Introduction to Machine Learning – Lecture Summary

Introduction

- Session led by Learnvista Private Limited, with participants including Vijay, Swapnil Awasthi, and Nakshh.
- Focus: Introduction to machine learning (ML), its necessity, basic concepts, practical examples, and upcoming course content.

Why Machine Learning?

- Motivation: Human decision-making is manageable with small data, but data volume has grown massively (from MBs to GBs per person/day), making manual analysis impossible.
- Examples: Tinder generating large behavioral datasets, personalized services requiring predictions for millions, Google's spam filter.

Data and Machine Learning

- Data quality is crucial (“garbage in, garbage out” concept). Good decisions require good data.
- Analogy: As eating low-quality food is harmful, feeding poor data to ML results in poor predictions.

Key Concepts and Terminology

- **Algorithm:** Step-by-step procedure or set of rules (explained using tea-making instructions).

- **Model:**
  A 'box' encapsulating the algorithm and data flow (analogy: putting potatoes in and getting chips out). The machinery (algorithm) transforms input into output.

- **Predictor/Input Variables:** Features used for predictions (e.g., scooter mileage, battery life).

- **Response/Output Variable:** The outcome or target to predict (e.g., price of scooter).

- **Training Data:** Data used to train the model (typically 70-80% of dataset).

- **Testing Data:** Data used to evaluate model performance.

The Machine Learning Project Lifecycle

1. **Define Objective/Problem Statement:** Understand what needs to be solved.

2. **Data Gathering:** Collect data from various sources (databases, APIs, third parties).

3. **Data Preparation:** Clean, deduplicate, and validate data (the majority of real-world effort).

4. **Data Exploration/EDA:** Perform exploratory analysis to answer business KPIs, visualize trends.

5. **Model Building:** Train the ML model using good quality data.

6. **Model Evaluation:** Test the model on testing data, assess prediction quality.

7. **Prediction/Deployment:** Use the model for decision-making or forecasting.

Types of Machine Learning

- **Supervised Learning:**
  Uses labeled output to form patterns (e.g., predicting placement based on IQ, Tom & Jerry analogy for labeled data).

- **Unsupervised Learning:**
  Works with unlabeled data, groups similar entities (clustering), forms its own labels (e.g., categorizing customers by characteristics for marketing strategies).

- **Reinforcement Learning:**
  Learning through experience and feedback (robot learning from obstacles, analogy with a child touching a wall).

Algorithms to be Covered

- **Supervised Learning:**

  - Classification: Logistic Regression (core), KNN, Naive Bayes, Decision Trees, Random Forest, Support Vector Machine (SVM).

  - Regression: Linear Regression.

- **Unsupervised Learning:**

  - Clustering: K-Means, Hierarchical Clustering.

- Logistic regression is under classification despite the name.

- Coding algorithms in Python uses libraries like sklearn/statistics; advanced knowledge not required for prebuilt modules.

Course Structure and Next Steps

- Emphasis: 60% math/theory, 30-40% Python/application.

- Review linear algebra, as it is essential for understanding algorithms.

- Initial algorithms (KNN, Naive Bayes) are easier; SVM is the most mathematically demanding.

- Shared materials: PDFs for notes and handouts (machine learning, linear algebra) available in the course folder.

- Encourage regular revision and active participation. Class schedule adjusted for vacations and breaks.

Resources and Reminders

- Confirm access to shared resources; contact coordinator for issues.

- No class on the following day; next class dates provided.

- Wishing everyone a happy new year and successful studies.