《面向机器学习的计量经济学》教学大纲

(2025－2026 年第 1 学期)

课程中文名称： 面向机器学习的计量经济学

英文名称： Econometrics powered by Machine Learning

授课教师： 张征宇、金泽群、卢晓晖

课程类别： 高年级研讨课

课程安排说明： 2025-2026 学年秋季，每周四上午第 3-4 节

授课地点：教技楼 215

教学学时分配表：

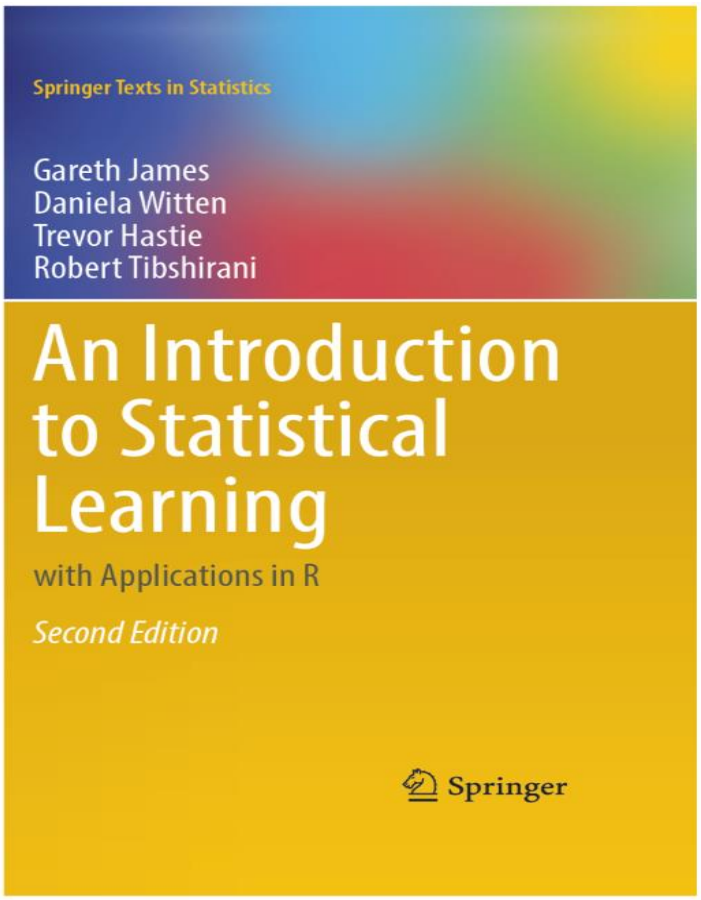
学分	总学时	理论教学学时	实践教学学时	实验教学学时
2	32	16	16	0

课件网址： https://canvas.shufe.edu.cn/courses/31401

教材和参考书目：

指定教材：

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, An Introduction to Statistical Learning with Applications in R, Second Edition, Springer.



参考书目：

Kevin P. Murphy, Machine Learning, A Probabilistic Perspective, The MIT Press.

严子中、张毅，计量经济学编程：以 Python 语言为工具，中国财政与经济出版社，2024 年 3 月。

预备知识：

本课程面向高年级本科生。学生在此之前应当已修完一学期的计量经济学（例如伍德里奇的现代计量经济学导论）；特别对多元线性回归模型的估计和推断有一定了解。除此以外，学过初级经济学，金融学原理、概率论与数理统计有助于更好地理解本课程内容。

先修课程：计量经济学。概率论和数理统计。

课程达成目标：

随着人工智能和大数据技术的快速发展，机器学习在包括计量经济学在内的各个经济学、金融学领域中获得广泛运用。在经济学研究中，机器学习可用于变量选择、数据降维、因果推断、反事实预测等用途，能够帮助经济学家、金融市场分析者更加高效地处理数据，探索其中的规律和关联。

本课程主要讲授目前已在学术研究与商业应用中获得普通认同，较为成熟，且具有广泛应用前景的机器学习方法。学完本课程后，学生可以了解常见机器学习方法与原理；例如 Lasso, Ridge, 弹性网、决策树、boosting, bagging, 随机森林、神经网络，K-means 与聚类方法，用于降维的主成分分析与因子模型。

本课程将用通俗易懂的语言帮助学生理解机器学习计量经济学的基本观点（预测和因果推断）、方法以及常见的应用场景。特别是理解机器学习与经典（以多元线性回归为主要内容的）计量经济学之间的区别；学完本课程，学生应当初步具备采用 R 或者 python 进行编程，将所学方法用于实际新问题研究的能力。

思政育人目标：

本课程在讲授机器学习计量经济学原理的同时，通过思政育人，实现知识传授、能力培养与价值思想塑造的有机统一。以下是三个案例。

案例一　样本内与样本外预测：科学理论既要解释过去，又能预测未来

在经济学实证研究中，经常会遇见这样一种现象：一个样本内拟合效果很好，模型的拟合优度 $R^2 = 1 - \frac{\text{SSR}}{\text{SST}}$ 很高的线性回归模型，其样本外预测却不准。若样本内拟合优度很高，有可能是线性回归模型包含了太多的无关的自变量或太多的“噪声”，这些外在的非本质的联系，不仅对预测未来没有帮助，反而降低了模型预测能力。科学理论不仅要能够解释过去（样本内拟合），还要能够预测未来（样本外预测）。只有揭示内在本质联系的理论，才具有真正的解释力和预测力。

机器学习通过划分训练数据和测试数据，并利用惩罚项（如 LASSO）控制模型复杂度，体现了在捕捉信号与抑制噪声之间取得平衡的方法论。这种方法论与马克思主义认识论中“实践是检验真理的唯一标准”相契合。科学理论不仅要能够解释过去（样本内拟合），还要能够预测未来（样本外预测）。只有揭示内在本

质联系的理论，才具有真正的解释力和预测力。毛泽东和马克思主义认识论强调，理论的真理性需要通过实践检验来验证和完善。经济学模型同样需要接受样本外数据的检验，以修正不完全性或错误，这与自然科学和社会科学的发展规律一致。党的十八大以来，我国经济政策的科学性和有效性通过实践检验得到验证，体现了从实际出发、尊重客观规律的思想方法。这与案例中强调的理论与实践相结合的核心思想高度一致。

案例二　从集成学习看“大我”与“小我”的统一

在机器学习领域，集成学习（Ensemble Learning）是一类非常重要的技术路径，其核心思想是将多个预测能力有限、表现各异的“弱学习器”通过某种策略（如投票、加权平均、Boosting 等）组合在一起，形成一个性能更佳的“强学习器”。这种方法背后的原理并不是简单的数量叠加，而是强调个体差异性、多样性与协同机制的重要性。这与中华优秀传统文化中“群策群力”、“三人行，必有我师焉”的智慧不谋而合，也与社会主义核心价值观中倡导的“团结协作、集体主义”高度契合。集成学习的成功依赖于一个朴素但深刻的哲理：个体的有限可以被多样性的协同所超越。从技术逻辑来看，这体现了“系统大于部分之和”的思想；从思想政治教育的视角看，它则蕴含着丰富的育人价值：个体与集体的关系、个体价值与集体贡献的统一、合作共赢与多元包容的理念。

集成学习在设计上并不要求每个模型都完美——相反，只要它们在某些方面具备“相对独立性”与“适度多样性”，即便表现有限，也能在集体中发挥重要作用。这种设计理念为新时代青年提供了积极的心理暗示和发展方向：不完美的“我”也能在集体中找到位置；多样性是力量而非障碍；强化自我，也要补位他人。集成学习的技术逻辑与我国协同治理理念相通，二者可以相互促进；在科技攻坚的关键时期，技术发展需服务于国家战略，而人才培养需兼顾能力与价值观，以实现科技自立自强和社会治理现代化的目标。

案例三　从主成分分析到矛盾论：探寻金融数据背后的哲学智慧

在计量经济学与机器学习交叉的领域中，主成分分析（PCA）作为关键的降维技术，不仅在数据处理上发挥着重要作用，更蕴含着深刻的哲学思想，与马克

思主义哲学中的矛盾论有着紧密的联系。以十只股票（贵州茅台、格力电器等）的日收益率数据为例，对这些数据进行 PCA 分析，得到的第一个主成分与十只股票的平均收益高度相关，方差比例为 45.04%，被称为市场成分。它在所有单位线性组合中解释了最多的方差，是影响股票收益率的关键因素，后续主成分按照重要性依次递减。在股票收益率数据分析中，PCA 确定的主成分与矛盾论，即“抓住主要矛盾，区分次要矛盾”的思想高度契合。

矛盾论强调矛盾的普遍性（矛盾无处不在）和特殊性（不同矛盾在不同阶段作用不同）。在金融分析中，既要把握市场整体趋势（主要矛盾），也要关注局部特征（次要矛盾），做到“具体问题具体分析”。 在复杂系统（如金融市场）中，需运用矛盾论思维，优先解决主要矛盾，同时兼顾次要矛盾，以实现更精准的分析与决策。

课程设置知识要求：

学习本课程需要具备以下方面的知识基础：

1. 数学基础：理解和掌握机器学习算法需要学生对概率论、数理统计、线性代数、微积分和最优化等知识有所了解。了解这些知识不仅可以更好理解算法工作的原理；还能在遇到实际问题时能够对原始算法略加修正，使算法效率更高。

2. 编程技能：熟练掌握至少一种编程语言，如 Python 或 R，是必须的。本书的指定教材“Introduction to Statistical Learning with Applications in R”提供了所有常用的机器学习估计量的 R 代码。本书的参考书目《计量经济学编程：以 Python 语言为工具》可以作为 Python 语言用于计量编程的入门书。

3. 对数据预处理的能力。为了将机器学习以及其他估计方法用于实际数据。须了解如何清洗、转换和标准化数据，包括缺失值处理、异常值检测和降维等技术。

4. 持续学习的能力。机器学习是一个知识更新迭代十分快速的领域，新的算法和工具不断涌现。保持好奇心和学习的热情，不断更新你的知识库。

5. 实证研究项目经验：通过实际项目来应用所学知识，解决实际问题，积累经验。这有助于你更好地理解理论知识，并提高解决复杂问题的能力。

课程设置能力要求：

本课程取名“面向机器学习的计量经济学”。学完本课程后，学生应当对机器学习和计量经济学两个科学领域之间的关系有正确的理解。机器学习的任务在于实现精准预测和模式识别。尤其在处理大规模数据集时，机器学习作用尤为显著。计量经济学则聚焦于因果推断和经济理论的验证。其目标在于通过数据证实经济模型的有效性，并对变量间的因果关系进行量化分析。计量经济学不仅重视模型的预测能力，更注重对模型的可解释性，以及因果机制的深度理解。

学完本课程，学生可以获得以下方面的能力：

1. 数据分析能力：机器学习和计量经济学都依赖于强大的数据分析能力。通过运用回归、聚类、分类等技术来提取信息和进行预测。

2. 运用因果推断框架，辅之以合适的机器学习方法，评估某项政策产生的净效果。近年来，机器学习被大量引入因果推断领域，随机森林等方法开始应用于识别因果关系，拓展了计量经济学的研究视野。

3. 计量经济学中存在形形色色的模型。哪一种模型最适用于当前数据？这需要对模型进行评估。计量经济学通过标准误、假设检验等工具进行评估，而机器学习则常利用交叉验证（cross validation），预测均方误差等指标来考察模型的性能。

4. 通过编程实现对大数据的计算。机器学习在处理庞大数据和复杂模型方面具有优势，而计量经济学的传统方法受到计算资源的约束。但随着硬件算力与算法技术的发展，计量经济学能借助机器学习来应对更复杂的分析任务。

考核形式：

期末考试采用论文方式，学生的最后的总分计算方法如下：

课后作业	共三次作业，每次 10%，共 30%
考勤与课堂参与	10%
结课论文	60%

期末考试（结题论文）要求：

本课程以案例教学为主，期末考试不再采用闭卷笔试方式，而是要求学生在

一定选题范围内，提交一项使用机器学习方法处理数据（包含程序代码）的小型研究论文。期末论文的基本要求如下：

1. 提交的成果须写有题目、作者姓名、学号。
2. 本文打算研究一个什么实际问题？例如研究某个 X 对某个 Y 的因果效应，或者利用大量 X 预测 Y。简要说明研究这个问题的现实意义。
3. 研究使用什么数据？数据的描述性统计。
4. 研究中至少采用两种机器学习方法。鼓励采用更多种不同的机器学习方法加以研究（有助于提高最后分数）
5. 产生实证结果的代码放在文章附录中。
6. 要对不同方法得到的结果进行比较。
7. 必要的参考文献。

学术诚实

涉及学生的学术不诚实问题主要包括考试作弊；抄袭；伪造或不当使用在校学习成绩；未经老师允许获取、利用考试材料。对于学术不诚实的最低惩罚是考试给予 0 分。其它的惩罚包括报告学校相关部门并按照有关规定进行处理。

课程教学要点

第一篇：基于线性回归的机器学习方法

第 1 章：引言：从线性回归到模型选择

1.1 经典线性回归模型复习

1.2 预测推断与因果推断

1.3 评价模型与估计方法的优劣

1.4 交错鉴定法

1.5 案例：各国增长速率预测。哪种预测方法最好？

第 2 章：Lasso

2.1 解释变量个数 p 较大时的 OLS

2.2 Lasso

2.3 惩罚系数的选择

2.4 其他带惩罚项的线性回归

2.5 案例：用 Lasso, Ridge 与弹性网预测各国增长速率。

第 3 章：利用机器学习进行因果推断：双重 Lasso

3.1 在线性模型中推断单个解释变量的偏效应：偏回归

3.2 双重 Lasso

3.3 机器学习如何助力估计高维解释变量回归系数？Neyman 正交性条件

3.4 案例：条件收敛假说（conditional convergence）：穷国增长速率是否更快

第二篇：非线性机器学习方法

第 4 章：决策树

4.1 非线性拟合与决策树

4.2 最优划分特征选择

4.3 过拟合与剪枝

4.4 案例：用决策树方法估计中国企业生产函数

第 5 章：集成学习与随机森林

5.1 集成学习概念

5.2 基于决策树的集成学习

5.3 案例：用集成学习估计中国企业生产函数

第 6 章：神经网络与深度学习

6.1 MP 神经元与感知机

6.2 多层神经网络

6.3 案例：利用月度频率数据预测美国宏观经济景气度。

第 7 章：主成分分析与因子模型

7.1 主成分分析

7.2 统计因子模型

7.3 案例：中国股票市场的方差因子分解

第 8 章：聚类与 EM 算法

8.1 K-means 聚类

8.2 高斯混合模型

8.3 EM 算法的一般形式

8.4 案例：分类中国上市公司股票

第三篇：机器学习方法高级篇

第 9 章：卷积神经网络

9.1 卷积神经网络基本结构

9.2 卷积与池化

9.3 案例：卷积神经网络与遥感卫星图像结合预测非洲地区贫困水平。

第 10 章：状态空间模型

10.1 状态空间模型

10.2 线性高斯状态空间模型

10.3 非线性和非高斯状态空间模型

10.4 状态空间模型中的参数估计

10.5 案例：带有趋势项和季节项的公司每股盈利建模

第 11 章：抽样方法与 MCMC

11.1 抽样方法

11.2 马尔可夫链蒙特卡罗（MCMC）

11.3 案例：估计贵州茅台的随机波动率