SUPERVISED, UNSUPERVISED & OTHER MACHINE LEARNING METHODS

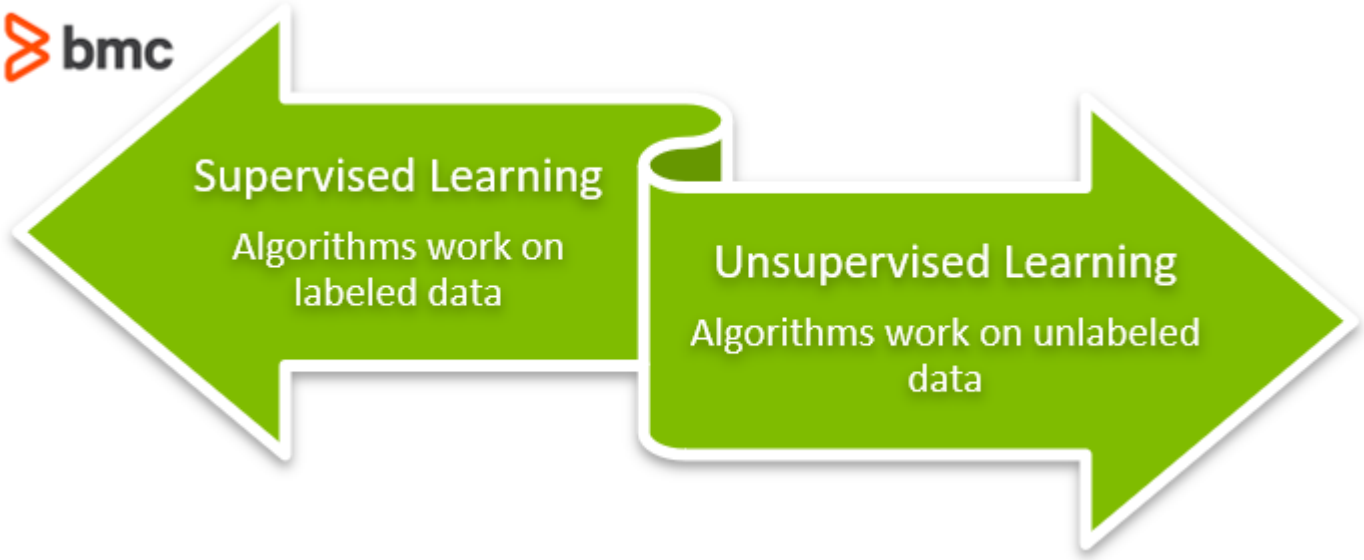
Machine learning is augmenting human capabilities and making things possible—things that just a few years back were considered impossible.

Take, for example, the protein folding problem. For about 50 years, the biology field assumed that solving this problem was beyond human capabilities. But with the might of AI and ML, folks at DeepMind were finally able to come up with a solution to this problem.

ML-based applications are ubiquitous these days and they continue to evolve day by day. Before long we might also manage to build a fully autonomous driving vehicle.

But then the question arises: how exactly do you make a machine learn?

Let's look at the two most well-known machine learning methods—supervised and unsupervised learning. We'll deep dive into how they both work, and we'll look at up-and-coming learning methods, too.



Machines can learn

We are very familiar with the paradigm of coding programs. Coding is akin to explicitly telling the machine what to do. The programmed machine cannot make a decision on its own. And it most certainly cannot handle a situation that it hasn't been programmed for.

This is like giving machines a fish when, really, we want to teach machines *how* to fish.

In the field of AI and ML the way machines are made to learn generally fall under two categories:

- Supervised learning
- Unsupervised learning

In a nutshell, the difference between these two methods is that in supervised learning we also provide the correct results in terms of labeled data. Labeled data in machine learning parlance means that we know the correct output values of the data beforehand.

In unsupervised machine learning, the data is not labeled. So, in unsupervised learning the machines are left to fend for themselves, you may ask? Not quite.

(Understand the role of data annotation in ML.)

How supervised machine learning works

The notion of 'supervision' in supervised machine learning comes from the labeled data.

With the help of labels, the predictions a machine learning model makes can be compared against the known correct values. This helps with gauging the accuracy of the model and calculation of loss. This in turn can be used as a feedback to the model to further improve its predictions. (This labeled data seems like the answer to all our problems, right? What could ever go wrong!)

But, as they say: with great power comes great responsibilities. We need to be careful with the extent we used the labels in during the supervised learning or in machine learning jargon how much we train our model.

The pitfall of too much training is overfitting. This is what happens when the ML model learns the training data so well that, when new data comes in, the model often fails to perform correctly.

(Unsupervised learning algorithm can also face overfitting, but it is more prevalent in supervised learning algorithms. Eagerness to train one more epoch, for the sake of better accuracy, often leads into overfitting.)

Broadly, supervised machine learning finds its application in 2 types of tasks:

- Classification
- Regression

Classification

In this type of tasks, the model tries to classify a given input into one of the data categories.

For example, classifying a tumor as malignant or benign. Here we train the model on the input data which has already been correctly labeled with either malignant or benign. We compare the generated output with these labels and re-train the model to achieve a robust model.

Regression

In this type of tasks, the model tries to predict a numerical value (real number).

An example of this is predicting housing prices given housing data. The key point here is that the output in this case is a real or continuous value—it's not one bucket or the other, as in classification. Again, we compare the predicted value with the known correct values and make further tweaks to improve the model.

As you can see, in both classification and regression, the labels themselves are, in a sense, "supervising" the training of the machine learning model.

How unsupervised learning works

On the other side of the aisle, unsupervised learning algorithms work on unlabeled data. This is where the notion of unsupervised comes from: there are no labels for the model to course correct against while training.

But the absence of labels does not mean that unsupervised learning methods wander aimlessly. Actually, these algorithms look for underlying patterns or connections within the data and uses that to help understand/analyze the data.

So, why would someone use unlabeled data in the first place, you may ask? There can be multiple reasons for this:

- Sometimes labeled data is simply not available.
- Often, the cost to label the data is very high.
- The size of the data is so huge that it is impossible to get labels added in a reasonable time.

This is not all bad. In fact, you'll often use unsupervised learning algorithms when you *don't know* what you are looking for.

For example, you have demographic data of various grocery shoppers from a city and you want to group the users into logical groups. Then the unsupervised learning algorithms can help identify the clusters/groups in the data. This is called clustering, and it's the most common application of

unsupervised learning algorithms.

Unsupervised learning algorithms are not limited to clustering tasks alone. Other applications are reducing dimensions and estimating density.

Clustering

As we saw in the example, clustering is where an algorithm finds similarities within the data points and groups similar data together. This can be based on:

- Distance (K-Means)
- Density (DBSCAN)
- And other characteristics

(Learn about common ML architectures, including K-Means.)

Dimensionality reduction

Often, the data contains way too many features—and not all features contribute equally to the predicting power of the model. So, using unsupervised learning algorithms can help to remove superfluous features from the dataset.

Density estimation

The aim of density estimation is to:

1. Discover relations among attributes in data.
2. Generate underlying probability density function based on the data.

One common use case for this method is anomaly detection.

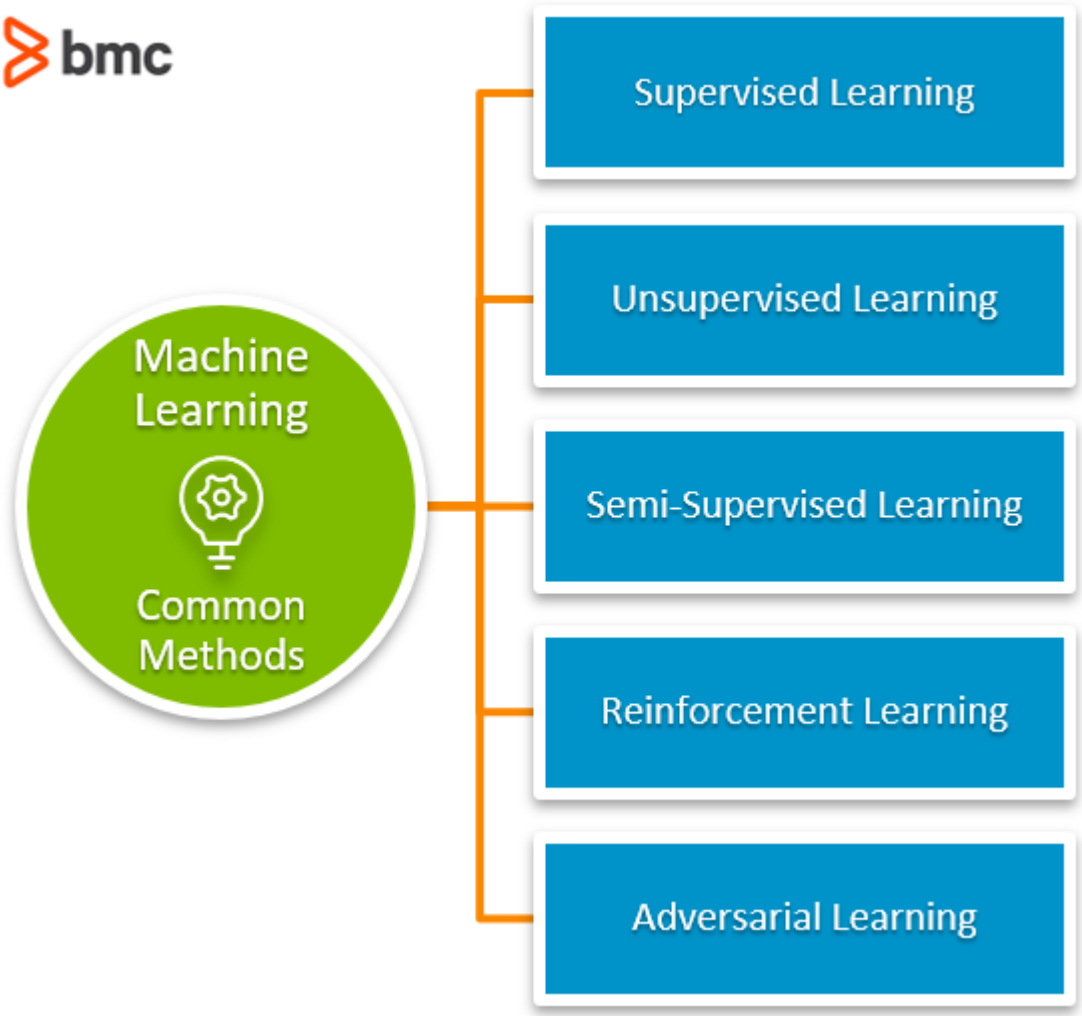
(Explore anomaly detection with ML.)

Supervised vs unsupervised learning algorithms

By now, we can say that the main difference between these two categories of algorithms lies in the labeling of the training data.

But you'll also need to consider other factors when building a machine learning pipeline, such as:

- **Using unsupervised methods on labeled data.** Doing so can identify hidden traits as a part of Exploratory Data Analysis (EDA).
- **Using supervised and unsupervised algorithms together.** For example, you can use unsupervised learning algorithms to reduce the dimensionality of the labeled data, and then proceed with supervised leaning algorithms.
- **Using manual effort to label data.** If the data is unlabeled and the use case desires highly accurate classification into specially defined categories, then you can apply manual effort to add labels to the input data. Often, this process is time consuming and expensive, but it's always available for situations that demand it.



What is semi-supervised learning?

Machine learning algorithms yearn for data. The more, the merrier. These days we are generating data at an astronomical rate. So far so good, but only a very small portion of this data is labeled. So, while unsupervised learning methods can identify patterns and create clusters, they cannot be guided to either:

- Create custom/or specific clusters
- Attain a certain accuracy threshold

Luckily, we don't always have to choose between large datasets *or* the customization and accuracy of supervised learning.

A new class of algorithms—semi-supervised algorithm—can learn from partially labeled datasets. These algorithms are especially useful in:

- Natural language processing (NLP) tasks like classifying text
- The field of medicine, like for protein sequence classification

These are both areas where we have loads of data, but it's mostly unlabeled. So, here's how semi-supervised machine learning works:

1. We add labels to a fraction of the data either algorithmically or via human labor.
2. Next, we use unsupervised learning algorithms to create clusters of similar data points.

3. Then we use that labeled data to further train the rest of the unlabeled data.

Semi-supervised ML is like an explorer getting the lay of the land upon arrival, and then seeking the help of locals to get to know individual areas more intimately.

One example of semi-supervised learning algorithm based system is Google's Expander, a technology that Google uses in many of its products including Gmail and Google Photos.

(See the similarities with human-in-the-loop ML.)

The path ahead

In this article we have covered a lot of ground on two major types of machine learning algorithms, finally arriving at semi-supervised learning, which epitomizes the notion of: why choose when you can have both?

Importantly, we are not limited to just these methods. Through continuous evolving, new methods are emerging to tackle new problems.

Reinforcement learning

For example, another class of machine learning approach is reinforcement learning. Instead of relying on labeled or unlabeled data, reinforcement learning employs the concept of rewards and penalties to make the machine learning model solve a problem on its own.

This learning setup is almost autonomous, with human intervention limited to altering the environment and tweaking rewards/penalties. Reinforcement learning finds wide application in building autonomous systems be it a self-driving car or video game playing bot.

Adversarial learning

As machine learning models are becoming mainstream, the threat of attack and hacking attempts is increasing. It's this scenario that adversarial learning, a sub-class of supervised learning, comes to the rescue.

Adversarial learning is widely used in making machine learning models robust and immune to attacks.

They say modern problems need modern solutions. So, as we are evolving in AI and ML, we find ourselves tackling new challenges. And the machine learning methods are also evolving to keep pace.

Related reading

- BMC Machine Learning & Big Data Blog
- Top Machine Learning Frameworks To Use Today
- Structured vs Unstructured Data: A Shift in Privacy
- What's a Deep Neural Network? Deep Nets Explained
- Artificial Artificial Intelligence & The Reality of AAI
- Data Ethics for Companies