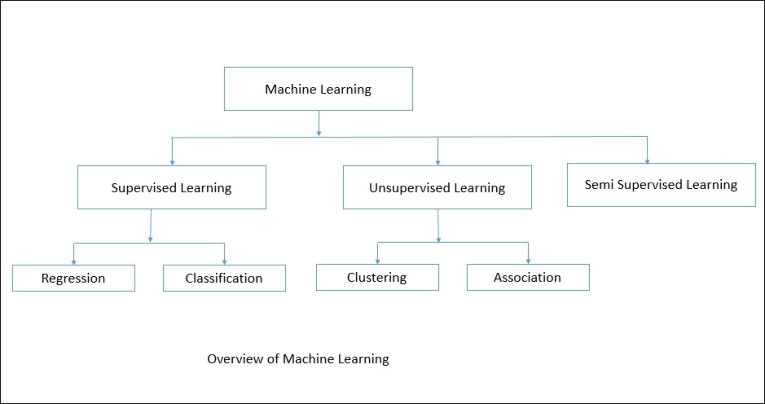
1. What is machine learning?

Machine learning is a stream of engineering which uses mathematics to allow machines to

make classifications, predictions, recommendations, and so on, based on the data provided to them.



1.1

Supervised learning

In supervised learning, we are generally given an input dataset, which is a historical record of

actual events. We are also given what the expected output should look like. Using the historical

data, we choose which factors contributed to the results. Such attributes are called features.

Using the historical data, we understand how the previous results were calculated and apply

that same understanding to the data on which we want to make predictions

Supervised learning can be again subdivided into:

Regression（回归）

Classification（分类）

1.1.1

Regression（回归）

In regression problems, we try to predict results using inputs from a continuous function.

Regression means predicting the score of one variable based on the scores of another

variable. The variable we will be predicting is called the criterion variable, and the variable from

which we will be doing our predictions is called the predictor variable. There can be more than

one predictor variable; in this case, we need to find the best fitting line, called the regression

line.

1.1.2

In classification, we predict the output in discrete results. Classification, being a part of

supervised learning, also needs the input data and sample output to be given. Here, based on

the features, we try to classify the results into sets of defined categories.

1.2

Unsupervised learning

Unsupervised learning does not give us any idea about how our results should look. Instead, it

allows us to group data based on the features of the attributes. We derive the clustering based

on the relationships among the records.

Unlike supervised learning, there is no validation we can do to verify our results, which means

there is no feedback method to teach us whether we did right or wrong. Unsupervised learning

is primarily based on clustering algorithms.

1.2.1

In order to understand clustering more easily, let's consider an example; let's say we have

20,000 news articles on various topics and we have to group them based on their content . In

this case, we can use clustering algorithms, which would group set of articles into small groups.

We can also consider the basic example of fruits. Let's say we have apples, bananas, lemons,

and cherries in a fruit basket and we need to classify them into groups. If we look at their

colors, we can classify them into two groups:

Red color group: Apples and cherries

Yellow color group: Bananas and lemons

Now we can do more grouping based on another feature, its size:

Red color and large size: Apples

Red color and small size: Cherries

Yellow color and large size: Banana

Yellow color and small size: Lemons

The following diagram shows a representation of clustering:

This way, by looking at more features, we can also do more clustering. Here, we don't have

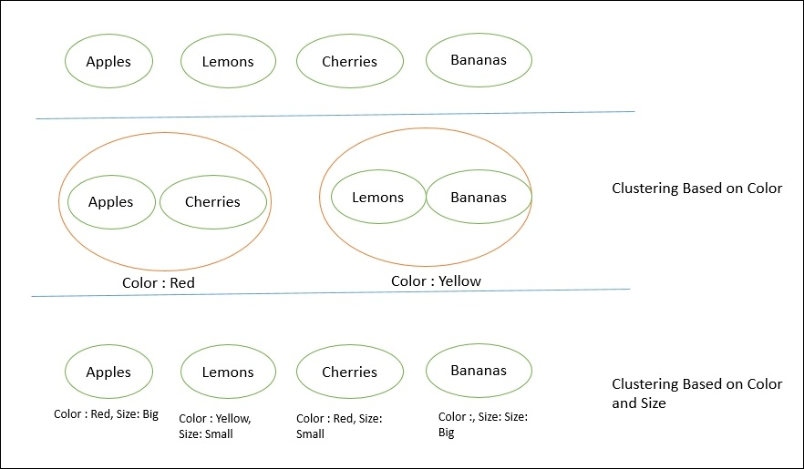
any training data and a variable to be predicted, unlike in supervised learning. Our only task is

to learn more about the features and cluster the records based on inputs.

1.2.2

Association problems are more about learning, and making recommendations by defining

association rules. Association rules could, for example, refer to the assumption that people who bought an iPhone are more likely to buy an iPhone case.



* 1. Semi-supervised learning

Many researchers and machine learning practitioners have found that,

when labeled data is used in conjunction with unlabeled data, the results are likely to be more

accurate.