Machine Learning

| Batch – Online Learning

Prof. Dr. Mostafa Elhosseini

Professor of Smart Systems Engineering

https://youtube.com/drmelhosseini

Agenda

- whether or not the system can learn incrementally تدريجيا from a stream of incoming data
  - Batch learning
  - Online learning
- Concept Drift
- Data Drift

Introduction

- Classify machine learning systems whether or not the system can learn incrementally from a stream of incoming data
  - Batch learning
  - Online learning

Batch Learning

- In **batch learning**, the machine learning model is trained using the entire training dataset at once. This means that the learning process happens in one step, using all available data.
  - **Data Size**: Usually applied when the dataset is of manageable size and can be fit into memory.
  - **Computation**: Training might be computationally expensive as the model processes the entire dataset in one go.
  - **Static Model**: Once trained, the model doesn't learn from new data unless it's retrained with a **new batch including the old and new data**.
  - **Use Cases**: Effective for situations where the data doesn't change frequently and the model doesn’t need to adapt to new data patterns over time.

Batch Learning

- First the system is trained, and then it is launched into production and runs without learning anymore; it just applies what it has learned. This is called **offline learning**.
- Model's performance tends to decay slowly over time, simply because the world continues to evolve while the model remains unchanged. This phenomenon is often called **model rot or data drift**.
- Static Model: Once trained, the model doesn't learn from new data unless it's retrained with a new batch including the old and new data – How often you do that
  - Classifies pictures of cats and dogs
  - Financial market

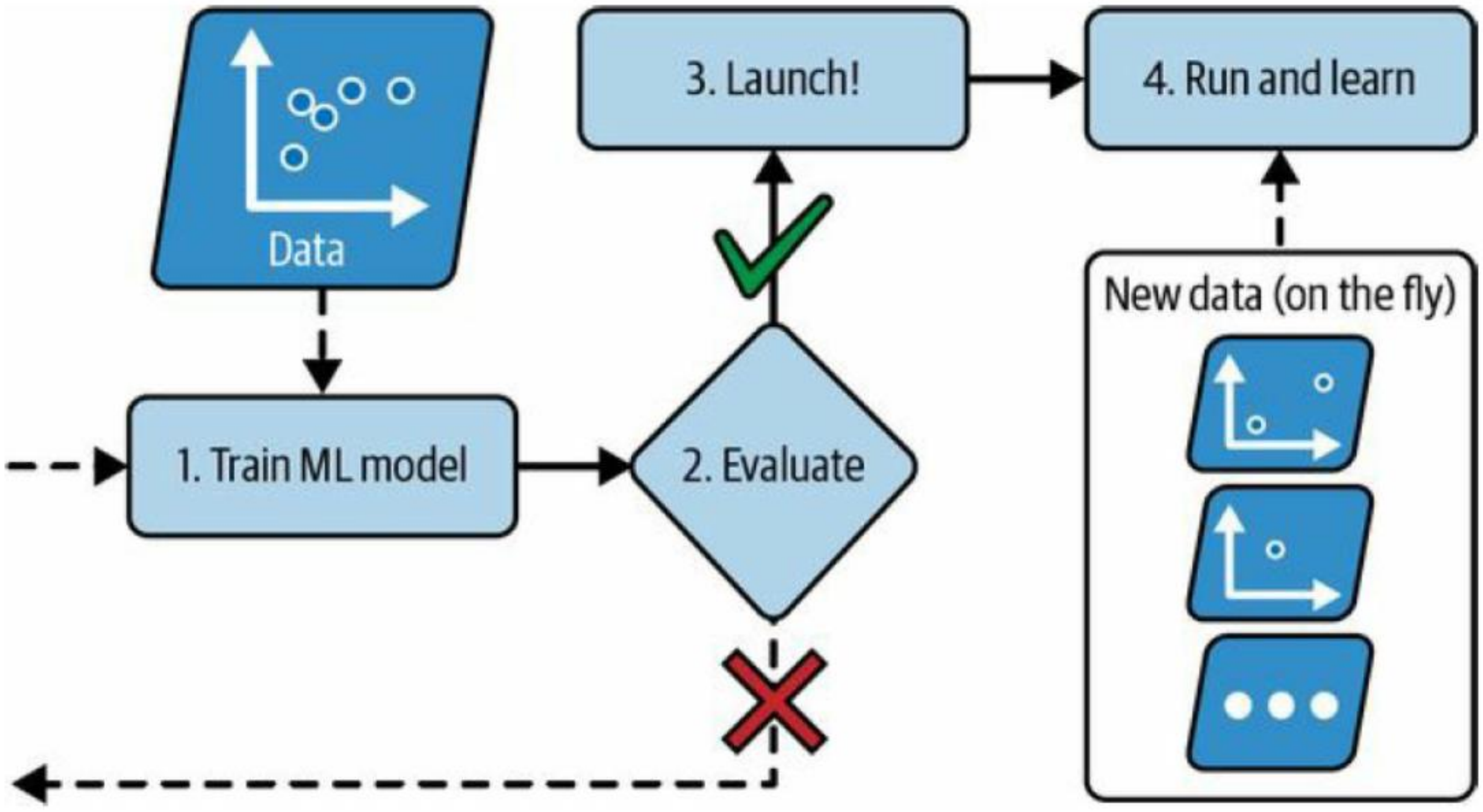
Batch Learning

- Medical Diagnosis System: Imagine a medical diagnosis system developed to identify diseases from medical images, like MRI or CT scans. The system is trained on a large dataset of labeled images, where each image is annotated with the disease it represents. This training is done in one comprehensive process, using all the available data. Once trained, the model is deployed in hospitals. It doesn't need to update frequently since the fundamental principles of diagnosing these diseases from images don't change often. In this scenario, batch learning is appropriate because the model can be trained on a fixed dataset and doesn't require continuous updates.

Online Learning

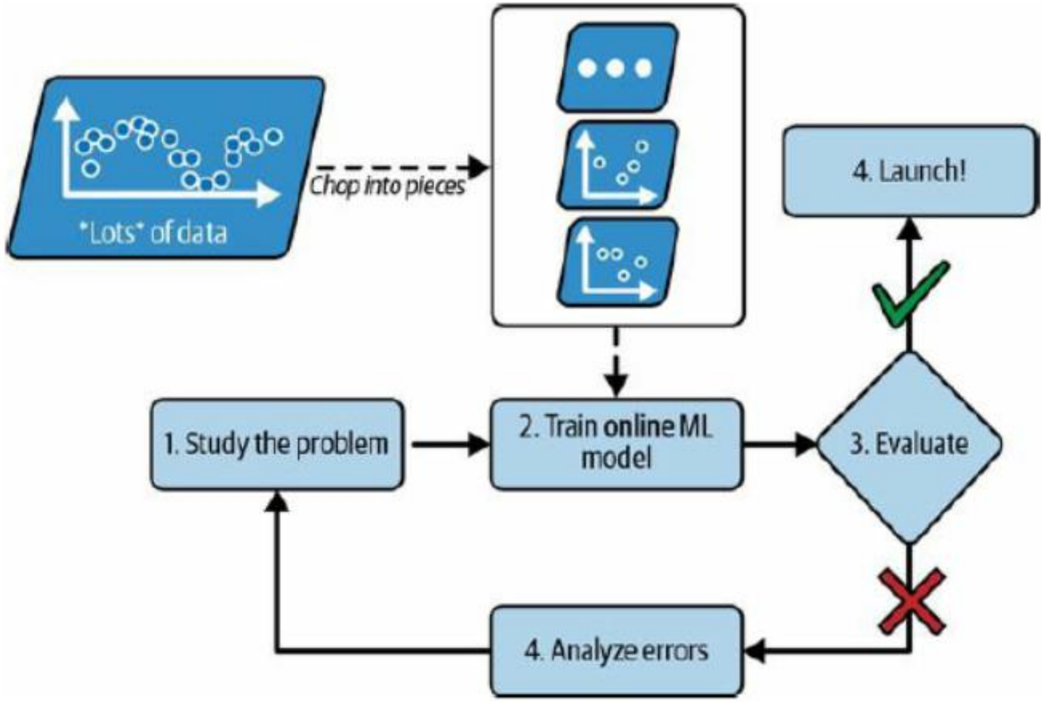
- In online learning, the machine learning model is trained incrementally by continuously feeding it data samples, often one at a time or in small groups (mini-batches).
  - **Data Stream**: Ideal for situations where data arrives in a continuous stream, or when the data is too large to be processed all at once.
  - Each learning step is fast and cheap, so the system can learn about new data on the fly, as it arrives
  - **Adaptability**: The model can adapt to new patterns in data over time, making it suitable for dynamic environments.
  - **Resource Efficiency**: More memory and computationally efficient as it doesn't require the entire dataset to be loaded into memory.
  - **Risk of Drift**: The model **might suffer from concept drift if the data patterns change over time**, requiring mechanisms to detect and adapt to such changes.

Online Learning



Online Learning

- Online learning algorithms can be used to train models on huge datasets that cannot fit in one machine's main memory (this is called **out-of-core** learning).
- The algorithm loads part of the data, runs a training step on that data, and repeats the process until it has run on all of the data



Online Learning

- Both **Concept Drift and Data Drift** are critical considerations in online learning systems, as these systems are continuously updated with new data. If bad or irrelevant data is fed into the system, it can quickly deteriorate the model's performance.

Online Learning

- To reduce this risk, you need to monitor your system closely and promptly switch learning off (and possibly revert to a previously working state) if you detect a drop in performance. You may also want to monitor the input data and react to abnormal data; for example, using an anomaly detection algorithm
- Regular monitoring, validation, and potentially retraining with more representative data are essential to maintain the performance of such systems

Online Learning

- **Concept Drift**: This refers to the scenario where the underlying relationships between input and output variables change. For instance, in a stock price prediction system, the factors influencing stock prices might evolve over time due to changes in the market, economic policies, or investor behavior.
- **Data Drift**: This is when the nature of the data itself changes, but the underlying concept remains the same. For example, if a model is trained to recognize objects in images and the quality of incoming images changes (like different lighting conditions or angles), the model's performance might degrade even though the task (object recognition) remains the same..

Online Learning

- **Concept Drift**: This refers to the scenario where the underlying relationships between input and output variables change. For instance, in a stock price prediction system, the factors influencing stock prices might evolve over time due to changes in the market, economic policies, or investor behavior.
- **Data Drift**: This is when the nature of the data itself changes, but the underlying concept remains the same. For example, if a model is trained to recognize objects in images and the quality of incoming images changes (like different lighting conditions or angles), the model's performance might degrade even though the task (object recognition) remains the same..

Concept Drift - Example

Email spam detection

- Initial Scenario: The model is initially trained on a dataset containing examples of spam and non-spam emails. At this time, spam emails might have certain recognizable patterns such as specific keywords (like "lottery," "free," "winner"), poor grammar, or formatting styles.
- Change Over Time: As spam filters become more effective, spammers change their tactics. They start using more sophisticated language, avoiding known spammy keywords, and employing techniques that make their emails look more like legitimate correspondence.

Concept Drift - Example

Email spam detection

- Concept Drift: The original patterns that the model learned to recognize as spam no longer apply. The nature of spam has evolved, but the model was trained on outdated examples. As a result, the model's performance in accurately detecting new spam emails starts to decline.
- Impact: The spam detection system begins to fail more frequently, either by flagging legitimate emails as spam (false positives) or by allowing spam emails to pass through (false negatives).

Concept Drift - Example

Email spam detection

- In this case, the concept of what constitutes a "spam email" has drifted over time. To maintain accuracy, the spam detection model would need to be periodically updated with new data that reflects the latest spamming tactics, a process often termed as retraining or model updating.

Data Drift - Example

Traffic monitoring and congestion prediction

- Initial Scenario: The model is initially trained on a dataset consisting of traffic images and videos collected under various conditions, such as different times of the day, weather conditions, and levels of traffic. The model learns to identify patterns of traffic flow and can predict congestion based on these inputs.
- Change in Data: Over time, the city upgrades its traffic cameras to higher resolution models, and some cameras are repositioned for better coverage. These changes result in the model receiving images and videos with different visual characteristics than those it was trained on. The new images might have higher clarity, different angles, and varying field-of-view compared to the original training data.

Data Drift - Example

Traffic monitoring and congestion prediction

- Data Drift: Although the task of monitoring traffic and predicting congestion remains the same, the data's nature (i.e., the images and videos from the traffic cameras) has changed. This change in the input data's distribution and characteristics can lead to a decrease in the model's performance, as it was trained on data with different visual properties.
- Impact: The model might start to misinterpret traffic conditions, either by overestimating or underestimating congestion levels, because it is less effective at processing the new types of images. This could lead to inaccurate traffic predictions and inefficient traffic management.

Data Drift - Example

Traffic monitoring and congestion prediction

- In this example of data drift, the underlying concept (traffic and congestion) hasn't changed, but the nature of the data fed into the model has. To address this, the model would need to be retrained or adapted to the new characteristics of the input data to maintain its accuracy and effectiveness..