140. Instance based Vs Model based learning

Instance Based Learning vs. Model Based Learning

Summary

- **Learning Approaches**: Machine learning models understand data through two primary methods: **Instance Based Learning** (memorizing) and **Model Based Learning** (generalizing) .
- **Instance Based Learning**: Relies entirely on the training data for prediction, functioning like a "domain expert" that references specific past examples rather than learning a general rule. It is characterized by **memorizing** the data.
- **Model Based Learning**: Focuses on discovering patterns and relationships within the data to create a **generalized model** or **decision boundary**. It is characterized by **learning the pattern** of the data.
- **Operational Differences**: The two approaches differ significantly in terms of training time, storage requirements, and scoring (prediction) speed.

Exam Notes

Memorization vs. Generalization

Question: What is the fundamental difference between how instance-based and model-based algorithms process information?

Answer: **Instance-based learning** focuses on **memorizing** the training data and uses specific data points to make predictions for new queries. **Model-based learning** focuses on **generalizing** rules from the data to learn **patterns** and create decision boundaries, allowing it to predict without referencing the original dataset.

Storage and Performance

Question: Which learning type is more efficient for storage and scoring speed, and why?

Answer: **Model-based learning** is generally more efficient. It requires **less storage** because it saves only the learned model parameters (serialized files) and can discard the training data. It also offers **faster scoring** for new instances because the mathematical rules are already established. Instance-based learning is slower at scoring because it must process the training data for every new query and requires storing the entire dataset.

Concept Overview

When solving regression or classification problems, the way a machine learning model "learns" defines its category. The distinction lies in whether the model depends continuously on the training data or if it distills that data into a mathematical rule.

Instance Based Learning (Memorizing)

Instance-based learning does not try to understand the underlying patterns immediately. Instead, it relies "religiously" on the training data.

- **Mechanism**: When a new query point arrives, the system looks at the **surrounding data** (neighbors) to make a decision.
- **Analogy**: It acts like a **domain expert** who recalls specific past experiences to solve a new problem rather than applying a general theory.
- **Example**: If a student plays many hours and studies few hours, the model looks at other students with similar hours. If the majority failed, the model predicts "fail" for the new student .
- **Key Algorithm**: **K-Nearest Neighbors (KNN)** is a classic example of this technique.

Model Based Learning (Generalizing)

Model-based learning attempts to understand the **math intuition** and patterns behind the data to create a generalized rule.

- **Mechanism**: The model analyzes the training data to discover patterns and establishes a **decision boundary** (e.g., a line or curve). Points above the boundary might be classified as "pass," and points below as "fail".
- **Generalization**: Once the pattern is learned, the model is **generalized**, meaning it can predict outcomes for new data using the learned rule without needing the original data points.
- **Serialization**: The learned model is often stored in a serialized file format (like **Pickle**, **HDF5**, or **.h5**) which contains the mathematical equations and parameters.

Key Differences

The following comparison highlights the operational distinctions between the two approaches.

1. Training Process

- **Model Based**: actively trains on the data to estimate model parameters and **discover patterns**.
- **Instance Based**: Does **not train** a model immediately. Pattern discovery is postponed until a query is actually received.

2. Generalization

- **Model Based**: Generalizes rules and creates a model **before** any scoring instance is seen.
- **Instance Based**: No generalization happens before scoring. It generalizes only for each specific scoring instance individually when it is seen.

3. Prediction (Scoring)

- **Model Based**: Predicts for unseen instances using the **stored model** (mathematical equation).
- **Instance Based**: Predicts for unseen instances using the **training data directly**.

4. Data dependency

- **Model Based**: You **can throw away** the input/training data after the model is trained because the patterns are preserved in the model file.
- **Instance Based**: You **cannot throw away** the data. The input/training data must be kept because every new query relies on parts or the full set of training observations.

5. Storage Requirements

- **Model Based**: Generally requires **less storage**. The model is saved as a small file (KB or MB) containing the serialized parameters.
- **Instance Based**: Generally requires **more storage** because the entire training dataset (potentially millions of records) must be retained.

6. Scoring Speed

- **Model Based**: Scoring is **generally fast**. Since the math equation is ready, inputs are processed immediately.
- **Instance Based**: Scoring **may be slow**. The system must calculate distances (e.g., Euclidean distance) and compare the new point against the existing data points at runtime.