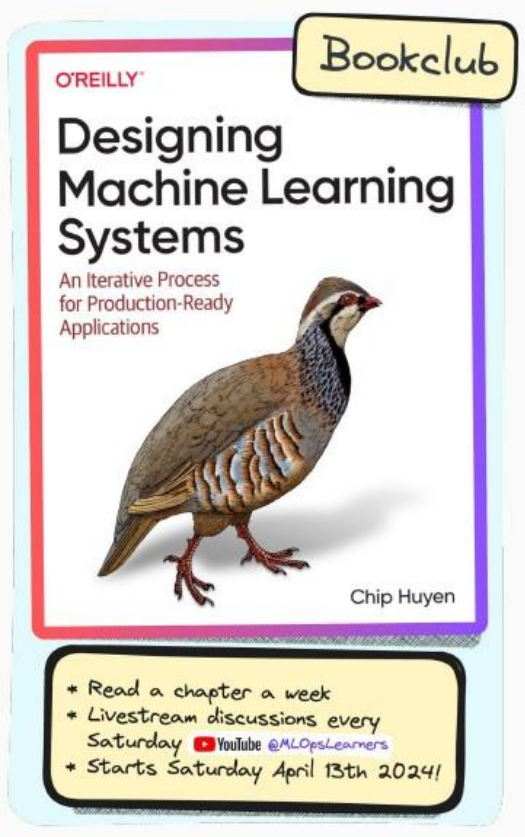


Chapter 6: Model Development and Offline Evaluation

Group 10 (Manifold Minds)

Saturday, June 1, 2024

YouTube

https://www.youtube.com/live/X8OOgKFhg4g

DMLS Book Club

Chapter 8: Data Distribution Shifts and Monitoring

Team Members

Mardiyyah Oduwole
Independent
Researcher

Besher Alkurdi
NLP Engineer
Huawei

Mau Nguyen
Research Assistant
HAI LAB - NAIST

Outline

- Team Introduction
- Chapter Takeaways
  - Causes of ML System Failures
  - Data Distribution Shifts
  - Monitoring and Observability
- Case studies
- Future discussion
  - Current situation with GenAI?

Chapter Takeaways

Causes of ML System Failures

1. Software system failures
    - Dependency failures: breaks in supporting software, package, lib,...
    - Deployment failures: deploy errors (wrong version, lack of permissions,...)
    - Hardware failures: CPU overheat, broken Memory,...
    - Downtime failures: Server shutdown, component service suspended,...

Chapter Takeaways

Causes of ML System Failures

2. ML-specific failures

- Mismatched data distribution: Prod data and train data are not in same distribution
- Edge cases: low failure rate but come with costly consequence
- Degenerate feedback loop: Systems' outputs give feedback to the inputs, making a loop of biased influence

Data Distribution Shifts

- A common cause of ML system failures
- Occurs when the data distribution changes over time
- Leads to less accurate predictions

Types of Data Distribution Shifts

- Covariate Shift: P(X) changes, P(Y|X) remains the same
    - Income level of users
    - Caused by biases in data selection, artificial alteration of training data, or active learning
- Label Shift: P(Y) changes, P(X|Y) remains the same
    - Breast cancer diagnosis, output distribution changes
- Concept Drift: P(Y|X) changes, P(X) remains the same
    - Housing prices, output distribution changes over time

Addressing Data Distribution Shifts

- Train models using massive datasets
    - Hope: anything will be in my data anyways
- Adapt trained models to target distribution without new labels
    - Very mathy :)
- Retrain models using labeled data from target distribution

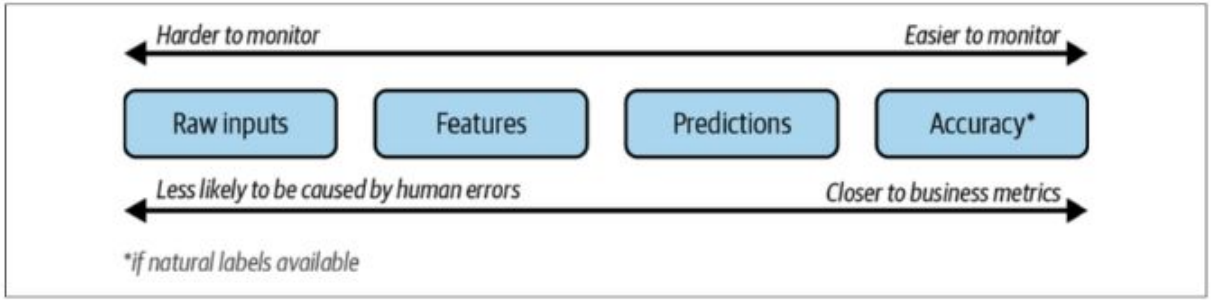
Some Advices for Data Distribution Shift

- Design systems to be robust to shifts
- Choose features carefully considering performance and stability
- Consider using separate models for different markets or distributions

Chapter Takeaways

Monitoring and Observability

- Monitoring refers to the act of tracking, measuring, and logging different metrics that can help us determine when something goes wrong.
  - Monitoring is all about metrics.
  - A model's accuracy-related metrics, predictions, features, and raw inputs.



Chapter Takeaways

Monitoring and Observability

- Observability: setting up visibility in our system to investigate issues.
- “Better visibility into understanding the complex behavior of software using [outputs] collected from the system at run time.”
- It also encompasses interpretability (helps us understand how an ML model works).

Case Studies

Apple Card is accused of gender bias. Here's how that can happen

New York (CNN Business) — Some Apple Card customers say the credit card's issuer, Goldman Sachs, is giving women far lower credit limits, even if they share assets and accounts with their spouse. But it's impossible to know if the Apple Card – or any other credit card – discriminates against women, because creditworthiness algorithms are notoriously opaque.

"It's such a mystery we are seeing," said Sara Rathner, travel and credit cards expert at NerdWallet. "Because we don't know exactly what those algorithms are looking for, it can be hard to say if there might be some bias built into them."

The New York Department of Financial Services is looking into the allegations of gender discrimination against users of the Apple Card. The allegations blew up on Twitter Saturday after tech entrepreneur David Heinmeier Hansson wrote that Apple Card offered him 20 times the credit limit as his wife, although they have shared assets and she has a higher credit score.

The ML algorithm may have used historical data that possibly contained gender biases, resulting in discriminatory credit assessments.

Ref: https://slate.com/technology/2022/01/ibm-watson-health-failure-artificial-intelligence.html

Case Studies

IBM Watson for Oncology

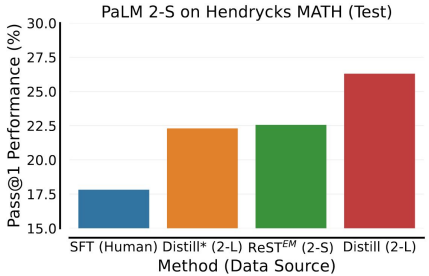
IBM's Watson for Oncology was intended to assist doctors in diagnosing and treating cancer, but it provided incorrect and unsafe treatment recommendations.

A doctor involved said that there wasn't enough data for the program to make good recommendations, and that Watson had trouble with the complexity of patient files.

"If you think about it, knowing what we know now or what we've learned through this, the notion that you're going to take an artificial intelligence tool, expose it to data on patients who were cared for on the upper east side of Manhattan, and then use that information and the insights derived from it to treat patients in China, is ridiculous. You need to have representative data. The data from New York is just not going to generalize to different kinds of patients all the way across the world"

Relation with current GenAI

- Adapting a target distribution is easy with GenAI!
  - With a small GPU at home large models trained on massive datasets can be adapted to a desired domain. See [“You can now train a 70b language model at home”](#).
  - RAG! (See [MLOPs Learners’ session on RAG](#))
- Using large generative models to create new, realistic synthetic data to help smaller (or large!) models perform better with unexpected data.



Source: Singh et al. (2024)