Day 1:

What Is Machine Learning?

- ➔ Machine Learning is the science (and art) of programming computers so they can learn from data.
- ➔ Here is a slightly more general definition: [Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed. —Arthur Samuel, 1959
- ➔ And a more engineering-oriented one: A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E. —Tom Mitchell, 1997
- ➔ For example, your spam filter is a Machine Learning program that can learn to flag spam given examples of spam emails (e.g., flagged by users) and examples of regular (non-spam, also called "ham") emails. The examples that the system uses to learn are called the training set. Each training example is called a training instance (or sample). In this case, the task T is to flag spam for new emails, the experience E is the training data, and the performance measure P needs to be defined; for example, you can use the ratio of correctly classified emails. This particular performance measure is called accuracy and it is often used in classification tasks.
- ➔ If you just download a copy of Wikipedia, your computer has a lot more data, but it is not suddenly better at any task. Thus, it is not Machine Learning.

Why Use Machine Learning?

- ➔ Machine Learning (ML) enables computers to learn from the data and make predictions or decisions without being explicitly programmed.
- ➔ Traditional Programming methods have limitations when dealing with complex task or task where rules are hard to define.
- ➔ ML excels in areas such as:

- Complex problems with large amounts of data
- Tasks requiring adaptation to new data
- Situations where explicit programming is impractical or impossible

➔ ML offers the ability to:

- Extract valuable insights from large datasets
- Make predictions or classifications based on patterns in data
- Automate decision-making processes

➔ ML techniques are applied across various domains:

- Finance: Fraud detection, stock market prediction
- Healthcare: Disease diagnosis, personalized medicine
- Marketing: Customer segmentation, recommendation systems
- Automotive: Autonomous driving, predictive maintenance

➔ ML has become increasingly important with the rise of Big Data, where traditional data processing methods are inadequate.

➔ ML algorithms can adapt and improve overtime leading to better performance in accuracy with experience in more data

➔ ML is essential for unlocking the potential of emerging technologies like artificial intelligence, robotics, and IoT (Internet of Things)

➔ Understanding ML concepts and techniques empowers individuals and organizations to harness the power of data for better decision-making and innovation.

Types of Machine Learning:

There are so many different types of Machine Learning systems that it is useful to classify them in broad categories based on:

1. Whether or not they are trained with human supervision (supervised, unsupervised, semi-supervised, and Reinforcement Learning)

2. Whether or not they can learn incrementally on the fly (online versus batch-learning)
3. Whether they work by simply comparing new data points to known data points, or instead detect patterns in the training data and build a predictive model, much like scientists do (instance-based versus model-based learning)

1. **Supervised Learning:**
    - Trains the model on labeled data, where each training example has an input-output pair.
    - Common tasks include regression (predicting a continuous value) and classification (predicting a class label).
    - For example, A typical supervised learning task is classification. The spam filter is a good example of this: it is trained with many example emails along with their class (spam or ham), and it must learn how to classify new emails.
      Another typical task is to predict a target numeric value, such as the price of a car, given a set of features (mileage, age, brand, etc.) called predictors. This sort of task is called regression. To train the system, you need to give it many examples of cars, including both their predictors and their labels (i.e., their prices).
      Note that some regression algorithms can be used for classification as well, and vice versa. For example, Logistic Regression is commonly used for classification, as it can output a value that corresponds to the probability of belonging to a given class (e.g., 20% chance of being spam).

2. **Unsupervised Learning:**
    - Works with unlabeled data, where the algorithm tries to learn patterns without explicit supervision.
    - Common tasks include clustering (grouping similar instances together) and dimensionality reduction (simplifying data while retaining important information).
    - For example, For example, say you have a lot of data about your blog's visitors. You may want to run a clustering algorithm to try to detect groups of similar visitors. At no point do you tell the algorithm which group a visitor belongs to: it finds those

connections without your help. For example, it might notice that 40% of your visitors are males who love comic books and generally read your blog in the evening, while 20% are young sci-fi lovers who visit during the weekends, and so on. If you use a hierarchical clustering algorithm, it may also subdivide each group into smaller groups. This may help you target your posts for each group.

- Here are some of the most important supervised learning algorithms (covered in this book): • k-Nearest Neighbors • Linear Regression • Logistic Regression • Support Vector Machines (SVMs) • Decision Trees and Random Forests • Neural networks

3. **Semi-supervised Learning:**
    - Uses a mix of labeled and unlabeled data for training.
    - Typically, the amount of labeled data is smaller than the amount of unlabeled data.
    - Some photo-hosting services, such as Google Photos, are good examples of this. Once you upload all your family photos to the service, it automatically recognizes that the same person A shows up in photos 1, 5, and 11, while another person B shows up in photos 2, 5, and 7. This is the unsupervised part of the algorithm (clustering). Now all the system needs is for you to tell it who these people are. Just one label per person, and it is able to name everyone in every photo, which is useful for searching photos.
    - Here are some of the most important unsupervised learning algorithms: • Clustering — K-Means — DBSCAN — Hierarchical Cluster Analysis (HCA) • Anomaly detection and novelty detection — One-class SVM — Isolation Forest • Visualization and dimensionality reduction — Principal Component Analysis (PCA) — Kernel PCA — Locally-Linear Embedding (LLE) — t-distributed Stochastic Neighbor Embedding (t-SNE) • Association rule learning — Apriori — Eclat

4. **Reinforcement Learning**
    - Involves an agent learning to interact with an environment by taking actions and receiving rewards or penalties.
    - The goal is to learn a policy that maximizes the cumulative reward over time.

- For example, For example, many robots implement Reinforcement Learning algorithms to learn how to walk. DeepMind’s AlphaGo program is also a good example of Reinforcement Learning: it made the headlines in May 2017 when it beat the world champion Ke Jie at the game of Go. It learned its winning policy by analyzing millions of games, and then playing many games against itself. Note that learning was turned off during the games against the champion; AlphaGo was just applying the policy it had learned.

5. **Batch Learning**
    - Trains the model using all available data at once.
    - Typically requires retraining the model from scratch on the entire dataset whenever new data is available.
6. **Online Learning**
    - Trains the model incrementally as new data becomes available.
    - Well-suited for applications with a continuous flow of data and limited computational resources.
7. **Instance – based Learning:**
    - Memorizes the training instances and uses them to make predictions.
    - Common algorithms include k-Nearest Neighbors (k - NN), where the prediction is based on the majority class among the k nearest neighbors.
8. **Model – based Learning:**
    - Generalizes from a set of training examples to make predictions on new instances.
    - Involves selecting a model architecture, learning its parameters from the training data, and then using the model to make predictions.