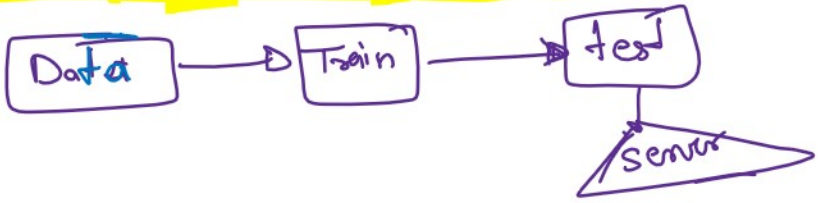
Batch vs online Machine Learning

02 November 2024 19:26

Batch Learning (Offline Learning)

Definition

Batch learning, also known as **Offline Learning**, is a conventional machine learning approach where a model is trained on the entire dataset in one go, without further updates or re-training.



Process

1. **Data Collection**: Gather the entire dataset at once.
2. **Training Phase**: Use the full dataset to train the machine learning model, allowing it to learn patterns from all the data at once.
3. **Deployment**: Once trained, the model is deployed to a server for production.
4. **Testing**: After deployment, the model runs predictions based on the patterns it learned during training.

→ Let Say we are gathering data after 24hrs than retrain again with combining previous data + new data to keep updated

Advantages

- **Consistency**: The model is trained on a fixed, comprehensive dataset, ensuring consistent learning.
- **Efficient for Stable Data**: Works well when data does not frequently change, making it ideal for stable environments.

Disadvantages

- **Not Suitable for Dynamic Data**: If data changes frequently, the model needs re-training to stay relevant.
- **High Computational Cost**: Training on large datasets requires significant computational resources and time.
- **Not Adaptable in Real-Time**: Batch learning cannot easily handle real-time updates or changes in the data patterns.

Examples

- **Recommendation Engines**: Batch training is often used for movie recommendations or e-commerce suggestions, where the model learns from historical user data.
- **Medical Diagnosis**: A model trained offline on medical records for diagnosing specific diseases.

→ But we need to retrain on new data on entire dataset.

Online Learning

Definition

Online Learning, also known as **Incremental Learning**, is a machine learning approach where a model updates itself continuously with new data points as they become available, without needing to re-train on the entire dataset.

Process

1. **Initial Training**: The model is initially trained on a small dataset.
2. **Continuous Updates**: As new data arrives, the model incorporates it, adjusting itself without a full retraining.
3. **Real-Time Adaptation**: The model adapts in real-time to any new trends, patterns, or shifts in data.

Advantages

- **Real-Time Adaptability**: Ideal for applications where data is dynamic and constantly changing.
- **Less Computational Cost**: Instead of re-training the whole model, only small, incremental updates are made.
- **Effective for Large, Streaming Data**: Works well with high-volume data environments, such as social media or stock markets.

Disadvantages

- **Risk of Drift**: The model may overfit or "drift" from accurate predictions if it overly adapts to recent but temporary trends.
- **More Complexity in Deployment**: Requires a stable and reliable infrastructure to ensure continuous data flow and model updates.

Examples

- **Spam Email Filtering**: Continuously learns and adapts to new spam patterns as they evolve.
- **News Feed Recommendation**: Adapts recommendations based on recent user behavior, ensuring relevance.

Comparison Between Batch and Online Learning

Feature	Batch Learning	Online Learning
Data Requirement	Full dataset available at once	Continuous data flow
Training Frequency	Trained once, then deployed	Continuously updated
Computational Cost	High for large datasets	Lower, as updates are incremental
Adaptability	Low; doesn't adapt to new data easily	High; adapts to new data in real-time
Ideal Use Case	Static data, stable environments	Dynamic data, real-time applications

Analogies for Better Understanding

Batch Learning Analogy

Think of **Batch Learning** as baking a large batch of cookies. You gather all the ingredients, mix them once, and bake them in one go. Once baked, the cookies are done and cannot be modified. If you want to make changes (e.g., add new ingredients), you'll need to start over with a new batch.

Online Learning Analogy

Imagine a **Personal Fitness Trainer** who tracks your progress daily and adjusts the workout routine based on your daily performance. The trainer doesn't rely on a single, fixed plan but continuously adapts to help you reach your goals in real time, similar to how Online Learning models constantly adjust to new data.

Revision Notes

- **Batch Learning** (Offline):
  - Trained on full dataset once.
  - Best for static data environments.
  - High computation, not adaptive in real-time.
  - Example: Movie recommendation model.
- **Online Learning**:
  - Continuously updates with new data.
  - Ideal for dynamic, changing data.
  - Adaptive but can suffer from drift.
  - Example: News feed algorithms.