DRIFT DETECTION

- "Drift" is a term used in machine learning to describe how the performance of a machine learning model in production slowly gets worse over time.
- This can happen for a number of reasons, such as changes in the distribution of the input data over time or the relationship between the input (x) and the desired target (y) changing.
- Drift can be a big problem when we use machine learning in the real world, where data is often dynamic and always changing

What Is Drift?

- Machine learning models are trained with historical data, but once they are used in the real world, they may become outdated and lose their accuracy over time due to a phenomenon called drift.
- Drift is the change over time in the statistical properties of the data that was used to train a machine learning model.
- This can cause the model to become less accurate or perform differently than it was designed to.
- In other words, "drift" is the decline in a model's ability to make accurate predictions due to changes in the environment in which it is being used.

Why Do Machine Learning Models Drift?

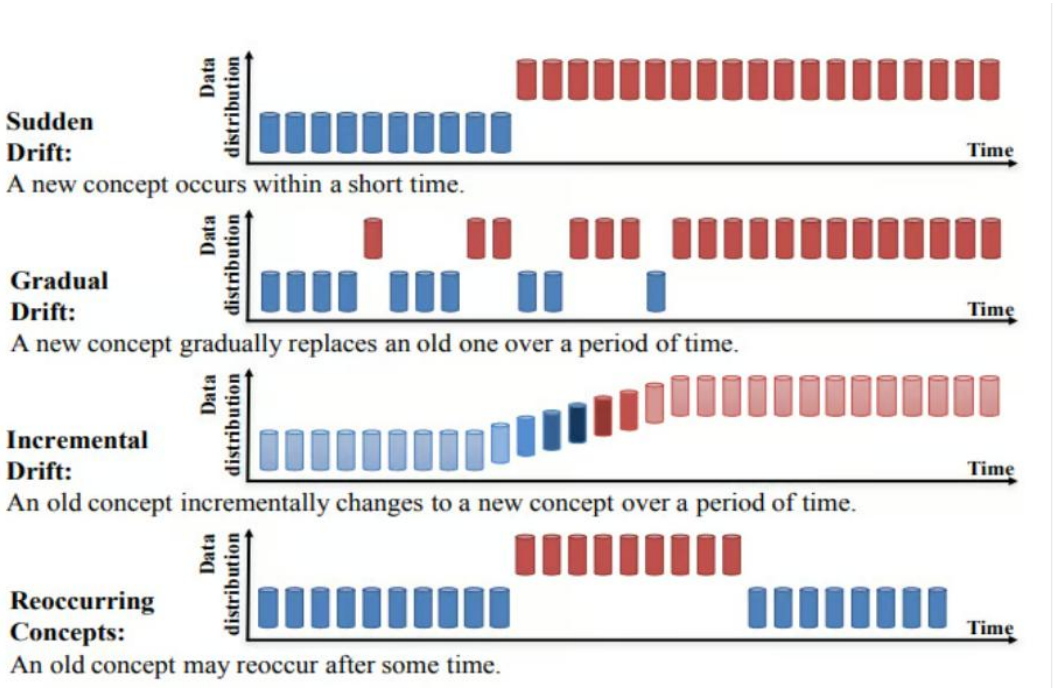
- There are several reasons why machine learning models can drift over time.
- One common reason is simply that the data that the model was trained on becomes outdated or no longer represents the current conditions.
- For example, consider a machine learning model trained to predict the stock price of a company based on historical data. If we train the model with data from a stable market, it might do well at first. However, if the market becomes more volatile over time, the model might not be able to accurately predict the stock price anymore because the statistical properties of the data have changed.
- Another reason for model drift is that the model was not designed to handle changes in the data. Some machine learning models can handle changes in the data better than others, but no model can avoid drift completely.

Types of Drift

1. Concept drift

- Concept drift, also known as model drift, occurs when the task that the model was designed to perform changes over time. For example, imagine that a machine learning model was trained to detect spam emails based on the content of the email. If the types of spam emails that people receive change significantly, the model may no longer be able to accurately detect spam.
- Concept Drift can be further divided into four categories (Learning under Concept Drift: A Review, Jie Lu et al.):

- Sudden Drift
- Gradual Drift
- Incremental Drift
- Recurring Concepts



2. Data drift

- Data drift, also known as covariate shift, occurs when the distribution of the input data changes over time. For example, consider a machine learning model that was trained to predict the likelihood of a customer purchasing a product based on their age and income. If the distribution of ages and incomes of the customers changes significantly over time, the model may no longer be able to predict the likelihood of a purchase accurately.
- It is important to be aware of both concept drift and data drift and take steps to prevent or mitigate their effects. Some strategies for addressing drift include continuously monitoring and evaluating the performance of a model, updating the model with new data, and using machine learning models that are more robust to drift.

3. LLM and Embedding Drift

- LLMs introduce a kind of drift that the methods above weren't designed for.
- The data isn't rows in a table — it's free-form text — and what shifts over time is usually the meaning of what users are asking.
- It is thought now there are three forms of LLM drift to monitor.
- Embedding drift
- Prompt or input drift

- Rubric drift

Embedding drift

- Embedding drift happens when the meaning of the text users send to a model changes, even when the surface text looks the same. LLMs convert text into long lists of numbers called embeddings, and drift can show up in those numbers without showing up in anything you'd usually be measuring.
- Imagine a customer support chatbot that originally handled mostly setup and onboarding questions. Then, six months later, the volume of messages, plus their average length, are unchanged, but users are now mostly asking about billing and cancellations. The text statistics look stable, but the embedding distribution has shifted. To detect this, teams compare batches of recent embeddings against a reference batch using a statistical distance measure.

Prompt or input drift

- Prompt drift is the same idea as embedding drift, but tracked one level up. Instead of comparing raw embeddings, you bucket incoming queries into categories — using a classifier or another LLM — and watch for shifts in the mix.
- Imagine an internal coding assistant launched for backend engineers might gradually start picking up traffic from data scientists asking pandas questions. The assistant may still answer competently, but the population it's serving is no longer the one it was tested on, and the system prompt or retrieval index may over time stop being a great fit.

Rubric drift

- Rubric drift is a change over time in the quality scores that an automated evaluator gives to a model's outputs. Many production teams now use an LLM as a judge, scoring each response on things like helpfulness, accuracy, or tone. When those scores start trending down for the same kinds of inputs, something has usually changed — the model behind an API, the documents being retrieved, or the user mix.
- What makes rubric drift especially useful is that it gives you a quality signal without needing ground truth labels, which are rarely available in real time for generative outputs.

How Do You Detect Drift?

- There are two ways we can detect drift:

1. Machine Learning Model-Based Approach: Model-based approach to detect whether the incoming input data has drifted or not.

2. Statistical Tests: There are many statistical tests to detect data drift. They are primarily divided into three categories:

- Sequential analysis methods
- A custom model to detect drift
- Time distribution method, which is very common.

- Time distribution-based methods use statistical methods to calculate the difference between two probability distributions to detect drift.
- These methods include the Population Stability Index, KL Divergence, JS Divergence, KS Test, and the Wasserstein Metric.

Algorithms for Detecting Data Drift

Kolmogorov-Smirnov (K-S) test

- The Kolmogorov-Smirnov (K-S) test is a nonparametric statistical test that is used to determine whether two sets of data come from the same distribution.
- It is often used to test whether a sample of data comes from a specific population or to compare two samples to determine if they come from the same population.
- The null hypothesis in this test is that the distributions are the same. If this hypothesis is rejected, it suggests that there is a drift in the model.
- The K-S test is a useful tool for comparing datasets and determining whether they come from the same distribution.

Population Stability Index

- The Population Stability Index (PSI) is a statistical measure that is used to compare the distribution of a categorical variable in two different datasets.
- The Population Stability Index (PSI) is a tool used to measure how much the distribution of a variable has changed between two samples or over time. It is commonly used to monitor changes in the characteristics of a population and to identify potential problems with the performance of a machine learning model.
- The PSI was originally developed to monitor changes in the distribution of a score in risk scorecards, but it is now used to examine distributional shifts for all model-related attributes, including both dependent and independent variables.
- A high PSI value indicates that there is a significant difference between the distributions of the variable in the two datasets, which may suggest that there is a drift in the model.
- If the distribution of a variable has changed significantly, or if several variables have changed to some extent, it may be necessary to recalibrate or rebuild the model to improve its performance.

Page-Hinkley method

- The Page-Hinkley method is a statistical method used to detect changes in the mean of a series of data over time. It is commonly used to monitor the performance of machine learning models and detect changes in the distribution of the data that may indicate model drift.
- To use the Page-Hinkley method, the first step is to define a threshold value and a decision function. The threshold value is a value above which a change in the mean is considered significant, and the decision function is a function that returns a value of 1 if a change has been detected and a value of 0 if no change has been detected.
- Next, the mean of the data series is calculated at each time step, and the decision function is applied to the data to determine if a change has occurred.

- If the decision function returns a value of 1, it indicates that a change has been detected and the model may be drifting.
- The Page-Hinkley method is a simple and effective way to detect changes in the mean of a data series over time. It is particularly useful for detecting small changes in the mean that may not be immediately apparent when looking at the data.