Einführung in MLOps

11 Processing- und Prediction Modes, Features

Tobias Mérinat
teaching2025@fsck.ch

Lucerne University of
Applied Sciences and Arts
HOCHSCHULE
LUZERN

Department of Information Technology
Lucerne University of Applied Sciences and Arts
6343 Rotkreuz, Switzerland

14. und 15. Februar 2025

Processing Modes

Drei grundlegende Arten, wie Daten verarbeitet werden

- Request-Response
- Batch Processing
- Stream Processing

Request-Response

- stateless
- synchron
- einfach skalierbar

Batch Processing

- stateful
- bounded
- parallelisierbar

Stream Processing

- stateful
- unbounded
- komplex

Zusammenfassung Processing Modes

- Batching und Streaming gemacht für stateful, high Volume Processing
- Request-Response
  - eher für User-Facing Apps verwendet
  - für Data Processing meist nicht geeignet, da im Normalfall stateless
  - kann aber auch die Resultate aus Batch-Processing nachladen und verwenden

Features und Datenkategorien

- **Features**: Die Inputs eines ML Modells
- **Feature-Engineering**: Berechnung der Features aus Rohdaten
- **Unstrukturierte Daten**: (eher weniger Feature Engineering)
- **Strukturierte Daten**: (eher mehr Feature Engineering)

Feature-Arten 1/2

- Batch-Features
- Streaming-Features
  - Real-Time (RT) Features
  - Near-Real-Time (NRT) Features

Feature-Arten 2/2

Feature-Typ	Berechnung	Compute Engine (Bsp.)	Latenz	Vorteile	Nachteile
Batch	Vorberechnet in Batch-Prozess	Spark	Stunden bis Tage	einfach, parallel-isierbar	keine aktuellen Features
RT	Zum Zeitpunkt der Vorhersage	Python, SQL	<1s	einfach, aktuelle Features	skaliert nicht
NRT	Vorberechnet in Streaming-Prozess	Flink, Quix	Sek. bis Min.	Aktuelle Features, skalierbar	schwierig umsetzbar

Lucerne University of Applied Sciences and Arts
HOCHSCHULE LUZERN

Batch-Features

- Setzt nur eine (häufig bereits bestehende) Batch-Processing Infrastruktur voraus
- Vorteile
  - Wenig komplex
  - Parallelisierbar
- Nachteile
  - Begrenzte Aktualität
  - u.U unnötige Berechnungen

Real-Time (RT) Features

- Direkt im Rahmen jeder Anfrage berechnet
- Vorteil
  - Top-Aktuell
  - Einfach umsetzbar
- Nachteil
  - Berechnungszeit erhöht direkt Latenz, skaliert deshalb schlecht

Near-Real-Time (NRT) Features

- Werden (im Gegensatz zu RT) asynchron berechnet
- Latenz wird nur durch Lookup-Zeit erhöht
- Unterschied zu Batch:
  - Tendentiell häufiger berechnet
  - Mittels Stream-Processing berechnet

Prediction Modes

Wir unterscheiden verschiedene Stufen

1. Offline-Prediction (Batch-Prediction)
2. Online-Prediction
    1. ausschliesslich mit Batch-Features
    2. mit Streaming- und Batch Features
3. Continual Learning

Lucerne University of Applied Sciences and Arts
HOCHSCHULE LUZERN

Offline-Prediction (Batch-Prediction)

- Berechnungen erfolgen periodisch und im Voraus (vor der Verwendung)
- Typische Anwendungsfälle:
  - Lead Scoring
  - Demand Forecasting
  - Inventory Management
- Voraussetzungen sind eine Batch-Infrastruktur
- Model Registry wäre eine mögliche Ausbaustufe

Online-Prediction mit Batch-Features

- Online Prediction: Wenn die Vorhersage erst gemacht wird, wenn sie benötigt wird
- Nur vorausberechnete Features werden verwendet
- Wenn Batch-Features (zeitlich) *akkurat genug* sind
- Neben der Batch-Infrastruktur wird schneller Key Value Store benötigt

Online-Prediction mit Online Features

- RT, NRT und Batch Features kombiniert
- Voraussetzung ist eine Stream Processing Infrastruktur

Retraining

- Fast immer ändern sich Umwelt und somit Input-Daten für Modell über die Zeit
- Dann müssen Modelle periodisch neu trainiert werden
- Dazu werden nicht nur Inputdaten (X) sondern auch das Target (y) benötigt
- Manchmal fällt das Target automatisch im Prozess an (sog. Natural Labels)
- Oft muss es aber manuell erarbeitet werden

Retraining Stufen

- Manuelles stateless Retraining
- Automatisiertes stateless Retraining
- Automatisiertes stateful Training
- Continual Learning

Manuelles stateless Retraining

- Retraining, wenn sich der Bedarf zeigt
- Labelling manuell
- Zusammenstellung der Trainingsdaten manuell
- Mit alten und neuen Daten oder nur mit neuen
- Stateless: Modell wird von Grund auf neu trainiert

Automatisiertes stateless Retraining

- Möglich mit Natural Labels
- Weniger fehleranfällig
- Komplex, wenn Inputdaten und Labels zeitlich stark versetzt anfallen

Automatisiertes stateful Training

- Auch *inkrementelles Lernen*
- Modelle werden mit aktuellen Daten *weitertrainiert*
- Setzt *Model Lineage* voraus

Continual Learning

- Vollkommene Automatisierung
- Nimmt die Performance zu stark ab, wird automatisch neu trainiert
- Preisfragen:
  - Wann sollte neu trainiert werden
  - Ist das neue Modell tatsächlich besser