Chapter 0 : Part 2

**ML : What? Why? How?**

Understanding the concepts.

**0.3 What Is Machine Learning?**

Machine learning is a data analytics technique that teaches computers to do what comes naturally to humans and animals: learn from experience.

* Machine Learning: Machine learning algorithms use computational methods to ***“learn” information directly from data*** without relying on a *predetermined* *equation* as a model. The algorithms adaptively ***improve their performance*** as the ***number of samples*** available for learning ***increases***. Deep learning is a specialized form of machine learning.

**0.4 Why Machine Learning Matters**

With the rise in big data, machine learning has become a key technique for solving problems in areas, such as:

1. ***Computational*** ***finance***, for credit scoring and algorithmic trading
2. ***Image*** ***processing*** and ***computer*** ***vision***, for face recognition, motion detection, and object detection
3. ***Computational*** ***biology***, for tumor detection, drug discovery, and DNA sequencing
4. ***Energy*** ***production***, for price and load forecasting
5. ***Automotive***, aerospace, and manufacturing, for predictive maintenance
6. ***Natural*** ***language*** processing, for ***voice*** ***recognition*** applications

* Machine learning algorithms find ***natural*** ***patterns*** in data that ***generate*** ***insight*** and help you make better decisions and predictions. They are used every day to make critical decisions in *medical diagnosis, stock trading, energy load forecasting*, and more.

For example, media sites rely on machine learning to sift through millions of options to give you song or movie *recommendations*. Retailers use it to gain insight into their customers’ purchasing behavior.

* When Should You Use ML: Consider using machine learning when you have a complex task or problem *involving a large amount of data* and *lots of variables*, but no existing formula or equation. For example, machine learning is a good option if you need to handle situations like these:
* Hand-written rules and equations are too complex—as in face recognition and speech recognition.
* The rules of a task are constantly changing—as in fraud detection from transaction records.
* The nature of the data keeps changing, and the program needs to adapt—as in automated trading, energy demand forecasting, and predicting shopping trends.

**0.5 How Machine Learning Works**

Walk through the three types of machine learning (***clustering***, ***classification***, and ***regression***). Machine learning uses two types of techniques:

1. Supervised learning: which **trains** a model on **known** **input** and output data so that it can **predict** **future** outputs,
2. Unsupervised learning: which **finds hidden patterns** or **intrinsic** **structures** in input data.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  Machine learning | |  |
| SUPERVISED LEARNING  Develop predictive model based on both input and output data | | UNSUPERVISED LEARNING  Group and interpret data based only on input data | |
| CLASSIFICATION | REGRESSION | CLUSTERING | |

**0.6 Supervised Learning**

* Supervised Learning: Supervised machine learning builds a model that makes *predictions* based on *evidence* in the *presence of uncertainty*.
* A *supervised learning algorithm* takes a *known set of input data* and *known* *responses* to the data (output) and trains a model to generate reasonable predictions for the response to new data.
* Use *supervised* *learning* if you have known data *for* the *output* you are trying to *predict*.
* Techniques for SUPERVISED LEARNING: Supervised learning uses CLASSIFICATION and REGRESSION techniques to develop machine learning models.

**0.6.1 Classification**

* Classification: Classification techniques *predict discrete responses*—for example, whether an *email is genuine or spam*, or *whether a tumor is cancerous or benign*.
* Classification models classify input data into categories. Typical applications include *medical* *imaging*, *speech* *recognition*, and *credit* *scoring*.
* Use classification if your data can be tagged, categorized, or separated into specific groups or classes. For example, applications for *hand-writing recognition* use *classification* to recognize letters and numbers.
* In image processing and computer vision, Unsupervised Pattern Recognition techniques are used for object detection and image segmentation.
* Common CLASSIFICATION Algorithms: Common algorithms for performing classification include

1. Support Vector Machine (svm),
2. Boosted and Bagged DECISION TREES,
3. K-Nearest Neighbor,
4. Naïve bayes,
5. Discriminant analysis,
6. Logistic Regression, and
7. Neural Networks.

**0.6.2 Regression**

Regression: Regression techniques *predict continuous responses*—for example, *changes in temperature* or *fluctuations in power demand*. Typical applications include electricity load forecasting and algorithmic trading.

Use regression techniques if you are working with a *data range* or if the nature of your *response is a real number*, such as *temperature* or the *time* until failure for a piece of equipment.

* Common REGRESSION Algorithms: Common regression algorithms include

1. Linear model,
2. Nonlinear model,
3. Regularization,
4. Stepwise Regression,
5. Boosted and Bagged Decision Trees,
6. Neural Networks, and
7. Adaptive Neuro-Fuzzy Learning.

* Example of supervised learning:
* Using Supervised Learning to Predict Heart Attacks: Suppose clinicians want to predict whether someone will have a heart attack within a year. They have data on previous patients, including ***age***, ***weight***, ***height***, and ***blood*** ***pressure***. They know whether the previous patients had heart attacks within a year. So the problem is combining the existing data into a model that can predict whether a new person will have a heart attack within a year.

**0.7 Unsupervised Learning**

Unsupervised learning finds ***hidden*** ***patterns*** or ***intrinsic*** ***structures*** in data. It is used to ***draw inferences from datasets*** consisting of input data without labeled responses.

**Clustering**

Clustering is the most common unsupervised learning technique. It is used for exploratory data analysis to find *hidden* *patterns* or *groupings* in data. Applications for cluster analysis include Gene Sequence analysis, Market Research, and Object Recognition.

* Example:
* For example, if a cell phone company wants *optimize* the *locations* where they build *cell phone towers*, they can use machine learning to ***estimate*** the ***number of clusters of people*** relying on their towers. A phone can only talk to one tower at a time, so the team uses Clustering Algorithms to design the ***best placement of cell towers*** to optimize Signal Reception for ***groups***, or ***clusters***, of their ***customers***.
* Common CLUSTERING algorithms:

1. K-Means and K-Medoids,
2. Hierarchical Clustering,
3. Gaussian Mixture models,
4. Hidden Markov Models,
5. Self-Organizing Maps,
6. Fuzzy C-Means Clustering, and
7. Subtractive Clustering.

**0.8 Which Machine Learning Algorithm to Use?**

Choosing the right algorithm can seem overwhelming—there are dozens of supervised and unsupervised machine learning algorithms, and each takes a different approach to learning.

* There is no best method or one size fits all. Finding the right algorithm is partly just trial and error—even highly experienced data scientists can’t tell whether an algorithm will work without trying it out.
* Algorithm selection also *depends* on the *size and type of data* you’re working with, the insights you want to get from the data, and how those insights will be used.

|  |  |
| --- | --- |
| * Here are some *guidelines* on choosing between *Supervised* and *Unsupervised* machine learning: * Choose Supervised Learning if you need to train a model to make a Prediction—for example, the future value of a continuous variable, such as temperature or a stock price, or a classification—for example, identify makes of cars from webcam video footage. * Choose *unsupervised* learning if you need to *explore your data* and want to *train* a *model* to find a good *internal* *representation*, such as ***splitting*** *data up into* ***clusters***. |  |