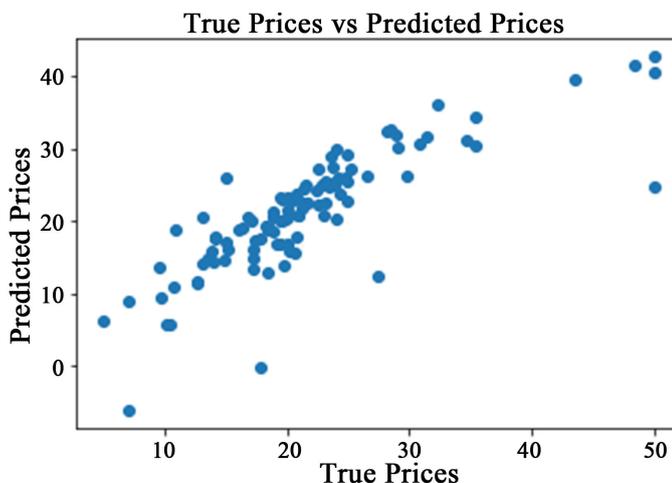


Figure 2. Scatter plot of the predicted vs. true housing prices in Boston
图 2. 波士顿房价预测值 vs 真实值散点图

从图2可以看到, 图中的点在一条45度的线附近分布, 这表示预测基本准确。需要注意的是, 如果点有较大的离散程度, 可能需要考虑模型的改进。此外, 较低的均方误差(24.29)和较接近于1的决定系数也表示模型的预测性能较好。需要说明的是, 限于篇幅, 没有报告模型截距项和系数(特征权重), 可以通过运行上述代码自动获得。

4. 结束语

如何在机器学习教学中平衡理论与实践, 使学生不仅能够深入理解机器学习算法的内在原理, 同时也能够熟练应用于实际问题, 成为一个值得探讨的现实问题。本文以线性回归模型为例, 既深入解释机器学习算法的理论原理, 又引导学生亲自动手构建模型、进行训练和预测, 并通过评价指标进行性能分析, 探讨了面向经管专业等文科学生的机器学习算法教学实践方法。

在当今日新月异的科技时代, 机器学习作为人工智能的核心领域, 其影响已经超越了学术界的范畴, 融入到了社会的各个层面。因此, 高校在培养未来人才时, 需要注重跨学科的培养, 让学生不仅能够具备扎实的理论基础, 还能够灵活运用所学知识解决实际问题, 使他们在未来能够更好地应对机器学习领域的挑战和变化。这也与教育部提出的人工智能创新行动计划以及新文科建设宣言的要求相契合, 为新时代的文科教育注入更多活力和创新。

通过本文的探讨, 希望为教育界提供一个新的思路, 即通过融合理论与实践的教学方法, 引导学生从算法的理论背景到实际应用的全过程。这将有助于学生在未来的职业生涯中更好地适应快速变化的科技环境, 为社会的发展做出积极贡献。总的来看, 机器学习算法教学实践是一个不断探索的过程, 相信随着技术的不断发展和教育理念的更新, 机器学习教育将会变得更加富有创新性和前瞻性。

基金项目

面向经管类专业本科生的《机器学习》课程体系建设与教学模式优化探索与实践(2021), 广东省教育教学研究改革重点项目; 面向农经专业本科生的《数据分析与挖掘》课程体系建设与教学模式优化研究(NJX22130), 农经教指委项目。基于“新农科”建设的农林经济管理专业人才培养模式探索与实践, 广东省教育教学研究改革重点项目; 教育部产学合作协同育人项目, 农业院校经管类专业大数据分析课程实习实训基地建设。

参考文献

- [1] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.
- [2] 赵娟娟, 强彦. Python 机器学习[M]. 北京: 机械工业出版社, 2019.
- [3] 张乐飞, 罗勇, 杜博. 机器学习教学改革与人工智能人才培养[J]. 中国大学教学, 2023(5): 18-21.
- [4] 文乐, 陈有华. 面向经管类专业的《机器学习》课程教学模式探索——以朴素贝叶斯分类算法为例[J]. 教育现代化, 2021, 8(60): 163-166.
- [5] 王重仁. 财经类高校机器学习课程教学探索与实践[J]. 中国管理信息化, 2023, 26(3): 212-215.
- [6] 王雷全, 吴春雷, 郭晓菲. 机器学习科研实践课程建设[J]. 电子世界, 2017(17): 50-51.
- [7] 陈强. 机器学习及 Python 应用[M]. 北京: 高等教育出版社, 2021.
- [8] 黄瑞章. 面向本科生的机器学习课程教学改革探讨[J]. 黑龙江科学, 2021, 12(3): 128-129.
- [9] 赵雪峰, 施珺. 面向本科生机器学习课程的教学探索[J]. 计算机教育, 2021(2): 170-174.
- [10] 邓子云. 深入机器学习[M]. 北京: 中国水利水电出版社, 2023.