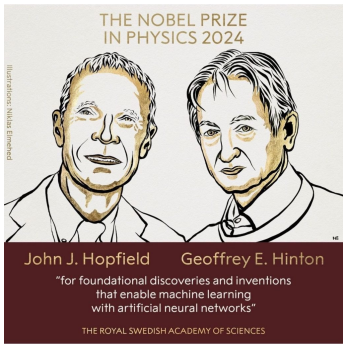
DOTE 6635: Artificial Intelligence for Business Research

Introduction

Renyu (Philip) Zhang

1

AI: Future of Human Civilization



华尔街见闻 首页 资讯 快讯 行情 日历 APP VIP会员 大师课 生活家

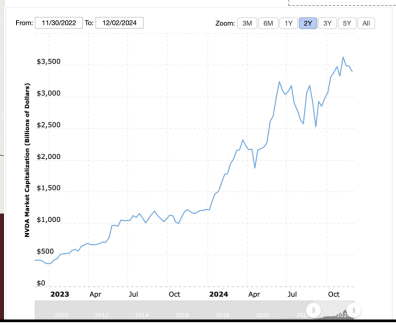
763 收藏

OpenAI完成最新一轮66亿美元融资 警告投资者不准支持马斯克xAI等劲敌

赵雨荷 10-03 00:07

摘要：

本轮融资也是史上规模最大的私人投资之一，由Thrive Capital领投，参与者还包括微软、英伟达、软银等，其中微软投资约7.5亿美元。本轮融资过后，OpenAI的估值达到1570亿美元，跻身全球前三大初创公司的行列。同时，OpenAI希望与投资者达成独家协议，防止马斯克的xAI和Anthropic等竞争对手获得战略合作机会和资本支持。



TechCrunch

Trump considers naming an 'AI czar'

Kyle Wiggers
Wed, November 27, 2024 at 1:59 PM PST • 1 min read

Incoming president Donald Trump is considering naming an "AI czar" in the White House, Axios reports.

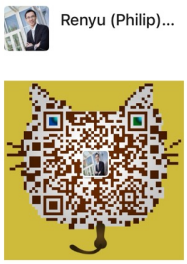
Should Trump appoint such a policy person, they'd be charged with helping to coordinate federal regulation and governmental use of AI. Importantly, an AI czar wouldn't require Senate confirmation, Axios notes — allowing them to get to work on the administration's goals faster.

2

2

Who Am I?

- A mostly harmless AI/data science scholar, teacher, and practitioner.
  - CUHK Business Professor & Kuaishou Economist
- How to use AI and data science to improve business decision making, especially for digitalized online platforms?



PKU (11') + WashU (16') Alum

NYUSH Ex-AP (16-21)

3

3

The Bitter Lesson

- References: http://www.incompleteideas.net/IncIdeas/BitterLesson.html https://www.youtube.com/watch?v=vbVfAqPI8ng
- The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin.
- Leveraging domain knowledge (short-term & specific) vs. Leveraging computation (long-term & general).
- Bitter lesson: Leveraging domain knowledge is self-satisfying and intellectually inspiring, but plateaus in the long-run or even inhibits further progress.

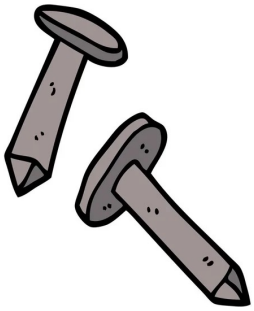
Prof. Richard Sutton

4

4

What is AI for Business Research?

Address or
Facilitate to Address



AI Technologies

Business Research Questions

5

5

What's Special About AI: Flywheel

Business Applications

Data

Compute

AI Models

We should work on the research that scales with data and compute.

6

6

Agenda

- **Course Introduction and Logistics**
- AI for Business Research Landscape

7

7

Purpose of this Course

1. Have a basic understanding of the fundamental concepts/methods in machine learning (ML) and artificial intelligence (AI) that are used (or potentially useful) in business research.
2. Understand how business researchers have utilized ML/AI and what managerial questions have been addressed by ML/AI in the recent decade.
3. Nurture a taste of what the state-of-the-art AI/ML technologies can do in the ML/AI community and, potentially, in your own research field.

8

8

What's New Beyond Last Year?

- Roughly 80%+ new content compared with last year.
  - Only the first two sessions (ML and DL introductions) are similar to those of last year.
- Topics: Large language models and AI-powered causal inference.
- I have learned more and deeper about AI as well ☺

9

9

Other Options to Learn AI

- To learn AI, you have a lot of other options:
  - Basic ML Intro by Andrew Ng: https://www.coursera.org/specializations/machine-learning-introduction
  - Basic Deep Learning (DL) Intro by Andrew Ng: https://www.coursera.org/specializations/deep-learning
  - Natural Language Processing by Chris Manning: https://web.stanford.edu/class/cs224n/
  - Computer Vision by Fei-Fei Li: http://cs231n.stanford.edu/
  - Deep Reinforcement Learning by Sergey Levine: https://rail.eecs.berkeley.edu/deeprlcourse/
  - Deep Learning Theory by Matus Telgarsky: https://mjt.cs.illinois.edu/courses/dlt-f22/
  - Machine Learning Fairness by Mortiz Hardt: https://fairmlbook.org/
  - Language Language Models by Danqi Chen: https://www.cs.princeton.edu/courses/archive/fall22/cos597G/
  - Short Courses on Generative AI: https://www.deeplearning.ai/short-courses/
  - See https://github.com/rphilipzhang/AI-PhD-S25 **for** more resources.

10

10

Why This Course?

- A fundamental and delegate trade-off: How much to cover vs. How deep to cover.
- This course provides a concise introduction to AI/ML topics relevant to applied business research.
- For each topic, we try to cover enough necessary knowledge that could:
  - Help you understand the key trade-offs and invent new applied methods (most likely without any theoretical guarantee);
  - Inform you about the literature development in the relevant domain;
  - Prepare you with the necessary sense to do rigorous business research using the relevant methods.
- We aim to cover conceptually important theories in AI/ML that can be applied in business research.
- We emphasize the combination of coding and theory so that you will be able to implement your ideas.

Impact of a CS Paper = Problem Importance * Technical Novelty * Performance Improvement

Impact of a Business Paper = Problem Importance * Identification Rigor * Insight Novelty

11

11

Why Not This Course?

- We have some assumptions on your prior knowledge:
  - Working knowledge in calculus, linear algebra, and stats;
  - Working knowledge of Python programming (but we have Cursor now…);
  - ML, causal inference, and econometrics: Better that you have some basic sense in them.
- We try to open doors and windows for you instead of preparing you to be a leading expert in a specific domain.
- I am trying my best to stay at the frontier, but some of the knowledge is outdated/constrained by academia.

Warning 0: At CUHK, we have an Econ course of similar topics (ECON 5180) without the coding emphasis.

Warning 1: This may be your MOST time-consuming course at CUHK by a wide margin.

Warning 2: We will mainly talk about the ideas and methods (with demos) in class, but you will need some coding skills to finish your homework and replication project.

12

12

Course Format

- We have a 2-hrs-and-45mins long course each week.
- For each session:
  - 15 mins: Homework discussions and review of previous content;
  - 105 mins: Theories and coding demos;
  - 30 mins: Student presentations.
- All coursework will be done in groups of at most TWO students.
  - Register your group members (and majors) and your group name by 11:59pm, Jan. 15, 2024.
  - Otherwise, we will match you with others (based on majors).
- You will need to evaluate your group mate's contribution in all the coursework.

13

13

Coursework and Grading

- Coursework:
  - Lecture notes scribing (each group will scribe the lecture note of one session/topic)
  - Paper replication and presentation (one paper replication and presentation per group each week)
  - Homework (one coding assignment each week, due two weeks after distribution; 5 assignments count)
  - Final Project (one final project based on your own choice).
- Grading:
  - See Syllabus.
- All homework/final project will be done in Python.

14

14

Coursework Materials

- GitHub: https://github.com/rphilipzhang/AI-PhD-S25
  - All course materials will be distributed on this GitHub Repository.
- Google Sheet: https://docs.google.com/spreadsheets/d/1ffRNlSFqki4vomz5UN0OseFNEoEXBY5kptS79wmgWs4/edit?usp=sharing
  - Group Registration
  - Lecture Notes Scribing Sign-up
  - Paper Replication and Presentation Sign-up
  - Project Presentation Sign-up
  - Homework Submission (use the link to your Google CoLab and opensource your code to your classmates by "Anyone with the link can view")
- Google CoLab: https://drive.google.com/drive/folders/1lYO4ni5B5AVkYZ3qVrVs2LoWxHProsxM
  - All code demos will be distributed via Google CoLab.
- Registered students please ask our TA, Leo Cao, to add your account to our course Google Sheet.

15

15

Course Communications

- Class Meeting: Tuesday, 12:30AM-3:15PM (@WMY 504)
- Office hour: By appointment, @CYT_911
- WeChat group: Online discussion forum.
- Instructor contact
  - Office: CYT_911
  - Email: philipzhang@cuhk.edu.hk
  - Tel: 852-3943-7763
  - WeChat: rphilip_zhang
- Teaching Assistant: Leo Cao
  - Email: yinglyucao@cuhk.edu.hk

16

16

Python Tutorial Sessions

- We have two optional Python tutorial sessions held online at Friday night, 7:00pm-9:00pm.
- Tutorial Instructor: Xinyu Li, MIS PhD Candidate @CUHK Business School, xinyu.li@link.cuhk.edu.hk
- Check the course GitHub Repo for CoLab and Zoom links.
- Session 1: Friday, Jan 17, 2024
  - Python Basics
- Session 2: Friday, Jan 24, 2024
  - PyTorch Basics & DOT Server
- Other References:
  - https://colab.research.google.com/drive/1hxWtr98jXqRDs_rZLZcEmX_hUcpDLq6e?usp=sharing
  - https://colab.research.google.com/drive/13HGy3-uIIy1KD_WFhG4nVrxJC-3nUUkP?usp=sharing
  - https://cs231n.github.io/python-numpy-tutorial/
  - https://colab.research.google.com/github/cs231n/cs231n.github.io/blob/master/python-colab.ipynb

17

17

Agenda

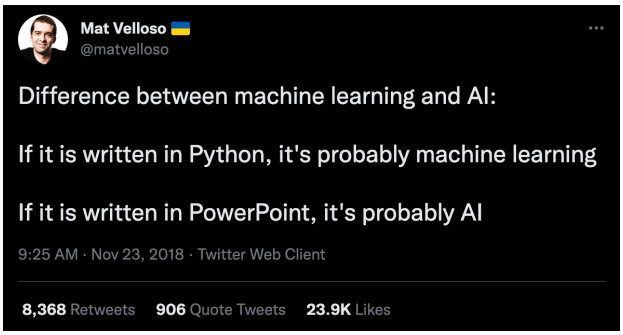
- Course Introduction and Logistics
- **AI for Business Research Landscape**

18

18

What is AI/ML?

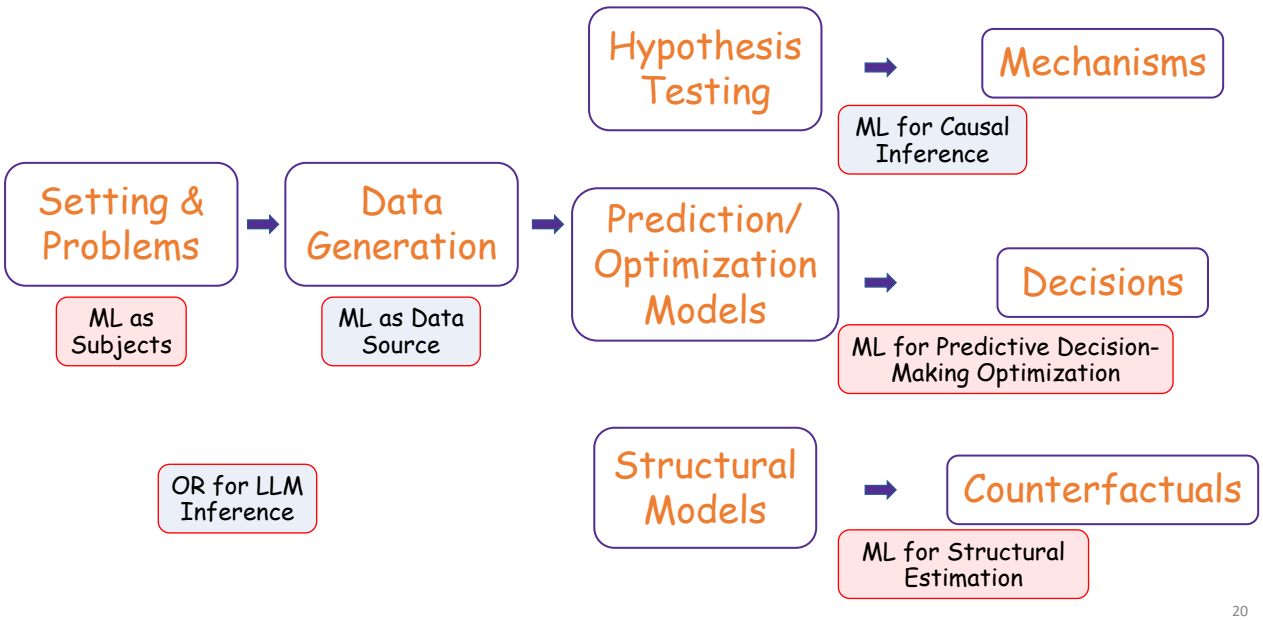
- ML is a CS subfield that automates computers to learn from data without explicitly programmed.
- Different names:
  - Data mining
  - Statistical learning
  - Data science



19

19

A Typical Applied Business Research Paper



20

20

Landscape of AI/ML for Business Research

- ML as Data/Data Source
  - Cohen M, Zhang R, Jiao K. (2022) Data aggregation and demand prediction. *Operations Research*, 70(5): 2597-2618.
- ML for Causal Inference
  - Ye, Z., Zhang, Z, Zhang, D. J., Zhang, H., Zhang, R. (2023) Deep Learning Based Causal Inference for Large-Scale Combinatorial Experiments: Theory and Empirical Evidence, *working paper* and *EC'23*.
- ML for Predictive Decision Making and Optimization
  - Ye, Z., Zhang, D. J., Zhang, H., Zhang, R., Chen, X., and Xu, Z. (2023) Cold start to improve market thickness on online advertising platforms: Data-driven algorithms and field experiments. *Management Science*, 69(7), 3838-3860.
- ML as Subjects
  - Zhang, X., Sun, C., Zhang, R., and Goh, K-Y (2024) The Value of AI-Generated Metadata for UGC Platforms: Evidence from a Large-scale Field Experiment, *working paper* and *CIST 2024*.
- ML for Structural Estimation

21

21

ML as Data/Data Source

Home > Marketing Science > Ahead of Print >

Engagement That Sells: Influencer Video Advertising on TikTok

Jeremy Yang, Juanjuan Zhang, Yuhan Zhang

Published Online: 20 Dec 2024 | https://doi.org/10.1287/mksc.2021.0107

Abstract

Many ads are engaging, but what makes them engaging may have little to do with the product. This problem can be particularly relevant to influencer advertising if influencers are motivated to promote themselves, not just the product. We develop an algorithm to measure the degree of effective engagement associated with the product and use it to predict the sales lift of influencer video advertising. We propose the concept of the product engagement score (PE score) to capture how engaging the product itself is as presented in a video. We estimate pixel-level engagement as a saliency map by training a deep three-dimensional convolutional neural network on video-level engagement data. We locate pixel-level product placement with an object detection algorithm. The PE score is computed as the pixel-level, engagement-weighted product placement in a video. We construct and validate the algorithm with influencer video ads on TikTok and product sales data on Taobao. We use variation in video posting time to identify video-specific sales lift and show that the PE score significantly and robustly predicts sales lift. We explore drivers of engagement and discuss how various stakeholders in influencer advertising can use the PE score in a scalable way to manage content, align incentives, and improve efficiency.

History: Olivier Toubia served as the senior editor.

Financial Machine Learning

Bryan Kelly[1] and Dacheng Xiu[2]

[1] *Yale School of Management, AQR Capital Management, and NBER; bryan.kelly@yale.edu*
[2] *University of Chicago Booth School of Business; dacheng.xiu@chicagobooth.edu*

ABSTRACT

We survey the nascent literature on machine learning in the study of financial markets. We highlight the best examples of what this line of research has to offer and recommend promising directions for future research. This survey is designed for both financial economists interested in grasping machine learning tools, as well as for statisticians and machine learners seeking interesting financial contexts where advanced methods may be deployed.

22

22

ML as Data Source

- Any recordable information that is not numerical can be analyzed with ML to answer business questions.

- References:
  - Text – Natural Language Processing (NLP)
  - Image/Video – Computer Vision (CV)
  - Sound – Deep Learning (DL)
  - Genetic information - Bioinformatics
  - And many more…

23

23

ML as Data Source

- Why do we use ML to understand unstructured data?
  - Cost reduction and scalability
  - Objectivity
  - Easy to built into other systems

- Issues with using ML to understand unstructured data:
  - Measurement errors
  - Interpretation

24

24

Issues with ML as Data Source

- Empirical model: $Y = a + b \cdot D + g(X) + \epsilon$
  - Key parameter of interest: b
- Outcome
  - Y may be generated through ML with error (of less concern).
- Treatment
  - D may be generated through ML with error which is correlated with ϵ
  - https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4480696; https://arxiv.org/abs/2402.15585
- Controls
  - X may be generated through ML with error.
  - X may be selected by ML with error.
  - Double machine learning can be applied for effective debias.

25

25

ML for Causal Inference

Journal of Marketing Research
Volume 61, Issue 3, June 2024, Pages 472-495

https://doi.org/10.1177/00222437231210267

Sage Journals

Article

Mega or Micro? Influencer Selection Using Follower Elasticity

Zijun Tian, Ryan Dew, and Raghuram Iyengar

Abstract

Influencer marketing, in which companies sponsor social media personalities to promote their brands, has exploded in popularity in recent years. One common criterion for selecting an influencer partner is popularity. While some firms collaborate with "mega" influencers with millions of followers, other firms partner with "micro" influencers with only several thousand followers, but who also cost less to sponsor. To quantify this trade-off between popularity and cost, the authors develop a framework for estimating the follower elasticity of impressions (FEI), which measures a video's percentage gain in impressions (i.e., views) corresponding to a percentage increase in the number of followers of its creator. Computing FEI involves estimating the causal effect of an influencer's popularity on the view counts of their videos, which is achieved through a combination of (1) a unique data set collected from TikTok, (2) a representation learning model for quantifying video content, and (3) a machine learning–based causal inference method. The authors find that FEI is always positive, averaging .10, but often nonlinearly related to follower size. They examine the factors that predict variation in these FEI curves and show how firms can use these results to better determine influencer partnerships.

Keywords

influencer marketing, causal inference, deep learning, representation learning, heterogeneous treatment effects, video data

https://causalml-book.org/

informs
https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–15
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Targeting for Long-Term Outcomes

Jeremy Yang,[a,*] Dean Eckles,[b,*] Paramveer Dhillon,[c] Sinan Aral[b]

[a] Harvard Business School, Boston, Massachusetts 02163; [b] Massachusetts Institute of Technology, Cambridge, Massachusetts 02142; [c] University of Michigan, Ann Arbor, Michigan 48109
*Corresponding authors
Contact: jeryang@hbs.edu, https://orcid.org/0000-0001-8639-5493 (JY); eckles@mit.edu, https://orcid.org/0000-0001-8439-442X (DE); dhillonp@umich.edu, https://orcid.org/0000-0002-0994-9488 (PD); sinan@mit.edu, https://orcid.org/0000-0002-2762-058X (SA)

Received: October 7, 2020
Revised: February 12, 2022
Accepted: February 27, 2022
Published Online in Articles in Advance: August 3, 2023

https://doi.org/10.1287/mnsc.2023.4881

Abstract. Decision makers often want to target interventions so as to maximize an outcome that is observed only in the long term. This typically requires delaying decisions until the outcome is observed or relying on simple short-term proxies for the long-term outcome. Here, we build on the statistical surrogacy and policy learning literatures to impute the missing long-term outcomes and then approximate the optimal targeting policy on the imputed outcomes via a doubly robust approach. We first show that conditions for the validity of average treatment effect estimation with imputed outcomes are also sufficient for valid policy evaluation and optimization; furthermore, these conditions can be somewhat relaxed for policy optimization. We apply our approach in two large-scale proactive churn management experiments at *The Boston Globe* by targeting optimal discounts to its digital subscribers with the aim of maximizing long-term revenue. Using the first experiment, we evaluate this approach empirically by comparing the policy learned using imputed outcomes with a policy learned on the ground-truth, long-term outcomes. The performance of these two policies is statistically indistinguishable, and we rule out large losses from relying on surrogates. Our approach also outperforms a policy learned on short-term proxies for the long-term outcome. In a second field experiment, we implement the optimal targeting policy with additional randomized exploration, which allows us to update the optimal policy for future subscribers. Over three years, our approach had a net-positive revenue impact in the range of $4–$5 million compared with the status quo.

History: Accepted by Eric Anderson, marketing.
Funding: This work was supported by Boston Globe Media.
Supplemental Material: The online appendix and data are available at https://doi.org/10.1287/mnsc.2023.4881.

Keywords: long-term effect • statistical surrogate • policy learning • targeting • proactive churn management

https://bookdown.org/stanfordgsbsilab/ml-ci-tutorial/

26

26

ML for Predictive Decision-Making &Optimization

informs
http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 70, No. 1, January–February 2022, pp. 309–328
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Crosscutting Areas

Customer Choice Models vs. Machine Learning: Finding Optimal Product Displays on Alibaba

Jacob Feldman,[a] Dennis J. Zhang,[a] Xiaofei Liu,[b] Nannan Zhang[b]

[a] Olin Business School, Washington University in St. Louis, St. Louis, Missouri 63130; [b] Alibaba Group Inc., Hangzhou 311100, China
Contact: jbfeldman@wustl.edu, https://orcid.org/0000-0002-5576-1953 (JF); denniszhang@wustl.edu (DJZ); xiaofei.liu@alibaba.com (XL); nannan.zhang@alibaba.com (NZ)

Received: November 5, 2019
Revised: November 18, 2019; August 25, 2020; February 2, 2021
Accepted: April 6, 2021
Published Online in Articles in Advance: October 26, 2021

Area of Review: OR Practice

https://doi.org/10.1287/opre.2021.2158

Abstract. We compare the performance of two approaches for finding the optimal set of products to display to customers landing on Alibaba's two online marketplaces, Tmall and Taobao. We conducted a large-scale field experiment, in which we randomly assigned 10,421,649 customer visits during a one-week-long period to one of the two approaches and measured the revenue generated per customer visit. The first approach we tested was Alibaba's current practice, which embeds product and customer features within a sophisticated machine-learning algorithm to estimate the purchase probabilities of each product for the customer at hand. The products with the largest expected revenue (revenue × predicted purchase probability) are then made available for purchase. Our second approach, which we developed and implemented in collaboration with Alibaba engineers, uses a featurized multinomial logit (MNL) model to predict purchase probabilities for each arriving customer. We used historical sales data to fit the MNL model, and then, for each arriving customer, we solved a cardinality-constrained assortment-optimization problem under the MNL model to find the optimal set of products to display. Our field experiments revealed that the MNL-based approach generated 5.17 renminbi (RMB) per customer visit, compared with the 4.04 RMB per customer visit generated by the machine-learning-based approach when both approaches were given access to the same set of the 25 most important features. This improvement represents a 28% gain in revenue per customer visit, which corresponds to a 4 million RMB improvement over the week in which the experiments were conducted. Motivated by the results of our initial field experiment, Alibaba then implemented a full-featured version of our MNL-based approach, which now serves the majority of customers in this setting. Using another small-scale field experiment, we estimate that our new MNL-based approach that utilizes the full feature set is able to increase Alibaba's annual revenue by 87.26 million RMB (12.42 million U.S. dollars).

Supplemental Material: The online appendix is available at https://doi.org/10.1287/opre.2021.2158.

Keywords: choice models • product assortment • machine learning • field experiment • retail operations

informs
https://pubsonline.informs.org/journal/mksc

MARKETING SCIENCE
Articles in Advance, pp. 1–22
ISSN 0732-2399 (print), ISSN 1526-548X (online)

Dynamic Coupon Targeting Using Batch Deep Reinforcement Learning: An Application to Livestream Shopping

Xiao Liu[a]

[a] Stern School of Business, New York University, New York, New York 10012
Contact: xl23@stern.nyu.edu, https://orcid.org/0000-0002-7093-8534 (XL)

Received: September 3, 2020
Revised: September 5, 2021; April 17, 2022
Accepted: June 23, 2022
Published Online in Articles in Advance: October 20, 2022

https://doi.org/10.1287/mksc.2022.1403

Abstract. We present an empirical framework for creating dynamic coupon targeting strategies for high-dimensional and high-frequency settings, and we test its performance using a large-scale field experiment. The framework captures consumers' intertemporal tradeoffs associated with dynamic pricing and does not rely on functional form assumptions about consumers' decision-making processes. The model is estimated using batch deep reinforcement learning (BDRL), which relies on Q-learning, a model-free solution that can mitigate model bias. It leverages deep neural networks to represent the high-dimensional state space and alleviate the curse of dimensionality. The empirical application is in a multibillion-dollar livestream shopping context. Our BDRL solution increases the platform's revenue by twice as much as static targeting policies and by 20% more than the model-based solution. The comparative advantage of BDRL comes from more effective and automatic targeting of consumers based on both heterogeneity and dynamics, using exceptionally rich, nuanced differences among consumers and across time. We find that price skimming, reducing discounts for attractive hosts, and increasing the coupon discount level at a faster rate for low spenders are effective strategies based on dynamics, consumer heterogeneity, and the two combined, respectively.

History: K. Sudhir served as the senior editor and John Hauser served as associate editor for this article.
Funding: Partial financial support was received from the NYU Center for Global Economy and Business.
Supplemental Material: The data files and online appendices are available at https://doi.org/10.1287/mksc.2022.1403.

Keywords: dynamic pricing • coupon • deep reinforcement learning • reference price • livestream shopping • targeting

27

27

ML as Subject

RESEARCH ARTICLE | CHATGPT

Experimental evidence on the productivity effects of generative artificial intelligence

SHAKKED NOY AND WHITNEY ZHANG Authors Info & Affiliations

SCIENCE · 13 Jul 2023 · Vol 381, Issue 6654 · pp. 187-192 · DOI: 10.1126/science.adh2586

62,070 2

Editor's summary

Automation has historically displaced human workers in factories (e.g., automotive manufacturing) or in performing routine computational tasks. Will generative artificial intelligence (AI) tools such as ChatGPT disrupt the labor market by making educated professionals obsolete, or will these tools complement their skills and enhance productivity? Noy and Zhang examined this issue in an experiment that recruited college-educated professionals to complete incentivized writing tasks. Participants assigned to use ChatGPT were more productive, efficient, and enjoyed the tasks more. Participants with weaker skills benefited the most from ChatGPT, which carries policy implications for efforts to reduce productivity inequality through AI. —EEU

informs
http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Vol. 65, No. 7, July 2019, pp. 2966–2981
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads

Anja Lambrecht,[a] Catherine Tucker[b]

[a] Marketing, London Business School, London NW1 4SA, United Kingdom; [b] Marketing, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142
Contact: alambrecht@london.edu, http://orcid.org/0000-0001-6766-1602 (AL); cetucker@mit.edu, http://orcid.org/0000-0002-1847-4832 (CT)

Received: November 28, 2017
Revised: March 2, 2018
Accepted: March 13, 2018
Published Online in Articles in Advance: April 10, 2019

https://doi.org/10.1287/mnsc.2018.3093

Abstract. We explore data from a field test of how an algorithm delivered ads promoting job opportunities in the science, technology, engineering and math fields. This ad was explicitly intended to be gender neutral in its delivery. Empirically, however, fewer women saw the ad than men. This happened because younger women are a prized demographic and are more expensive to show ads to. An algorithm that simply optimizes cost-effectiveness in ad delivery will deliver ads that were intended to be gender neutral in an apparently discriminatory way, because of crowding out. We show that this empirical regularity extends to other major digital platforms.

History: Accepted by Joshua Gans, business strategy.
Funding: Supported by a National Science Foundation Career Award [Grant 6923256].

Keywords: algorithmic bias • online advertising • algorithms • artificial intelligence

AI/ML as subjects: Economics of AI; Machine human collaboration; ML fairness/discrimination; ML and labor market; data privacy; Data and ML in IO; AI as a species, etc.

28

28

ML for Structural Estimation

Econometrica, Vol. 91, No. 6 (November, 2023), 2041–2063

AN ADVERSARIAL APPROACH TO STRUCTURAL ESTIMATION

TETSUYA KAJI
University of Chicago Booth School of Business

ELENA MANRESA
Department of Economics, New York University

GUILLAUME POULIOT
University of Chicago Harris School of Public Policy

We propose a new simulation-based estimation method, adversarial estimation, for structural models. The estimator is formulated as the solution to a minimax problem between a generator (which generates simulated observations using the structural model) and a discriminator (which classifies whether an observation is simulated). The discriminator maximizes the accuracy of its classification while the generator minimizes it. We show that, with a sufficiently rich discriminator, the adversarial estimator attains parametric efficiency under correct specification and the parametric rate under misspecification. We advocate the use of a neural network as a discriminator that can exploit adaptivity properties and attain fast rates of convergence.

KEYWORDS: Structural estimation, generative adversarial networks, neural networks, simulation-based estimation, efficient estimation.

Home > Marketing Science > Ahead of Print >

Estimating Parameters of Structural Models Using Neural Networks

Yanhao (Max) Wei, Zhenling Jiang

Published Online: 16 Aug 2024 | https://doi.org/10.1287/mksc.2022.0360

Abstract

We study an alternative use of machine learning. We train neural nets to provide the parameter estimate of a given (structural) econometric model, for example, discrete choice or consumer search. Training examples consist of datasets generated by the econometric model under a range of parameter values. The neural net takes the moments of a dataset as input and tries to recognize the parameter value underlying that dataset. Besides the point estimate, the neural net can also output statistical accuracy. This neural net estimator (NNE) tends to limited-information Bayesian posterior as the number of training datasets increases. We apply NNE to a consumer search model. It gives more accurate estimates at lighter computational costs than the prevailing approach. NNE is also robust to redundant moment inputs. In general, NNE offers the most benefits in applications where other estimation approaches require very heavy simulation costs. We provide code at: https://nnehome.github.io.

History: Manchanda Puneet served as the senior editor.

29

29

OR for LLM Inference

Foundational Model in Large Language Model (LLM) Inference: Online Batching and Scheduling

Ishai Menache
Microsoft Research, ishai@microsoft.com

Konstantina Mellou
Microsoft Research, kmellou@microsoft.com

Marco Molinaro
Microsoft Research, mmolinaro@microsoft.com

Zijie Zhou
Massachusetts Institute of Technology, zhou98@mit.edu

In the rapidly evolving field of artificial intelligence, Large Language Models (LLMs) play a pivotal role across various applications, driving the need for efficient computational strategies for LLM inference. LLM inference, the process by which a trained model generates text one word at a time in response to input prompts, is both costly and energy-intensive, consuming substantial amounts of electricity and water. Optimizing LLM inference can significantly reduce these costs and promote sustainability. This paper presents the pioneering **Operations Research for LLM** methodology, which optimizes the online batching and scheduling of LLM inference tasks within a single GPU worker. We introduce the first foundational model that accurately reflects the dynamics of LLM inference, addressing challenges such as job precedence and unpredictable token generation. Our analysis reveals that the commonly used Greedy Prioritize Prompt algorithm has a regret bound of $\Omega(T)$, where T represents the number of requests. We then propose a novel algorithm, *Opportunity Cost Coupling* (OCC), which achieves an $O(1)$ regret under the assumption of identical prompt sizes. The proof involves a technique called compensated coupling. This work is the first to apply this proof technique to the adversarial arriving setting. Numerical simulations on a public conversation dataset demonstrate that OCCmaintains constant average latency, unlike other algorithms which exhibit an at least linearly increasing trend in average latency.

Fundamental Modeling for LLM Inference with Exploding KV Cache Demands

Patrick Jaillet
Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, jaillet@mit.edu

Jiashuo Jiang
HKUST, jsjiang@ust.hk

Chara Podimata
Sloan School of Management, Massachusetts Institute of Technology podimata@mit.edu

Zijie Zhou*
Operations Research Center, Massachusetts Institute of Technology, zhou98@mit.edu

In the rapidly advancing field of artificial intelligence, Large Language Models (LLMs) are crucial for many applications, demanding efficient computational strategies for inference. LLM inference, where a trained model generates text one word at a time in response to prompts, is resource-intensive, consuming significant electricity and water. This paper models LLM inference, focusing on reducing redundant computations through a special memory-saving mechanism. This mechanism temporarily stores information from each word the model processes into the KV (key-value) cache to avoid recalculating it repeatedly. However, as more words are processed, this storage can quickly reach its limit. When this happens, the system incurs substantial extra costs by reprocessing tasks. We optimize batching and scheduling strategies to manage KV cache memory usage and minimize the inference latency to improve efficiency and sustainability.

We address this challenge by first analyzing a semi-online model, where all prompts arrive initially and must be processed sequentially. For this case, we develop a polynomial-time algorithm that achieves exact optimality. Next, we examine the fully online setting with sequential prompt arrivals. For adversarial sequences, we demonstrate that no algorithm can achieve a constant competitive ratio. For stochastic arrivals, we present a fast algorithm that guarantees constant regret, using a novel framework based on compensated coupling to prove it.

Finally, we use the Vidur simulator on a public conversation dataset Zheng et al. (2023) to compare our algorithm with parametrized benchmark algorithms on 2 linked A100 GPUs with the Llama-70B model. After optimizing the benchmark parameters, we find that in high-demand scenarios, our algorithm's average latency increases only one-third as fast as the best benchmark, and in low-demand cases, it grows at one-eighth the rate. From a practical perspective, meeting a given average latency requirement in the high-demand setting would require over 8 A100 GPUs for the best benchmark, while our algorithm achieves this with only 2 GPUs, substantially reducing costs and energy use, and promoting sustainability in LLM deployment.

30

30

Tentative Course Schedule

- Introduction to Supervised Learning (1)
- Introduction to Deep Learning (1)
- Large Language Models (4)
- Causal Inference (4)
- Economics and Ethics of AI (1)

Note: Tentative schedule subject to changes. See Syllabus and GitHub repo for details.

31

31

Who Are You?

- What is your name?
- Which department are you from?
- Why are you here?
- What do you expect from this course?
- What else do you want me to cover?

32

32