机器学习与因果推断专题

1. 课程概览

A. 课程概要

- **时间**：2024 年 11 月 9-10 日；16-17 日
- **方式**：网络直播 + 45 天回放
- **授课教师**：司继春 (T1-T3) || 张宏亮 (T4-T6)
- **软件**：Python + R + Stata
- **PDF 课纲**：查看大纲　　预读资料
- **课程主页**：https://www.lianxh.cn/ML.html（往期答疑和板书）
- **报名链接**：https://www.wjx.top/vm/eiJVjQY.aspx#
- **助教招聘**：10 名, 截止时间：10 月 8 日，**点我报名**

B. 授课教师简介

司继春，上海财经大学博士，目前任教于上海对外经贸大学统计与信息学院，主要研究领域为微观计量经济学、产业组织理论，成果见诸 Journal of Business and Economic Statistics、《中国人口科学》、《系统工程理论与实践》等期刊。司老师专长于机器学习，尤其是基于机器学习的因果推断前沿方法，有多个大型数据分析项目的实战经验。业余时间里，司老师也经常在知乎上耐心作答，用通俗的语言普及统计和计量知识。他的知乎专栏名为「慧航」，关注者逾 31w，获赞超过 17w。他总能抽丝剥茧，把复杂的问题讲得清清楚楚。

张宏亮，美国麻省理工学院 (MIT) 博士，浙江大学经济学院新百人计划研究员，博士生导师。主要从事经济学微观实证研究，尤其偏爱因果推断方法在劳动经济学、公共经济学、发展经济学、城市经济学等领域的应用。研究成果见诸 International Economic Review (**IER**), Journal of European Economic Association (**JEEA**), Journal of Public Economics (**JPubE**), Journal of Development Economics (**JDE**, 2 篇), Journal of Urban Economics (**JUE**) 等专业领域顶级期刊。

C. 课程特色

- **懂原理、会应用**。本次课程邀请了两位老师合作讲授，目的在于最大限度地实现理论与应用的有机结合。为期四天的课程，分成两个部分：第一部分讲解常用的机器学习算法和适用条件，以及文本分析和大语言模型；第二部分通过精讲 4-6 篇发表于 Top 期刊的论文，帮助大家理解各类机器学习算法的应用场景，以及它们与传统因果推断方法的巧妙结合。
- **以 Top 期刊论文为范例**。目前多数人的困惑是不清楚如何将传统因果推断方法与机器学习结合起来。事实上，即便是 MIT 和 Harvard 的大牛们也都在「摸着石头过河」。为此，通过论文精讲和复现来学习这部分内容或许是目前最有效的方式了。张宏亮老师此前在浙江大学按照这一模式教授了「因果推断和机器学习」课程，效果甚佳：学生们能够逐渐建立起研究设计的理念，并在构造识别策略时适当地嵌入机器学习方法。

2. 课程详情

> **文献打包下载：** https://www.jianguoyun.com/p/DerGXTEQtKiFCBi1jeMFIAA

T1. 机器学习基础及无监督学习

随着大数据和机器学习技术的发展，学术研究中越来越多地应用这些工具来解决复杂问题，特别是在因果推断领域。通过利用大规模数据集，研究者不仅能够提升模型的预测能力，还能更有效地识别变量间的因果关系，估算政策的异质性效应。

本专题从机器学习的基础入手，明确其目标和关键问题，并区分监督学习和无监督学习两大类，以便大家对机器学习的整体架构有个全面的了解。同时，我们还将演示如何使用 scikit-learn 库来训练和评估机器学习模型。

本节内容将从两个方面展开。

其一，我们将从机器学习最基础也是最关键的性能度量出发，系统讲解模型的评估方法，进而讨论如何通过交叉验证等正则化技术来优化模型选择。学员将掌握预测效果的评估方式，这也是机器学习模型构建中的核心步骤。

其二，我们引入无监督学习技术，涵盖主成分分析、流形学习和聚类等方法。这些技术在数据预处理中起到重要作用，例如，主成分分析用于降维，而流形学习作为主成分分析的扩展，在更复杂的数据处理任务中具有应用价值。通过这些内容的学习，学员将理解机器学习的标准工作流程，从数据处理到模型构建的全过程。

- 预测问题与泛化能力
- 模型的评价方法
- 交叉验证和正则化
- 无监督学习：主成分分析和嵌入 (embedding)
- 无监督学习：聚类分析
- 案例：迁徙距离的嵌入分析

T2. 监督学习与因果推断

本专题讲解监督学习与因果推断的结合及应用。传统的因果推断方法在处理复杂、非线性数据面临诸多挑战，而监督学习技术如 Lasso、决策树和集成学习通过其强大的预测与拟合能力，能有效应对这些挑战。结合监督学习，特别是正则化技术，能够解决因果推断中的高维数据问题，从而提升模型的解释性与准确性。

课程内容从基础的线性回归和 Logistic 回归入手，逐步引入 **Lasso** 等正则化方法，Lasso 通过变量选择与约束处理高维数据，使得模型更具鲁棒性。接下来，我们将介绍基于 **树** 的方法，如决策树模型，这些模型能够有效捕捉数据中的复杂交互和非线性关系。随后，我们将讲解 **集成学习**，如随机森林和梯度提升树，它们通过结合多个弱学习器增强模型的表现。

课程还将重点介绍 **双重机器学习 (Double Machine Learning, DML)**，这一方法能够巧妙地将机器学习的预测能力应用于参数估计和统计推断。DML 在异质性处理效应的因果推断中表现尤为突出，能够更好地控制混杂因素，提升因果效应估计的精确度。

通过本课程，学员将学会如何应用监督学习方法来进行因果推断，掌握 Lasso、决策树、集成学习和 DML 等前沿方法。这将帮助学员在实际数据分析中，既能提升预测模型的效果，又能进行稳健的因果推断。

- 线性回归与 Lasso
- 分类问题与 Logistic 回归
- 决策树算法

- 集成学习与 Boosting
- 随机森林
- 双重机器学习的原理介绍
    - Double Selection Lasso 方法
    - Neyman 正交化
    - 交叉拟合 (cross fitting) 方法
- 处理效应与无混淆分配假设
- 基于 DML 的处理效应估计方法
- 案例：
    - **Deryugina**, T., Heutel, G., Miller, N. H., Molitor, D., & Reif, J. (2019). The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction. American Economic Review, 109(12), 4178–4219. Link (rep), PDF, Appendix, Google, -cited-.
- 参考文献：
    - **Chernozhukov**, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1), C1–C68. Link, PDF, Google, Replication.
    - **Gilchrist**, D. S., & Sands, E. G. (2016). Something to talk about: Social spillovers in movie consumption. *Journal of Political Economy*, 124(5), 1339-1382. Link, PDF, Google, -cited-.
    - **Knaus**, M. C., Lechner, M., & Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *The Econometrics Journal*, 24(1), 134-161. Link, PDF, Google, -cited-.
    - **Ahrens**, A., Hansen, C. B., Schaffer, M. E., & Wiemann, T. (2024). **ddml**: Double/debiased machine learning in Stata. **The Stata Journal**, 24(1), 3–45. Link, PDF, Google.
    - **Ahrens**, A., Hansen, C. B., & Schaffer, M. E. (2023). pystacked: Stacking generalization and machine learning in Stata. **The Stata Journal**, 23(4), 909–931. Link, PDF, Google.

T3. 文本分析与大语言模型 (LLMs)

随着文本数据在社会科学研究中的应用日益广泛，文本分析工具变得愈发重要。通过从文本中提取结构化信息，研究者能够识别情感、主题，甚至构建量化变量用于因果推断分析。

本专题首先介绍传统的文本分析技术，包括**词袋模型**、**TF-IDF**、**词嵌入**等方法，辅以 Python 中的强大工具 (如 NLTK、Jieba、Pandas、NumPy 和 Scikit-learn)。这些工具可以帮助我们从大量文本中提取有用的信息，如情感分析和文本相似性度量。同时，我们还将探讨如何通过这些方法构建与经济学或金融学相关的指数，如政策不确定性指数或情绪指数。

接着，我们将引入一些开源的 **大语言模型 (LLMs)**，如通义千问、Llama3.1 等，展示其在文本处理和变量提取中的应用。LLMs 在理解上下文和生成文本方面表现出色，尤其在处理大规模文本数据时，它们能够发现复杂的语义关系，有助于从文本中提取潜在的因果变量。

通过结合传统文本分析与大语言模型，学员将学习如何利用这些工具提升实证研究的深度和广度，从而为因果推断提供更加丰富的证据。

- 词袋模型与 TF-IDF
- 词嵌入与文本相似性度量
- 情感分析与不确定性指数

- 大语言模型的文本处理能力
- 大语言模型 (LLMs)
  - qwen-通义千问、Llama3.1 简介
  - 部署和使用方法
- 参考文献：
  - **Anand, V.**, Bochkay, K., Chychyla, R., & Leone, A. (2020). Using Python for Text Analysis in Accounting Research. Foundations and Trends? in Accounting, 14(3–4), 128–359. Link，PDF，Google.
  - **Benguria, F.**, Choi, J., Swenson, D. L., & Xu, M. J. (2022). Anxiety or pain? The impact of tariffs and uncertainty on Chinese firms in the trade war. *Journal of International Economics*, 103608. Link，PDF，Google.
  - **Gentzkow, M.**, Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3), 535–574. Link，PDF，Appendix，Google，-cited-.
  - **姚加权**, 冯绪, 王赞钧, 纪荣嵘, & 张维. (2021). 语调、情绪及市场影响: 基于金融情绪词典. 管理科学学报, 24(5), 26–46. -Link-，-PDF-

T4. 机器学习应用一：精准预测

预测问题和因果问题是经济政策制定与评估面临的两个核心问题。本单元的课程将首先理清经济学中预测问题和因果问题的本质差异及关联，并带领大家**探索机器学习在其擅长的预测领域的各类预测方法。**

在本单元的学习中，张宏亮老师将首先简要概括/回顾分类回归树 (Classification and Regression Tree，CART) 、 最小绝对值收敛和筛选算子 (Least Absolute Shrinkage and Selection Operation，LASSO) 、随机森林 (Random Forest) 、梯度提升算法 (Gradient Boosting) 和集成学习 (Essemble Learning) 等机器学习预测算法。然后，他将通过**多个实证案例来讲解上述算法的应用场景、假设条件和优劣。**

本专题将包括如下五个应用案例：

- Kleinberg et al. (AERPP, 2015)演示机器学习的Lasso模型如何应用于预测手术的术后预期寿命以 **实现有效避免术后预期寿命短的高风险患者接受该类手术。**
- Mullainathan & Spiess (JEP，2017) 将普通最小二乘法 (OLS) 和各种机器学习算法 (regression tree，lasso，random forest，ensemble method) 同时运用于预测美国的房价，并通过比较各个模型在样本外预测 (out-of-sample prediction) 的表现，**展示出机器学习算法在预测精准度上的改进。**
- Einav et al. (Science, 2018) 使用美国医疗保险 (Medicare) 赔付数据构建了一个预测老年人在未来 1 年内的死亡风险的包括lasso，random forest和grandient boosting的集成机器学习模型 (ensemble method)，并通过测算医疗开支与预测的死亡率之间的实证关系，**驳斥了坊间认为医疗保险开支过度浪费在短时间内注定死亡的病人身上的观点。**
- Kleinberg et al. (QJE, 2018) 使用 2008-2013 年纽约市法院保释申请的数据构建了一个gradient boosting的机器学习算法预测申请人保释后可能对社会带来的风险，发现 **基于机器学习算法预测的假释决定能帮助降低犯罪率。**
- Aiken et al. (Nature, 2022) 将机器学习的方法应用于手机大数据来快速识别 COVID-19 疫情的一项紧急现金援助计划的目标群体，凸显出 **机器学习方法在改善人道主义援助的针对性上的重要潜力。**

最后，张宏亮老师将指导大家对 Mullainathan & Spiess (JEP，2017) 文章的实证结果进行复现。

专题亮点：

- 理清预测问题和因果问题的本质差异及关联；
- 探索机器学习的精准预测之道；
- 机器学习算法 (CART, LASSO, Random Forest, Gradient Boosting, Ensemble) 的介绍；
- 机器学习算法在实证预测案例中的应用；
- Mullainathan & Spiess (2017, JEP) 的复现。

相关论文：

- Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction Policy Problems. **American Economic Review**, 105(5), 491–495. Link (rep), PDF, Appendix, Google.
- Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. **Journal of Economic Perspectives**, 31(2), 87–106. Link (rep), PDF, Appendix, Google.
- Einav, L., Finkelstein, A., Mullainathan, S., & Obermeyer, Z. (2018). Predictive Modeling of U.S. Health Care Spending in Late Life. **Science**, 360(6396), 1462–1465. Link, PDF, Google.
- Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human Decisions and Machine Predictions. **Quarterly Journal of Economics**. Link (rep), PDF, Google.
- Aiken, E., Bellue, S., Karlan, D., Udry, C., & Blumenstock, J. E. (2022). Machine Learning and Phone Data can Improve Targeting of Humanitarian Aid. **Nature**, 603(7903), 864–870. Link, PDF, Google.

T5. 机器学习应用二：增强因果推断

近年来，机器学习与因果推断方法的结合，使我们可以更好地识别、理解经济现象背后的因果关系，并不断实现政策优化。本专题的目标是：**在传统因果推断的基础上**，如何用机器学习方法增强对因果关系的分析、解读和预测。

本专题将介绍**若已经满足因果识别的前提条件**，如何借助机器学习展开更深入、更细致的分析。将重点介绍两个应用领域：

- **借助机器学习方法估算异质性处置效应。**基于此，我们可以进行更为细致的政策评价，如处置效应的地区差异、事变特征，也可以据此提出更有针对性的政策建议。
- **借助机器学习方法识别「有效的政策干预对象」。**其意义在于：一项政策对不同个体的作用效果存在差异，若能有针对性地实现政策的精准实施，则可以有效提高政策效果。

张宏亮老师将首先介绍 Wagner & Athey (JASA, 2018) 在随机森林 (Random Forest) 基础上提出的因果森林 (Causal Forest) 算法，并通过“家庭能源使用干预计划” (Knittel & Stolper, AERPP, 2021) 和“青年就业计划” (Davis & Heller, ReStat, 2020) 这两个随机实验案例来演示**机器学习辅助因果推断方法在估算异质性处置效应上的应用。**

在此基础上，他将进一步通过“美国职业安全与健康管理局”的随机检查对减少工作场所重大伤害发生率的现实案例 (Johnson et al., AEJ：Applied, 2023)介绍**机器学习辅助因果推断方法在优化政策干预对象选择、支持更有效的政策制定领域的应用**：

最后，他将指导大家对 Johnson et al.(AEJ：Applied, 2023) 这篇文章的实证结果进行复现。

专题亮点：

- 因果森林算法；
- 基于因果森林算法的「异质性处置效应估算」；
- 基于机器学习的「目标干预对象选择」；
- Johnson et al. (2023, AEJ) 的复现

相关论文：

- Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. **Journal of the American Statistical Association**, 113(523), 1228–1242. Link, PDF, Google.
- Davis, J. M. V., & Heller, S. B. (2020). Rethinking the Benefits of Youth Employment Programs: The Heterogeneous Effects of Summer Jobs. **Review of Economics and Statistics**, 102(4), 664–677. Link, PDF, Google.
- Knittel, C. R., & Stolper, S. (2021). Machine Learning about Treatment Effect Heterogeneity: The Case of Household Energy Use. **American Economic Review Papers and Proceedings**, 111, 440–444. Link (rep), PDF, Appendix, Google.

- Johnson, M. S., Levine, D. I., & Toffel, M. W. (2023). Improving Regulatory Effectiveness through Better Targeting: Evidence from OSHA. **American Economic Journal: Applied Economics**, 15(4), 30–67. Link (rep), PDF, Appendix, Google.

T6. 机器学习应用三：助力实现因果推断的前提条件

在与因果推断的结合上，机器学习方法除了可以增强传统因果推断方法的分析能力，还可以帮助满足实现因果推断所需的前提条件。本专题将介绍机器学习方法在 **模型构造** 和 **精准预测受政策影响的群体** 这两个领域助力因果推断条件的满足的应用。

如何选择控制变量和工具变量是传统因果推断中常常面临的难题，在高维数据中会更加棘手。**机器学习方法非常擅长应对此类问题**。在本专题中，张宏亮老师将首先以 Belloni et al. (RES, 2013) 和 Angrist & Frandsen (JoLE, 2022) 为基础，介绍机器学习如何应用于高维模型选择以满足实现因果推断的条件，包括：

- 高维线性回归模型 (含 DiD 模型) 中控制变量的选择；
- IV 模型中控制变量和工具变量的选择。

在此基础上，张宏亮老师将以 Lowes & Montero (AER, 2021) 为例，演示“后双重选择” (Post-Double-Selection, PDS) LASSO 方法在 OLS 模型和 IV 模型中选择控制变量上的应用，并指导大家对 Belloni et al. (JEP, 2014) 这篇文章的结果进行复现。

对于那些只对少数人产生影响的政策，如果不能将受影响的个体从样本中精准地区分出来，就导致评估结果失真、统计精度不足。为此，张宏亮老师将以 Cengiz et al. (JoLE，2022) 为例，介绍如何应用机器学习方法来 **精准预测受政策影响的群体**，从而实现**对仅影响少数人的政策进行有效的因果效应评估**。在这篇论文中，Cengiz et al.(2022) 首先通过机器学习方法建立了一个预测模型，预测出个体劳动者受最低工资上调影响的概率，进而挑选出受最低工资上调影响概率最高的 10% 的劳动者，最后结合事件研究法，分析「最低工资上调事件」对他们的工资水平和就业的影响。

专题亮点：

- 机器学习在高维模型选择中的应用；
- PDS Lasso 方法及其应用；
- 机器学习精准预测受政策影响的群体；
- 如何评估只对少数人有影响的政策？
- JEP (2014) 文章实证结果的复制。

相关论文：

- Belloni, A., Chernozhukov, V., & Hansen, C. (2013). Inference on Treatment Effects after Selection among High-Dimensional Controls. **Review of Economic Studies**, 81(2), 608–650. Link (rep), PDF, Google.
- Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-Dimensional Methods and Inference on Structural and Treatment Effects. **Journal of Economic Perspectives**, 28(2), 29–50. Link (rep), PDF, Appendix, Google.
- Lowes, S., & Montero, E. (2021). The Legacy of Colonial Medicine in Central Africa. **American Economic Review**, 111(4), 1284–1314. Link (rep), PDF, Appendix, Google.
- Angrist, J. D., & Frandsen, B. (2022). Machine Labor. **Journal of Labor Economics**, 40(S1), S97–S140. Link (rep), PDF, Appendix, Google.
- Cengiz, D., Dube, A., Lindner, A., & Zentler-Munro, D. (2022). Seeing beyond the Trees: Using Machine Learning to Estimate the Impact of Minimum Wages on Labor Market Outcomes. **Journal of Labor Economics**, 40(S1), S203–S247. Link (rep), PDF, Google.

3. 报名信息

- **主办方：** 太原君泉教育咨询有限公司
- **标准费用** (含报名费、材料费)：
  - **全价**：3600 元/班/人
- **优惠方案**：
  - **专题课/现场班老学员**：9 折，3240 元/人
  - **学生 (需提供学生证/卡照片)**：9 折，3240 元/人
  - **连享会会员：** 8.5 折，3060 元/人
- **温馨提示：** 以上各项优惠不能叠加使用。
- **联系方式：**
  - 邮箱： wjx004@sina.com
  - 王老师：18903405450 (微信同号)
  - 李老师：18636102467 (微信同号)

> **报名链接：** https://www.wjx.top/vm/eiJVjQY.aspx#

> 长按/扫描二维码报名：

缴费方式

> **方式 1：对公转账**

- 户名：太原君泉教育咨询有限公司
- 账号：35117530000023891 (晋商银行股份有限公司太原南中环支行)
- **温馨提示：** 对公转账时，请务必提供「**汇款人姓名-单位**」信息，以便确认。

> **方式 2：扫码支付**

> **温馨提示：**

- 扫码支付后，请将「**付款记录**」截屏发给王老师-18903405450（微信同号）

4. 听课指南

4.1 软件和课件

听课软件：支持手机，ipad ，平板以及 windows/Mac 系统的笔记本电脑，但**不支持台式机以及平板式的电脑。**

> **特别提示：**

- 为保护讲师的知识产权和您的账户安全，系统会自动在您观看的视频中嵌入您的「用户名」信息。
- 一个账号绑定一个设备，且听课电脑不能外接显示屏，请大家提前准备好自己的听课设备。
- 本课程为虚拟产品，**一经报名，不得退换。**
- 为保护知识产权，课程不允许以任何形式录屏及传播。

4.2 实名制报名

本次课程实行实名参与，具体要求如下：

- 高校老师/同学报名时需要向连享会课程负责人 **提供真实姓名，并附教师证/学生证图片**；
- 研究所及其他单位报名需提供 **能够证明姓名以及工作单位的证明**；
- 报名即默认同意「连享会版权保护协议条款」。

5. 助教招聘

- **名额：** 10 名
- **任务：**
  - **A. 课前准备**：完成 2 篇推文，风格参见连享会主页 www.lianxh.cn；
  - **B. 开课前答疑**：协助学员安装软件和使用课件，在微信群中回答一些常见问题；
  - **C. 上课期间答疑**：针对前一天学习的内容，在微信群中答疑 (8:00-9:00，19:00-22:00)；
  - Note: 下午 5:30-6:00 的课后答疑由主讲教师负责。
- **要求：** 热心、尽职，熟悉 Stata/Python/R 的基本语法和常用命令，能对常见问题进行解答和记录
- **特别说明：** 往期按期完成任务的助教可以直接联系连老师直录。
- **截止时间：** 2024 年 10 月 8 日 (将于 10 月 10 日公布遴选结果于 课程主页，及连享会主页 lianxh.cn)

> **申请链接：** https://www.wjx.top/vm/h4w7hQd.aspx#