

04. What is Data Drift? Concept Drift?

Tags	Concept Drift Data Drift MLOps
date	@August 7, 2021
url	

Motivation

Googling

[What is data drift and how does MLOps contribute to handling it](#)

[What is Data drift](#)

[ML Ops as an enabler to detect data drift and easily retrain ML models](#)

Motivation

- what is data drift
- what is concept drift
- why is it important to understand data drift and concept drift in MLOps
- How can I automate the ML pipeline?

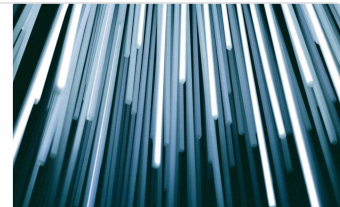
Googling

What is data drift and how does MLOps contribute to handling it

What is data drift and how does ML Ops contribute to handling it? - VIQTOR DAVIS

Machine Learning (ML) models are unique and optimized for specific problems, yet their model performance can fluctuate over time. ML models used in organizations' operational processes continuously require monitoring to revalidate their performance on an ongoing

[https://www.viktordavis.com/en-us/media/data-drift#:~:text=Data%20drift%20is%20a%20c](https://www.viktordavis.com/en-us/media/data-drift#:~:text=Data%20drift%20is%20a%20change,current%20real%20time%20production%20data)



Machine Learning (ML) models are unique and optimized for specific problems, yet their model performance can fluctuate over time. ML models used in organizations' operational processes continuously require monitoring to revalidate their performance on an ongoing basis. One needs to monitor whether a ML model's performance remains acceptable and whether patterns of new incoming data are still accurately being captured by the model. One of the top reasons why performance of ML models decreases over time is caused by data drift. In this article we will dive into data drift and how ML Ops contributes to handling it. This article is part of a series of blog posts on ML Ops. See a previous post for a general introduction to ML Ops.

What is Data drift

Data drift is a change in the distribution of data over time. In case of operational machine learning models, data drift is the change in the distribution of a baseline data set on which the model was

trained and the current real-time production data. Real-time production data distributions can drift from the baseline data set over time due to changes in the real world or changes in measures. We can broadly classify 3 types of drift: concept drift, data drift, and upstream data changes.

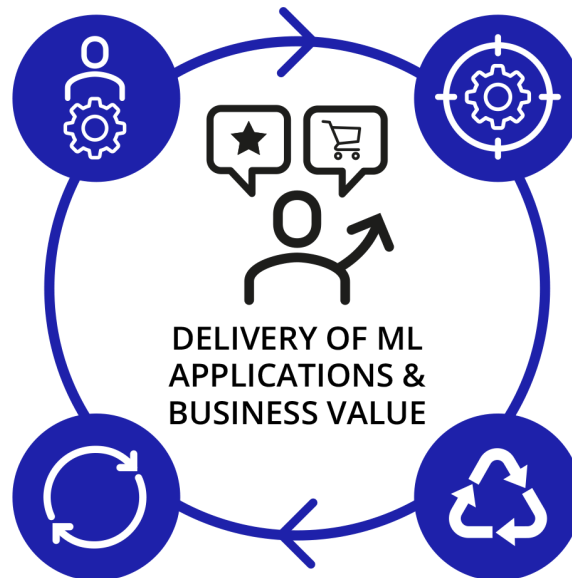
- Concept drift happens when the statistical properties of the target variable itself change. The meaning of what you are trying to predict changes and therefore the model will not work well for this updated definition. For example, the definition of what is considered a fraudulent transaction could change over time.
- Data drift happens when the statistical properties of the underlying variables that predict an outcome change. A classic example is the natural drift in data due to seasonality.
- Upstream Data Change happens when there is a change in the data pipeline upstream which has an impact on the model performance. For example, camera's being replaced and as a consequence the units of measurement change.

ML Ops as an enabler to detect data drift and easily retrain ML models

Given that the performance of a ML model can change over time once it is deployed into production, the best course of action is to monitor for changes in performance and retrain models when needed. By implementing the ML Ops principles: automation, reusability, reproducibility and manageability, you can protect your models from undesired performance degradation. ML Ops (as discussed in our ML-Ops [solution](#)) draws on DevOps principles and practices. It consists of best practices for the delivery of ML models and enables you to address the *mentioned types of* data drift.

MANAGEABILITY
Ability to enforce model governance and track changes to models and code throughout the development lifecycle. This ensures consistent value delivery of ML solutions.

AUTOMATION
Automation of the ML pipeline components is required to assure a rapid and repeatable pipeline and ensure constant, consistent, and efficient delivery of business value.



REPRODUCIBILITY
ML pipelines, together with the data sources, code, models, libraries, and SDK's, are versioned and maintained such that they can be reproduced.

REUSABILITY
To fit with principles of continuous delivery, pipelines are built such that models & code are packed consistently into training and target environments, and the same configuration can be re-used producing the same results.

Using ML Ops pipelines and monitoring tools, both concept drift (distribution of predictions) and data drift (both data- and feature contribution distributions) can be monitored. When drift in a model is detected, the next step is identifying which features are causing the drift. It can be the case that several features have drifted but not have caused a meaningful drift in the model because these features have a low importance on the model. Identifying the feature that causes the drift and are of great importance to the model, are crucial to the performance of the model and should therefore receive better attention when retraining your model.

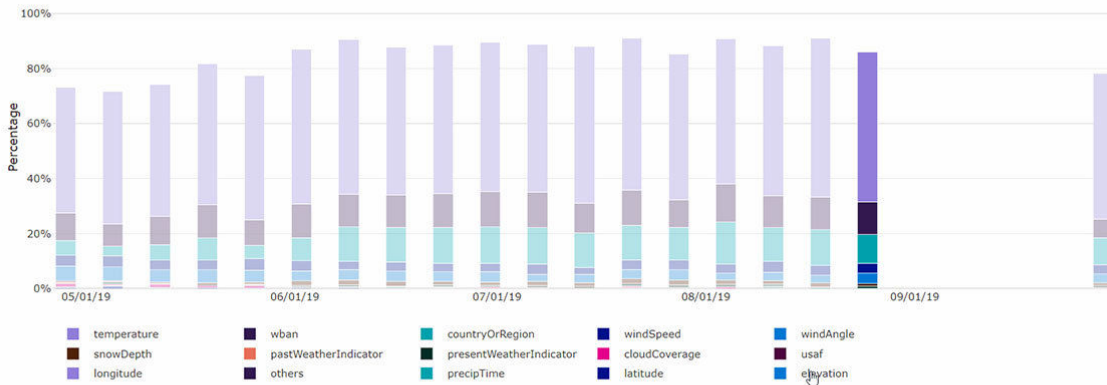
- **concept drift** : distribution of predictions
- **data drift** : data-and feature contribution distribution

⇒ 데이터가 바뀌었네? ⇒ 데이터 드리프트

컨셉이 바뀌었네? ⇒ 컨셉 드리프트

weather-monitor-demo

Settings Analyze existing data Refresh



Feature level drift details as of 2019-08-25

View metrics for a specific date by selecting the date in the chart above. Click on the feature name to view feature level trends
Loading

When ML model development, deployment and monitoring have been established with the ML Ops principles in mind, retraining your model becomes “a piece of cake”. Configuring the newly perceived data (combined with your baseline dataset) as your new training data, automatically triggers the training pipeline that trains, evaluates and validates the new models. Trained models with an increased performance, are automatically deployed via the continuous deployment pipeline. Ensuring the best performing model is in production!


ML Ops 원칙을 염두에 두고 ML 모델 개발, 배포 및 모니터링이 설정되면 모델 재교육은 "piece of cake"(식은죽 먹기)가 됩니다. 새로 인식된 데이터(기준 데이터 세트와 결합)를 새 훈련 데이터로 구성하면 새 모델을 훈련, 평가 및 검증하는 훈련 파이프라인이 자동으로 트리거됩니다. 향상된 성능으로 훈련된 모델은 지속적인 배포 파이프라인을 통해 자동으로 배포됩니다. 최고의 성능을 발휘하는 모델이 생산 중인지 확인하십시오!

• A-ha!

- 새로 인식된 데이터(기준 데이터 세트와 결합) ⇒ 새 훈련 데이터 구성
- 새 모델을 훈련, 평가 및 검증하는 훈련 파이프라인 트리거
- 향상된 성능으로 훈련된 모델은 지속적인 배포 파이프라인을 통해 자동 배포

Detect data drift on datasets (preview) - Azure Machine Learning

Learn how to monitor data drift and set alerts when drift is high. With Azure Machine Learning dataset monitors (preview), you can: Analyze drift in your data to understand how it changes over time. Monitor model data for differences between

 <https://docs.microsoft.com/en-us/azure/machine-learning/how-to-monitor-datasets?tabs=python>

