Algorithm

Supervised Algorithm

The majority of practical machine learning uses supervised learning.

Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

$$Y = f(X)$$

The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance.

Supervised learning problems can be further grouped into regression and classification problems.

- **Classification**: A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.
- **Regression**: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Some popular examples of supervised machine learning algorithms are:

1. Linear regression for regression problems.

2. Random forest for classification and regression problems.

3. Support vector machines for classification problems.

Unsupervised Algorithm

Unsupervised learning is where you only have input data (X) and no corresponding output variables.

The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

These are called unsupervised learning because unlike supervised learning above there is no correct answers and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data.

Unsupervised learning problems can be further grouped into clustering and association problems.

- **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
- **Association**: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

Some popular examples of unsupervised learning algorithms are:

1. k-means for clustering problems.
2. Apriori algorithm for association rule learning problems.

Semi-supervised Algorithm

Problems where you have a large amount of input data (X) and only some of the data is labeled (Y) are called semi-supervised learning problems.

These problems sit in between both supervised and unsupervised learning.

A good example is a photo archive where only some of the images are labeled, (e.g. dog, cat, person) and the majority are unlabeled.

Many real world machine learning problems fall into this area. This is because it can be expensive or time-consuming to label data as it may require access to domain experts. Whereas unlabeled data is cheap and easy to collect and store.

You can use unsupervised learning techniques to discover and learn the structure in the input variables.

You can also use supervised learning techniques to make best guess predictions for the unlabeled data, feed that data back into the supervised learning algorithm as training data and use the model to make predictions on new unseen data.

Summary

In this post you learned the difference between supervised, unsupervised and semi-supervised learning. You now know that:

- **Supervised**: All data is labeled and the algorithms learn to predict the output from the input data.
- **Unsupervised**: All data is unlabeled and the algorithms learn to inherent structure from the input data.
- **Semi-supervised**: Some data is labeled but most of it is unlabeled and a mixture of supervised and unsupervised techniques can be used.

Reinforcement learning

Reinforcement learning (**RL**) is an area of machine learning concerned with how software agents ought to take actions in an environment in order to maximize the notion of cumulative reward. Reinforcement learning is one of three basic machine learning paradigms, alongside supervised learning and unsupervised learning.

Reinforcement learning, due to its generality, is studied in many other disciplines, such as game theory, control theory, operations research, information theory, simulation-based optimization, multi-agent system, swarm intelligence, statistics, and genetic algorithm. In the operations research and control literature, reinforcement learning is called *approximate dynamic programming,* or *neuro-dynamic programming.*

Data

Labeled data

Labeled data consists of unlabeled data with a description, label or name of features in the data. E.g. In a labeled image dataset, an image is labeled as it is a cat's photo and it's a dog's photo.

Unlabeled data

Unlabeled data consists of data which is either taken from nature or created by human to explore the scientific patterns behind it. Some examples of unlabeled data might include photos, audio recordings, videos, news articles, tweets, x-rays, etc. The main concept is there is no explanation, label, tag, class or name for the features in data.