Objectives

In this Lecture you will learn:

- What is machine learning.
- The role of dataset in the machine learning.
- Machine learning categories.
- Regression and classification.

1. Introduction

Our imaginations have long been captivated by visions of machines that can learn and **imitate human intelligence**.

We use machine learning ML programs to **discover** new music that we might enjoy, and to find exactly the shoes we want to purchase online.

ML programs allow us to dictate commands to our smart phones, and allow our thermostats to set their own temperatures.

ML programs can decipher sloppily-written mailing addresses better than humans, and can guard credit cards from fraud more vigilantly.

ML is the design and study of software artifacts **that use past experience** to **inform future decisions**. It is also the study of programs that **learn from data**.

The fundamental goal of machine learning is to generalize, or to **induce an unknown rule from examples** of the rule's application.

2. Learning from Experience

Machine learning systems are often described as learning from experience either with or without supervision from humans.

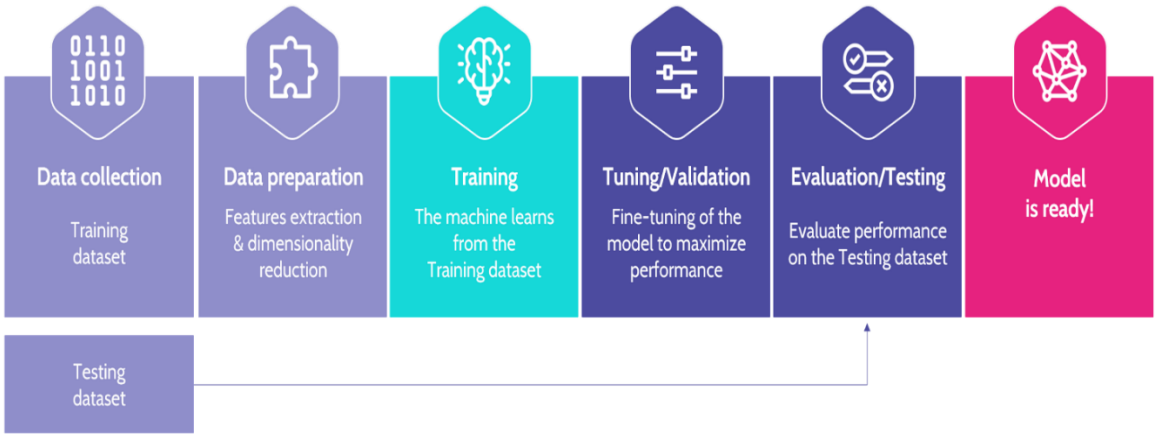

Fig1. General steps of ML Algorthim.

Generally all ML algorthim uses **training**, **testing**, and **validation** data to build the final model.

A **training set** is a collection of observations (data). These observations comprise the experience that the algorithm uses to **learn**.

The test set is used to **evaluate** the performance of the model using some performance metric. It is important that no observations from the training set are included in the test set. If the test set does contain examples from the training set, it will be difficult to assess whether the algorithm has learned to generalize from the training set or has simply memorized it. A program that generalizes well will be able to effectively perform a task with new data.

In contrast, a program that memorizes the training data by learning an overly-complex model could predict the values of the response variable for the training set accurately, but will fail to predict the value of the response variable for new examples. Memorizing the training set is called overfitting. Balancing generalization and memorization is a problem common to many machine learning algorithms.

In addition to the training and test data, a third set of observations, called a **validation** or **hold-out set**, is sometimes required. The validation set is used to tune variables called **hyperparameters** that control how the algorithm learns from the training data.

It is common to partition a single set of supervised observations into training, validation, and test sets. There are no requirements for the sizes of the partitions, and they may vary according to the amount of data available. It is common to allocate between fifty and seventy-five percent of the data to the training set, ten to twenty-five percent of the data to the test set, and the remainder to the validation set.

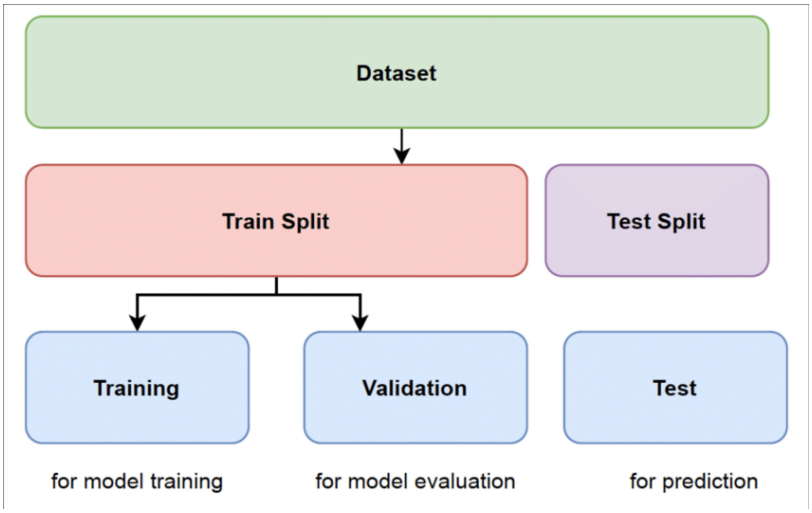

Fig2. Dataset Splitting.

There three machine learning types are supervised, unsupervised, and reinforcement learning. We would discuss two type as this course suggested.

2.1 Supervised Learning

In supervised learning problems, a program **predicts an output** for an **input** by learning from pairs of **labeled inputs and outputs**.

Supervised learning is a technique in which we fit the model with both inputs(**features**) and **outputs**(**labels**) of the data "learning something new under the supervision of a teacher".

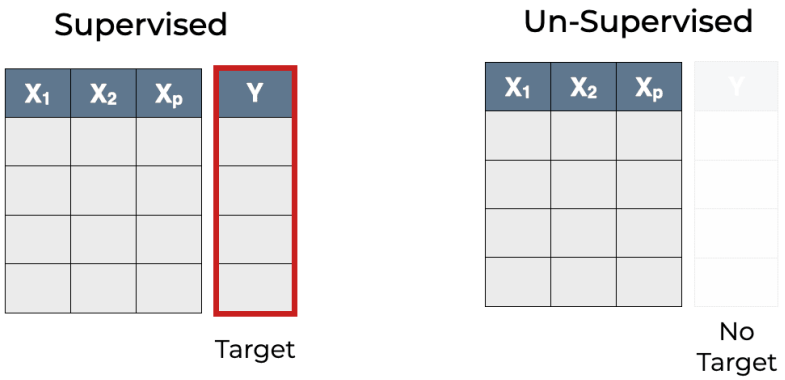

Fig. 3. Labeled and unlabaled dataset.

An unlabled data consists of samples of natural or human-created artifacts that you can obtain relatively easily from the world. Some examples of unlabeled data might include photos, audio recordings, videos, news articles, tweets, x-rays (if you were working on a medical application), etc. There is no "explanation" for each piece of unlabeled data -- it just contains the data, and nothing else.

A labeled data typically takes a set of unlabeled data and augments each piece of that unlabeled data with some sort of meaningful "tag," "label," or "class" that is somehow informative or desirable to know. For example, labels for the above types of unlabeled data might be whether this photo contains a horse or a cow, which words were uttered in this audio recording, what type of action is being performed in this video, what the topic of this news article is, what the overall sentiment of this tweet is, whether the dot in this x-ray is a tumor, etc.

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

labels

Fig.4 . Labeled Data.

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio
1	41	2	6	19	0.124	1.073	NBA001	6.3
2	47	1	26	100	4.582	8.218	NBA021	12.8
3	33	2	10	57	6.111	5.802	NBA013	20.9
4	29	2	4	19	0.681	0.516	NBA009	6.3
5	47	1	31	253	9.308	8.908	NBA008	7.2
6	40	1	23	81	0.998	7.831	NBA016	10.9
7	38	2	4	56	0.442	0.454	NBA013	1.6
8	42	3	0	64	0.279	3.945	NBA009	6.6
9	26	1	5	18	0.575	2.215	NBA006	15.5
10	47	3	23	115	0.653	3.947	NBA011	4
11	44	3	8	88	0.285	5.083	NBA010	6.1
12	34	2	9	40	0.374	0.266	NBA003	1.6

unlabeled

Fig. 5. Unlabeled Data.

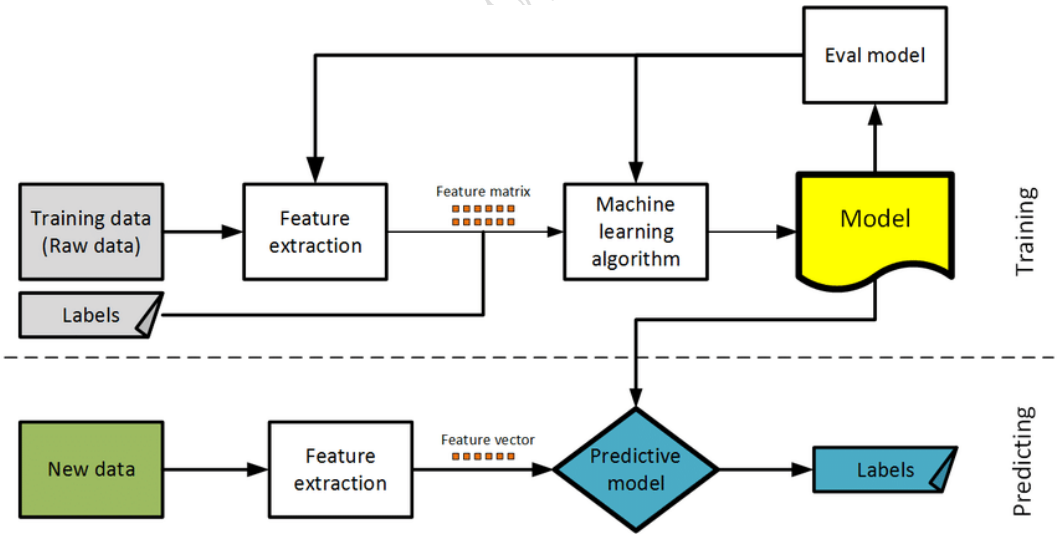

Fig. 6. Supervised ML algorthim General Structure.

There are two types of supervised learning techniques: **regression** where the results are continuous values (numbers), and **classification** where our result set consist of categories (words).

2.2 Unsupervised Learning

Unsupervised learning is the case where we fit the model without known outputs. In unsupervised learning, a program does not learn from labeled data. Instead, it attempts to

discover patterns in data. For example, assume that you have collected data describing the heights and weights of people. An example of an unsupervised learning problem is dividing the data points into groups. A program might produce groups that correspond to men and women, or children and adults.

There are three types of unsupervised learning techniques: **Clustering** where data is grouped in a meaningful way. **Dimensionality Reduction** where high-dimensional data is represented with low-dimensional data. **Association** where the relationships between variables in a big dataset is discovered.

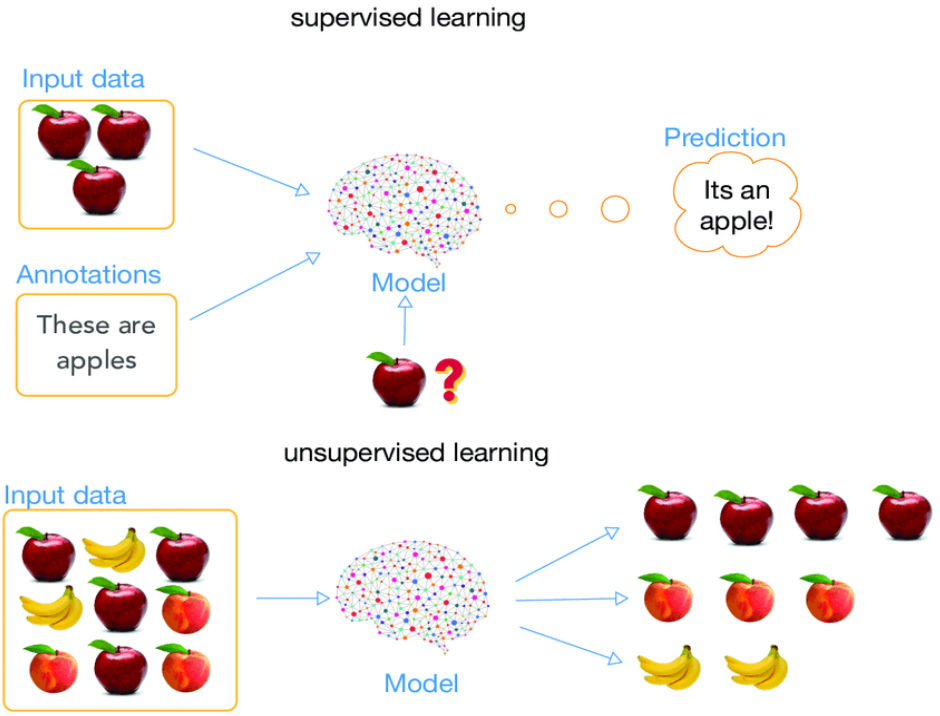

Fig. 7. Supervised vs Unsupervised.

3. features & feature vector

In supervised learning , the dataset is the collection of labeled examples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ Each element xi among N is called a **feature vector. A feature vector** is a vector in which each dimension j = 1, . . . ,D contains a value that describes the example somehow. **Feature vectors** are used to represent numeric or symbolic characteristics, called features, of an object in a mathematical, easily analyzable way.

Feature: is a list of numbers eg: age, name, height, weight etc., that means every column is a feature in relational table.

Feature Vector is representation of particular row in relational table. Each row is a feature vector, row 'n' is a feature vector for the 'n'th sample.

ID	First Name	Last Name	Email	Year of Birth
1	Peter	Lee	plee@university.edu	1992
2	Jonathan	Edwards	jedwards@university.edu	1994
3	Marilyn	Johnson	mjohnson@university.edu	1993
6	Joe	Kim	jkim@university.edu	1992
12	Haley	Martinez	hmartinez@university.edu	1993
14	John	Mfume	jmfume@university.edu	1991
15	David	Letty	dletty@university.edu	1995

Feature Vector

Fig. 8. Feature Vector .

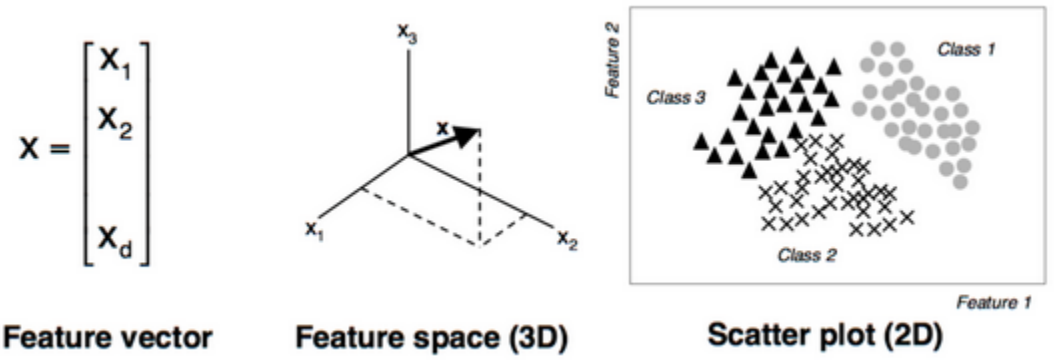

Fig. 9. Vectors & spaces.

Machine Learning Regression is a technique for investigating the relationship between independent variables or features and a dependent variable or outcome. It's used as a method for predictive modelling in machine learning, in which an algorithm is used to predict continuous outcomes. Examples are sales forecasting, demand forecasting, stock price prediction, Weather forecasting.

3.1 Regression

One example is the Simple Linear Regression SLR is a type of Regression algorithms that models the relationship between a dependent variable and a single independent variable. The relationship shown by a Simple Linear Regression model is linear or a sloped straight line, hence it is called Simple Linear Regression.

ID	X	Y
ID1	1	1.2
ID2	2	2.1
ID3	3	3.0
ID4	4	3.9
ID5	5	4.8
IDnew	6	?

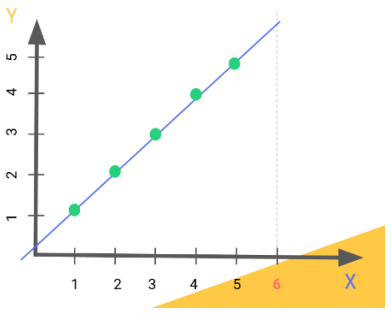

Fig. 10. SLR.

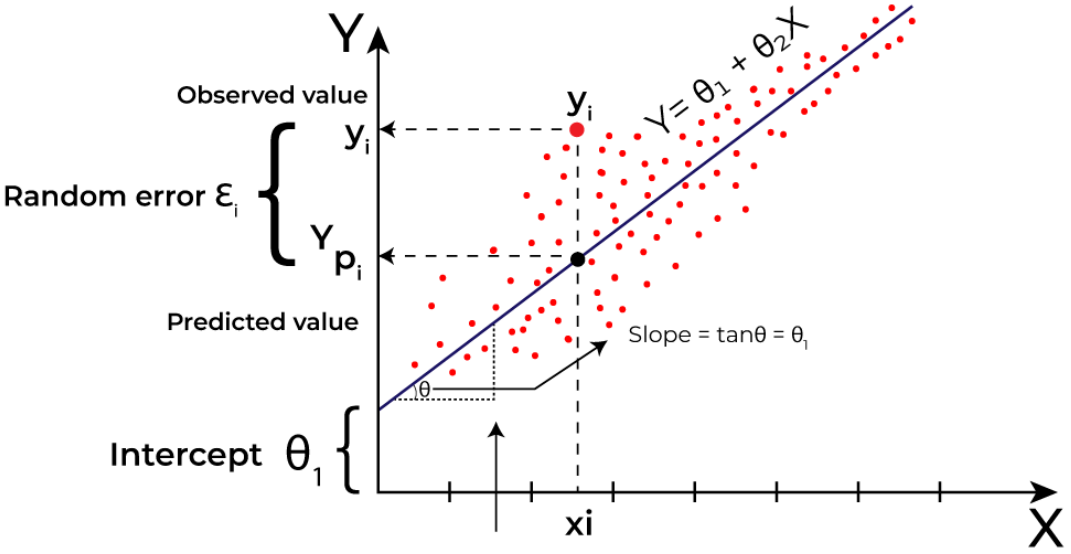

Fig. 11. Mathmatics behine the SLR .

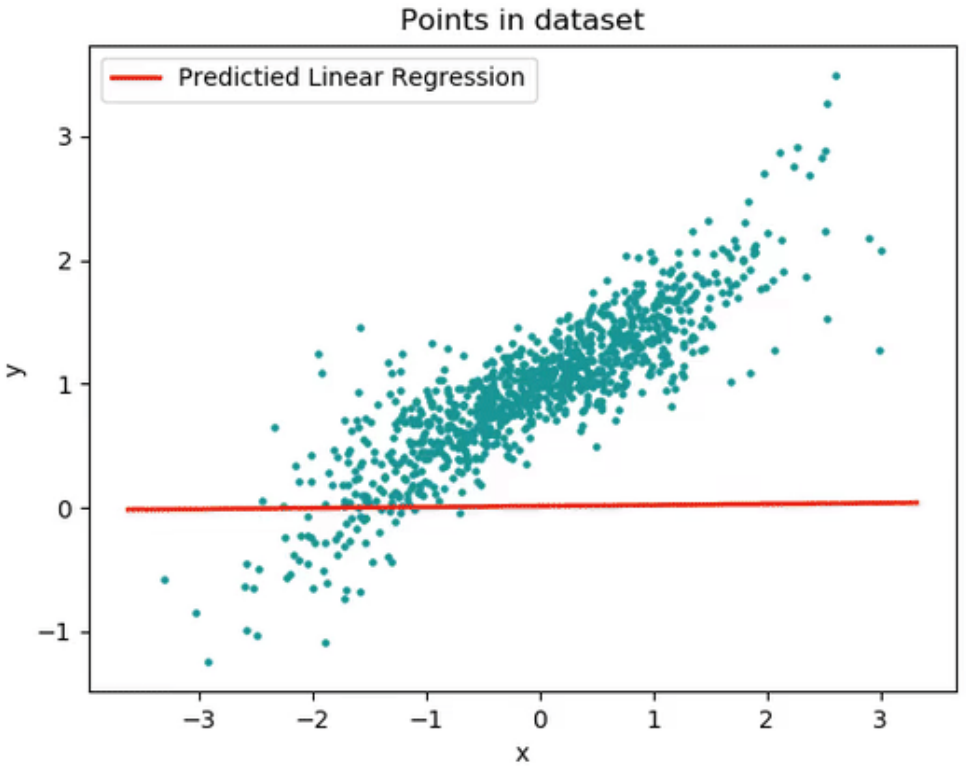

Fig. 12. Mathmatics behine the SLR .