

ReliableML 2024

технические итоги года

Дмитрий Колодезев @ ДатаЕлка 2025-01-18
Промсофт, Новосибирск

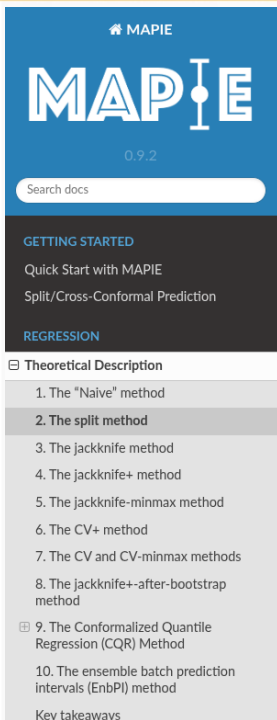
О чем мы тут

- Ведем канал https://t.me/reliable_ml
- Организовываем секцию ReliableML
- Раньше — про интерпретируемость и АБ тесты
- В этом году еще и Карьеру и Датастратегию
- Как (аккуратно) повергнуть ML лицом к бизнесу
- Взгляд с разных сторон — я больше из окопов
- Ира больше с капитанского мостика

Conformal Prediction

- На небесах только и говорят о Conformal Prediction
- Вы тоже говорите
- Оценка неопределенности
 - Для классификатора — подмножество классов
 - Для регрессии - диапазон
- Гарантированно покрывает
- Не делает предположений о распределении

Conformal Prediction - 2



MAPIE 0.9.2

Search docs

GETTING STARTED

Quick Start with MAPIE

Split/Cross-Conformal Prediction

REGRESSION

- Theoretical Description
- 1. The "Naive" method
- 2. The split method
- 3. The jackknife method
- 4. The jackknife+ method
- 5. The jackknife-minmax method
- 6. The CV+ method
- 7. The CV and CV-minmax methods
- 8. The jackknife+ -after- bootstrap method
- 9. The Conformalized Quantile Regression (CQR) Method
- 10. The ensemble batch prediction intervals (EnBP) method

Key takeaways

» Theoretical Description Regression : contents

[View page source](#)

Theoretical Description

The `mapie.regression.MapieRegressor` class uses various resampling methods based on the jackknife strategy recently introduced by Foygel-Barber et al. (2020) [1]. They allow the user to estimate robust prediction intervals with any kind of machine learning model for regression purposes on single-output data. We give here a brief theoretical description of the methods included in the module.

Before describing the methods, let's briefly present the mathematical setting. For a regression problem in a standard independent and identically distributed (i.i.d) case, our training data $(X, Y) = \{(x_1, y_1), \dots, (x_n, y_n)\}$ has an unknown distribution $P_{X,Y}$. We can assume that $Y = \mu(X) + \epsilon$ where μ is the model function we want to determine and $\epsilon_i \sim P_{\epsilon|X}$ is the noise. Given some target quantile α or associated target coverage level $1 - \alpha$, we aim at constructing a prediction interval $\hat{C}_{n,\alpha}$ for a new feature vector X_{n+1} such that

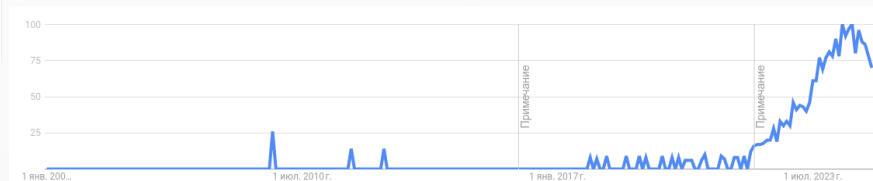
$$P\{Y_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1})\} \geq 1 - \alpha$$

All the methods below are described with the absolute residual conformity score for simplicity but other conformity scores are implemented in MAPIE (see [Theoretical Description for Conformity Scores](#)).

1. The "Naive" method

The so-called naive method computes the residuals of the training data to estimate the typical error obtained on a new test data point. The prediction interval is therefore given by the prediction obtained by the model trained on the entire training set \pm the quantiles of the conformity scores of the same training set:

- Любые модели
- Любые распределения
- Статистические гарантии покрытия
- Магия
- На собеседах дуреют от этой прикормки



<https://mapie.readthedocs.io/en/latest/>

Causal Inference в массы

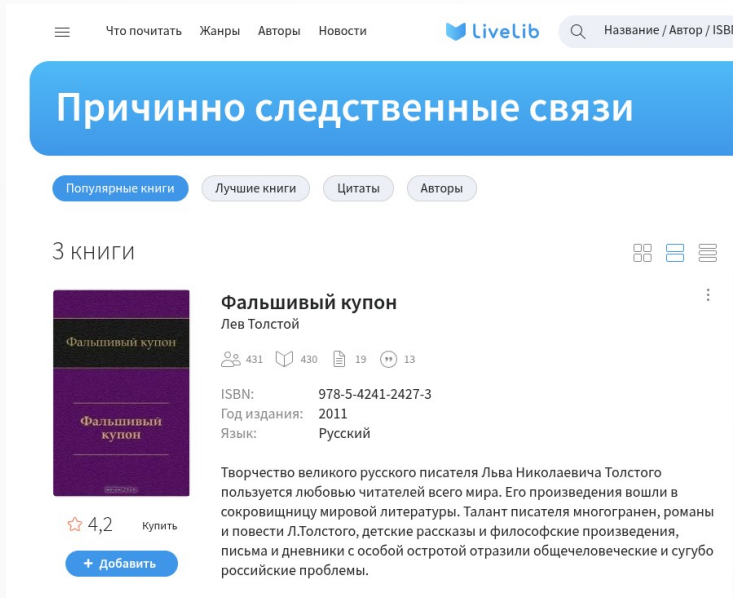
- Год назад все тюнили лам
- Теперь эластичность цены
- Эффект от коммуникаций
- Объем продаж, если ...
- Отдача от рекламы
- Линейная, брат, регрессия



То есть АБ тесты - это всего лишь OLS-регрессия? — Иван Комаров

Книжки, курсы, библиотеки

Раньше



The screenshot shows the LiveLib website interface. At the top, there is a navigation bar with a menu icon, the text 'Что почитать', and categories 'Жанры', 'Авторы', and 'Новости'. The LiveLib logo and a search bar are also present. Below the navigation is a large blue banner with the text 'Причинно следственные связи'. Underneath the banner are several filter buttons: 'Популярные книги', 'Лучшие книги', 'Цитаты', and 'Авторы'. The main content area displays '3 книги' and a grid of book covers. The first book shown is 'Фальшивый купон' by Лев Толстой. The book cover is purple and white. To the right of the cover, the book's title and author are listed, along with statistics: 431 users, 430 pages, 19 reviews, and 13 ratings. Below this, the ISBN (978-5-4241-2427-3), publication year (2011), and language (Russian) are provided. A short description follows, highlighting Tolstoy's talent and the book's focus on human problems. At the bottom left of the book entry, there is a rating of 4.2 and a 'Купить' button. A '+ Добавить' button is located at the bottom of the book entry.

Что почитать Жанры Авторы Новости LiveLib Название / Автор / ISBN

Причинно следственные связи

Популярные книги Лучшие книги Цитаты Авторы

3 книги

Фальшивый купон
Лев Толстой

431 430 19 13

ISBN: 978-5-4241-2427-3
Год издания: 2011
Язык: Русский

Творчество великого русского писателя Льва Николаевича Толстого пользуется любовью читателей всего мира. Его произведения вошли в сокровищницу мировой литературы. Талант писателя многогранен, романы и повести Л.Толстого, детские рассказы и философские произведения, письма и дневники с особой остротой отразили общечеловеческие и сугубо российские проблемы.

4,2 Купить

+ Добавить

Теперь



The image shows the cover of a book titled 'Причинно-следственный анализ' (Causal Inference Analysis) by Mateush Fakur and Artem Gruzdev. The cover has a white background with a central graphic of a stylized eye and a hand holding a gear, surrounded by a dashed circle. The text is in red and black. At the bottom, there are logos for 'G' and 'AMK'.

Матеуш Факур, Артем Груздев

Причинно-следственный анализ

для смелых и честных

G AMK

Тайком готовлю интенсив по Causal Inference, пишите

<https://kolodezev.ru/causalily.html>

Causal Inference — ML Way

- Наш ответ эконометристам
- Закопайте statsmodels ols
- А нет, выкопайте
- А, все уже было в EconML
- А, все равно круто



 Grigory Ch 18:37
#causality #textbooks #causalml-book #изканала

В твиттере/X только и разговоров что о вышедшем в феврале 24-го учебнике Черножукова (<http://causalml-book.org/>). Как и писал ранее, сделал о нём. Дублирую из канала сюда:

Прочитал несколько глав подробно и остальные наискосок и готов поделиться мнением.

Достоинства:

- Core. В книге повествование начинается через adjustment формулу общего вида, что позволяет сравнивать разные стратегии идентификации между собой (например графическую, балансовую, экспериментальную и структурную);
- Нотация. Очень многие темы собраны вместе и имеют, плюс минус, близкую нотацию (например double robust методы сравниваются с conditional ignorability).
- Диапазон тем. Есть большое число достаточно глубоких топигов (например связи регрессии с treatment effects)

Applied Causal Inference Powered by ML and AI

Victor Chernozhukov* Christian Hansen† Nathan Kallus‡
Martin Spindler§ Vasilis Syrgkanis¶

July 28, 2024

Publisher: Online
Version 0.1.1

* MIT

† Chicago Booth

‡ Cornell University

§ Hamburg University

¶ Stanford University

<https://causalml-book.org/>

Интерпретируемость моделей

Учебник по интерпретируемости трансформеров

- Home
- [1.1] Transformer from Scratch
- [1.2] Intro to Mech Interp
- [1.3.1] Toy Models of Superposition...
- [1.3.2] Interpretability with SAEs
- [1.4.1] Indirect Object Identification
- [1.4.2] Function Vectors & Model St...
- [1.5.1] Balanced Bracket Classifier
- [1.5.2] Grokking & Modular Arithme...
- [1.5.3] OthelloGPT
- Monthly Algorithmic Problems

- Home
- Reference page



Chapter 1: Transformer Interpretability

Anthropic про интерпретируемость

https://kolodezev.ru/interpretable_llm.html

Интерпретируемость моделей

- ~~Финальная бумажка про чебуреки~~
- A Comprehensive Guide to Explainable AI: From Classical Models to LLMs
<https://arxiv.org/abs/2412.00800>
- shapiq: Shapley Interactions for Machine Learning
<https://github.com/mmschlk/shapiq>
- Prototypical Concept-based Explanation
<https://maxdreyer.github.io/pcx/>

LLM тащат в обычный прод

/llms

Steering Large Language Models with Pydantic



Jason Liu

2024/01/04

In the last year, there's been a big leap in how we use advanced AI programs, especially in how we communicate with them to get specific tasks done. People are not just making chatbots; they're also using these AIs to sort information, improve their apps, and create synthetic data to train smaller task-specific models.

<https://pydantic.dev/articles/llm-intro>

<https://github.com/instructor-ai/instructor>

Прорыв в антихрупкости* ML

- На изоляте собирают пайплайны с LLM
- С ~~блекджеком~~ инструментами и агентами
- С промптами и моделями по API
- Инженеры и постановщики учатся жить с этим
- Мы наконец-то начали строить гибкие системы
- Мы учимся их мониторить
- И они (иногда) работают

* Термин Антихрупкость украден у [А. Бындю](#) и [Н. Талеба](#)

LLM в Causal Inference

Are LLMs Capable of Data-based Statistical and Causal Reasoning? Benchmarking Advanced Quantitative Reasoning with Data

Xiao Liu¹, Zirui Wu¹, Xueqing Wu², Pan Lu², Kai-Wei Chang² and Yansong Feng^{1*}

¹Wangxuan Institute of Computer Technology, Peking University

²Computer Science Department, University of California, Los Angeles

{xlisa, ziruiwu, fengyansong}@pku.edu.cn

{xueqing.wu, panlu, kwchang}@cs.ucla.edu

Abstract

Quantitative reasoning is a critical skill to analyze data, yet the assessment of such ability remains limited. To address this gap, we introduce the Quantitative Reasoning with Data (QRDATA) benchmark, aiming to evaluate Large Language Models' capability in statistical and causal reasoning with real-world data. The benchmark comprises a carefully constructed dataset of 411 questions accompanied by data sheets from textbooks, online learning materials, and academic papers. To compare models' quantitative reasoning abilities on data

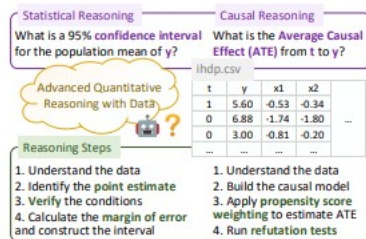


Figure 1: Examples of advanced quantitative reasoning questions and reasoning steps.

Model	Size	Overall Acc.	Multiple Choice / Numerical	Statistical / Causal
Random		23.0	38.1 / 0.0	15.0 / 27.2
<i>Table Question Answering</i>				
TableLlama (Zhang et al., 2023)	7B	9.5	14.9 / 1.2	3.5 / 12.6
<i>Chain of Thought Prompting</i>				
Llama-2-chat (Touvron et al., 2023)	7B	<u>22.1</u>	<u>35.1</u> / <u>2.5</u>	<u>20.4</u> / <u>23.0</u>
AgentLM (Zeng et al., 2023)	7B	<u>24.3</u>	<u>40.0</u> / <u>0.0</u>	<u>19.0</u> / <u>27.1</u>
WizardMath (Luo et al., 2023)	7B	<u>25.1</u>	<u>40.0</u> / <u>2.5</u>	<u>18.3</u> / <u>28.6</u>
CodeLlama-instruct (Roziere et al., 2023)	7B	<u>19.2</u>	<u>31.0</u> / <u>1.2</u>	<u>14.1</u> / <u>21.9</u>
Deepseek-coder-instruct (Guo et al., 2024)	6.7B	18.7	30.2 / 1.2	15.5 / 20.4
Gemini-Pro (Team et al., 2023)	-	31.9	46.8 / 9.2	25.4 / 35.3
GPT-4 (OpenAI, 2023)	-	36.5	57.3 / 4.9	24.6 / 42.8
<i>Program of Thoughts Prompting</i>				
Llama-2-chat	7B	3.7	6.1 / 0.0	7.8 / 1.5
AgentLM	7B	3.2	4.4 / <u>1.2</u>	8.5 / 0.4
WizardMath	7B	16.8	16.5 / <u>17.2</u>	<u>31.7</u> / 8.9
CodeLlama-instruct	7B	17.3	24.2 / <u>6.7</u>	<u>19.7</u> / 16.0
Deepseek-coder-instruct	6.7B	<u>37.0</u>	<u>40.7</u> / <u>31.3</u>	<u>45.8</u> / <u>32.3</u>
Gemini-Pro	-	26.8	27.8 / 25.2	39.4 / 20.3
GPT-4	-	49.1	46.0 / 54.0	72.5 / 36.8
<i>ReAct-style Prompting</i>				
Llama-2-chat	7B	11.4	18.5 / 0.6	4.2 / 15.2
AgentLM	7B	10.7	16.9 / <u>1.2</u>	5.6 / 13.4
WizardMath	7B	18.7	21.0 / 15.3	19.7 / 18.2
CodeLlama-instruct	7B	15.8	22.2 / 6.1	16.2 / 15.6
Deepseek-coder-instruct	6.7B	27.3	29.4 / 23.9	38.7 / 21.2
Gemini-Pro	-	38.9	47.6 / 25.8	41.5 / 37.5
GPT-4	-	57.4	<u>59.2</u> / 54.6	69.0 / <u>51.3</u>
<i>Code Interpreter Assistants</i>				
GPT-3.5 Turbo	-	41.4	42.3 / 39.9	60.6 / 31.2
GPT-4	-	<u>57.9</u>	56.9 / <u>59.5</u>	<u>78.9</u> / 46.8
Human (subset)		76.0	71.9 / 83.4	88.8 / 68.8

<https://arxiv.org/abs/2402.17644v2>

LLM в СІ лучше среднего джуна

Data Description

The CSV file `ihdp.csv` contains data obtained from the Infant Health and Development Program (IHDP). The study is designed to evaluate the **effect of home visit from specialist doctors on the cognitive test scores of premature infants**. The confounders `x` (`x1-x25`) correspond to collected measurements of the children and their mothers ...

Question

What is the **Average Treatment Effect (ATE)** of the dataset? Please round the final answer to the nearest hundredth.

Correct Reasoning Steps:

1. Check rows of the dataset to understand its structure

```
import pandas as pd
data = pd.read_csv('ihdp.csv')
print(data.head())
```

t	y	x1	x2	...
1	5.60	-0.53	-0.34	...
0	6.88	-1.74	-1.80	...
0	3.00	-0.81	-0.20	...
...

Sandbox Execution Results:

3. Recall related method and apply to this scenario

ATE can be estimated using propensity score weighting:

...

```
ihdp_estimate = ihdp_model.estimate_effect(
    ihdp_identified_estimand,
    method_name="backdoor.propensity_score_weighting"
)
print('Estimated effect:', ihdp_estimate.value)
```

Estimated effect: 4.02

2. Build a causal model based on the data description

```
from dowhy import CausalModel
common_causes = ['x1', ..., 'x25']
ihdp_model = CausalModel(
    data=data, treatment='t', outcome='y',
    common_causes=common_causes
)
```

4. Run refutation test to validate the estimate

The estimate should not change if we add an independent random variable as a common cause to the dataset.

```
ihdp_refute_random_common_cause = ihdp_model.refute_estimate(
    ihdp_identified_estimand, ihdp_estimate,
    method_name="random_common_cause"
)
print('New effect:', ihdp_refute_random_common_cause.new_effect)
```

New effect: 4.02

Final Answer: 4.02

Figure 2: An example question of QRDATA and correct reasoning steps.

Все как у людей

Easier Said than Done. The unsatisfactory performance on causal questions appears to differ from the findings of some previous papers (Willig et al., 2022; Kıcıman et al., 2023), which show that LLMs are capable of causal reasoning in tasks like causal discovery *without data*. To figure out the reason for the inconsistency, we conduct an experiment to evaluate if the provided data interfered with the models. Specifically, we evaluate the best three models on multiple-choice causal questions of QR-DATA *without giving them the data*⁸, and compare their performance with the original setting. We use the best reasoning method for each model: ReAct-style prompting for Gemini-Pro, and code interpreter assistants for GPT-3.5 Turbo and GPT-4.

As shown in Table 4, two out of three models perform better without data access, revealing that it is easier for these LLMs to speculate causal re-

Question	Which cause-and-effect relationship is more likely? Please answer with A, B, or C. A. R L5 radiculopathy causes L L1 radiculopathy B. L L1 radiculopathy causes R L5 radiculopathy C. No causal relationship exists
Gold Answer	C
GPT-4 Prediction with Data Access	
Step 1:	# Display the first few records to understand the data structure
Code Interpreter	<code>data.head()</code>
Step 2:	I will examine the data for the following patterns: - If patients with “R L5 radiculopathy” also frequently have “L L1 radiculopathy” and not the other way around, it may suggest that A is more likely. - Conversely, if patients with “L L1 radiculopathy” also frequently have “R L5 radiculopathy” and not the other way around, it may suggest that B is more likely. - If there is no clear pattern or both conditions occur independently, it may suggest that C is more likely.
Step 3:	# Probability of L L1 given R L5
Code Interpreter	<code>P_L_L1_given_R_L5 = co_occurrence_count / R_L5_count if R_L5_count else 0</code> # Probability of R L5 given L L1 <code>P_R_L5_given_L_L1 = co_occurrence_count / L_L1_count if L_L1_count else 0</code> <code>P_L_L1_given_R_L5, P_R_L5_given_L_L1</code>
Step 4:	Based on these conditional probabilities, it is more likely that “L L1 radiculopathy” causes “R L5 radiculopathy” over the other proposed causal direction. Final answer: B.
GPT-4 Prediction without Data Access	
Prediction	Given both ‘R L5’ and ‘L L1’ radiculopathies are on different sides of the body and different spinal levels, it’s not common for one to cause the other according to standard medical understanding. It is more likely that they occur independently due to separate causes. It’s likely that if both conditions are present in a patient, they arose independently or due to a systemic issue affecting the spine broadly, rather than one causing the other. Final answer: C.

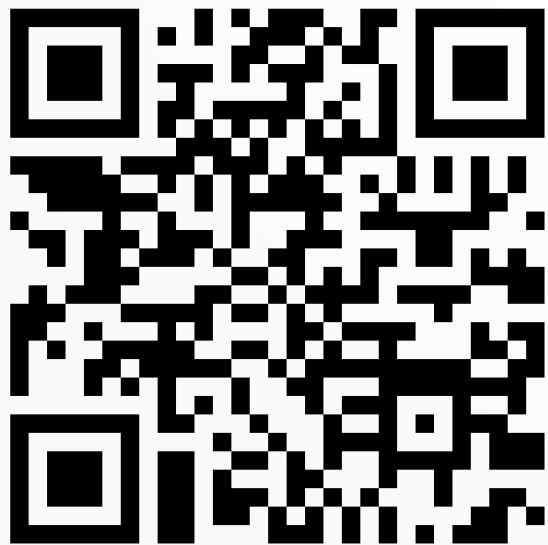
Ну ЭТО ОНИ ТОЧНО НЕ СМОГУТ

Potential Risks. Our evaluation reveals that current LLMs are imperfect in conducting quantitative reasoning with data. We suggest readers not trust solutions provided by LLMs.

- Ждем LLM с Causal Inference до майских праздников
- «Будущее Causal Inference за онтологиями» (с)
- Внутри LLM есть онтологии, на любой вкус, ч. и т.д.

Вопросы

Слайды тут



dkolodezev



dmitry_kolodezev



Promsoft



Reliable ML

<https://kolodezev.ru/download/dataelka2025.pdf>