AI/ML - Session 1 (26th August 2024)

By - Heta Rahul Patel

Introduction to Data Science

- **Definition**: Data science involves collecting, processing, and analyzing data to extract insights. Data science involves extracting insights from data using scientific methods, algorithms, and systems. It combines fields like statistics, computer science, and domain expertise.
- **Workflow**:
  - **Data Collection**: Scraping data from APIs, databases, or web sources.
  - **Data Cleaning**: Handling missing data, removing duplicates.
  - **Exploration**: Identifying patterns using statistics and plots.
  - **Modeling**: Applying machine learning to make predictions.
  - **Interpretation**: Converting results into actionable insights.
- **Examples**: Python libraries commonly used are pandas, NumPy, and matplotlib.

Takeaways:

- Data science bridges raw data and business insights.
- It's both exploratory (finding patterns) and predictive (forecasting).
- Think of data science as using detective work on data to reveal patterns and drive decisions.

Real-World Use Cases

- **E-commerce (Amazon)**: Product recommendations based on user behavior (collaborative filtering, deep learning).
- **Streaming (Netflix/Spotify)**: Tailoring movie suggestions based on your past viewing history (recommendation engines).
- **Finance**: Fraud detection through pattern recognition in transaction data (anomaly detection).
- **Voice Assistants (Alexa/Siri)**: Understand and respond to commands (voice recognition, natural language processing).
- **Autocorrect and Predictive Text**: Next word suggestion, auto-correct in mobile devices (federated learning).
- **Ride-Hailing Apps (Uber/Ola)**: Predicting demand, setting surge pricing, and optimizing driver routes (time series forecasting, reinforcement learning).
- **Targeted Ads**: Personalized ads based on your browsing and/or purchase history (collaborative filtering, recommendation systems).

- **Personalized Feeds (Instagram/Facebook)**: Analyze your interactions (likes, comments, shares) to prioritize posts that are most likely to engage you (content-based filtering, rank aggregation).

Takeaways:

- Data science touches many industries, driving personalization, efficiency, and predictions.
- We interact with AI/ML daily, and it has become an invisible part of our everyday lives.

Data Explosion

- Today, we generate more data in a single day than was created in the entire history of humanity up until 2005.
- **Growth Stats**: In 2000, the total digital data generated was around 2 exabytes. In 2024, we generate over 2.5 quintillion bytes *per day*.
- To provide a more specific comparison:
  - Early 2000s: Data generation was relatively low, primarily driven by personal computers and early internet usage.
  - Mid-2000s: The rise of social media platforms like Facebook and YouTube began to significantly increase data generation.
  - Since 2010s: The proliferation of smartphones, IoT devices, cloud computing, and streaming services led to an explosion in data production.
- Link: How much data do we generate daily?

Takeaways:

- The data explosion has created a new reality. Scalable tools and data science are essential for navigating this new era of Big Data.
- Since we live in a data-driven world, the ability to harness and analyze massive datasets is critical for success.
- Data science allows us to extract value and derive insights from this flood of data.

Role of Data Scientist

- **Responsibilities**: Data collection, model building, and storytelling with data.

- **Skillset**: Proficient in Python, SQL, stats, machine learning, and domain knowledge.
- Example: A data scientist working for an online retailer
  - Analyze customer purchase data to identify patterns and trends.
  - Use this information to create personalized product recommendations, optimize marketing campaigns, or improve inventory management.

Takeaways:

- Data scientists are data storytellers. They bridge the gap between technical models and business insights.
- Successful data scientists combine programming prowess with domain expertise and strong communication skills.
- They translate complex data models into actionable insights that non-technical stakeholders can understand and implement.

Rule-Based Systems vs Machine Learning

- **Rule-Based**:
  - **What it is**: Hard-coded rules, designed for specific tasks (e.g., "If-This-Then-That" logic).
  - **Example**: Spam filters that block emails containing specific keywords.
  - **Limitation**: Cannot adapt to new patterns—rules must be manually updated.
- **Machine Learning**:
  - **What it is**: Learns from data and makes predictions or decisions based on that learning.
  - **Example**: Spam filters that learn from user behavior and adapt to new spamming tactics.
  - **Strength**: Adaptable and improves with more data; however, it can be a black-box.

Takeaways:

- Rule-based systems are rigid but interpretable. Great for simple tasks, but not adaptable.
- Machine learning is flexible and scales better with complex data, though it requires careful monitoring to avoid bias.

Types of Machine Learning

- **Supervised Learning**:
  - **What it is**: Trains on labeled data to make predictions.
  - **Real-Life Example**: Predicting house prices using historical sales data.
  - **Algorithms**: Linear regression, decision trees, support vector machines.
- **Unsupervised Learning**:
  - **What it is**: Finds hidden patterns in unlabeled data.
  - **Real-Life Example**: Grouping customers by purchasing behavior for targeted marketing.
  - **Algorithms**: K-means clustering, principal component analysis.
- **Semi-Supervised Learning**:
  - **What it is**: Mixes a small amount of labeled data with a large amount of unlabeled data.
  - **Real-Life Example**: Google Photos labeling images—some images are manually labeled while others are inferred.
  - **Algorithms**: Label propagation, semi-supervised SVM.
- **Reinforcement Learning**:
  - **What it is**: Learns by interacting with an environment and receiving feedback.
  - **Real-Life Example**: Self-driving cars adjusting their behavior based on road conditions.
  - **Algorithms**: Q-learning, Deep Q-Networks (DQN).

Takeaways:

- Supervised learning is like studying with a teacher's answers.
- Unsupervised learning is discovering patterns on your own.
- Semi-supervised learning bridges the gap when labeled data is scarce.
- Reinforcement learning is trial and error in real-time.

Summary

- **Data Science**: Extracts insights from data through analysis, modeling, and prediction.
- **Machine Learning**: Powers everything from recommendations to self-driving cars.
- **Rule-Based Systems vs ML**: Rule-based systems work well for static conditions, while ML excels in adaptive, dynamic environments.
- **Types of ML**: Each type of machine learning serves different problems—from structured prediction to exploring unknown data and learning through action.

This journey has covered the core concepts in data science and machine learning, emphasizing real-world applications and tools that you can explore further.