A Guide to Machine Learning

Supervised or unsupervised; leveraging computer algorithms to discover the best method for finding malicious content.

BLUVECTOR®

CYBER THREAT HUNTING PLATFORM

Two Ways to Learn

Machine learning is the process of computer programs becoming more accurate at a task due to exposure to data, called training instances. What task is being performed and what type of data is provided for training is critical to deciding how and what techniques to use in the machine learning process.

The most important aspect of the data is whether it is "labeled". Labeled means that someone has assigned a category of interest to each training instance. The best label is dependent on what task the machine learning program is seeking to accomplish. For example, if one wants a program to distinguish between cats and dogs labels of "cat" and "dog" are sufficient. If, however, one wishes to distinguish between *breeds* of cat or dog labels such as; Pug, Dalmatian, American Shorthair, Siamese, Collie, German Shepard are required (See Figure 1).

If, however, we instead want to distinguish between young and old animals we would use labels related to age; newborn, adolescent, adult, senior. The labels, the training instances, and the desired task are inextricably linked.

Labeling may seem trivial for something as familiar as dogs and cats but in general it can be a difficult, expensive and time consuming process to attain enough training instances of each label to produce highly accurate machine learning models.

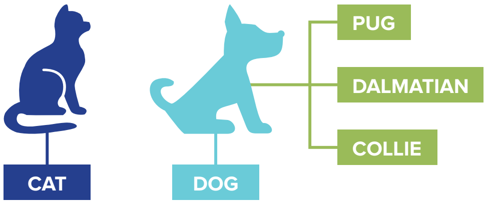

FIGURE1. The difference in specificities of labels.

Subject matter experts are used to manually look at every training instance to determine its label. The number of training instances required can be in the tens or hundreds of thousands. Some labeling, such as determining if something is normal or abnormal (a task known as anomaly detection), can even be dependent on the environment in which the determination is made. For example, a cat at a dog show may be labeled as *abnormal* while a cat at an animal shelter may be considered *normal*.

Supervised Learning

Since the existence or absence of labeled training data is so fundamental to the machine learning process, special terms are used to describe the two scenarios[1]:

- When labeled training instances are used for learning the process is said to be **supervised**
- If no label instances are used, the process is said to be **unsupervised**.

One way to think about supervised learning is that during the learning process, a "teacher" is available that will tell the algorithm when it is **predicting** labels correctly and when it is making mistakes. That teacher is the subject matter expert or experts who labeled all the training instances used by the machine learning algorithm. The machine uses the teacher to make more accurate predictions. In Figure 2 below, the teacher directs the machine to learn to predict whether a sample has a "hole" and label each sample with its prediction.

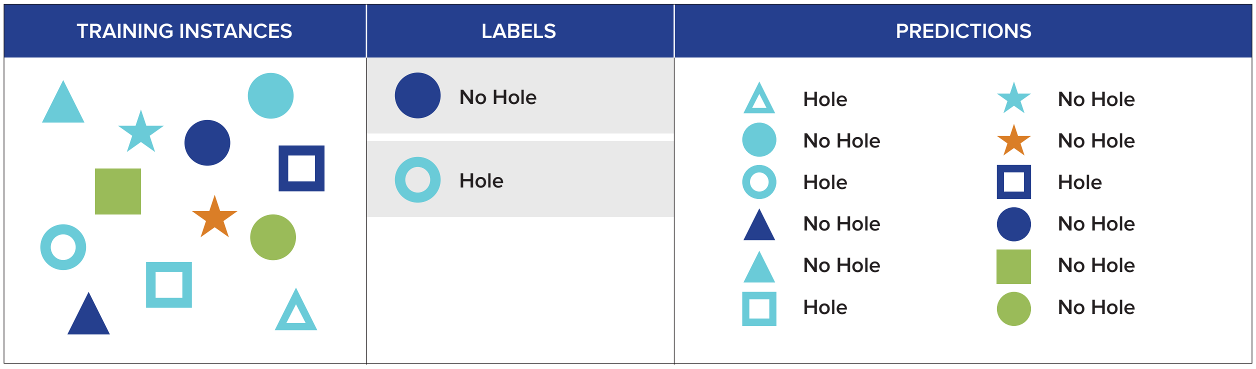

FIGURE 2. Example of how an unsupervised learning algorithm could sort training instances.

[1] *In addition to supervised and unsupervised learning, there are two other categories of machine learning that are much less common. First, semi-supervised learning is a hybrid approach in which relatively few training instances are labeled but most are not. Second, reinforcement learning is a variant of supervised learning concerned with teaching software agents how to interacting with complex environments by providing rewards and penalties over time (it's akin to learning by trial-and-error).*

Unsupervised Learning

When no labels are present a learning algorithm's job is to devise a method for sorting the training instances. In unsupervised learning, the algorithm is left to draw a reasonable conclusion given the training instances. There is no "right" conclusion but rather many reasonable ones. How the training instances are sorted depends on :

1. The features used to describe the instances to the algorithm.
2. The parameters of the sort such as how many groups are desired, minimum group size or density, whether the sorting should be hierarchal.

Figure 3 below shows how a learning algorithm may sort some training instances of colored shaped similar to those shown in Figure 2. Three different outputs/models are shown illustrating how multiple reasonable groupings can be formed from even this trivial example. One major drawback of unsupervised learning is there is no way to predict or generate descriptions of groups the algorithm forms. So, a user may know there is some similarity between instances placed into the same group or cluster but they may not understand why or what makes them similar.

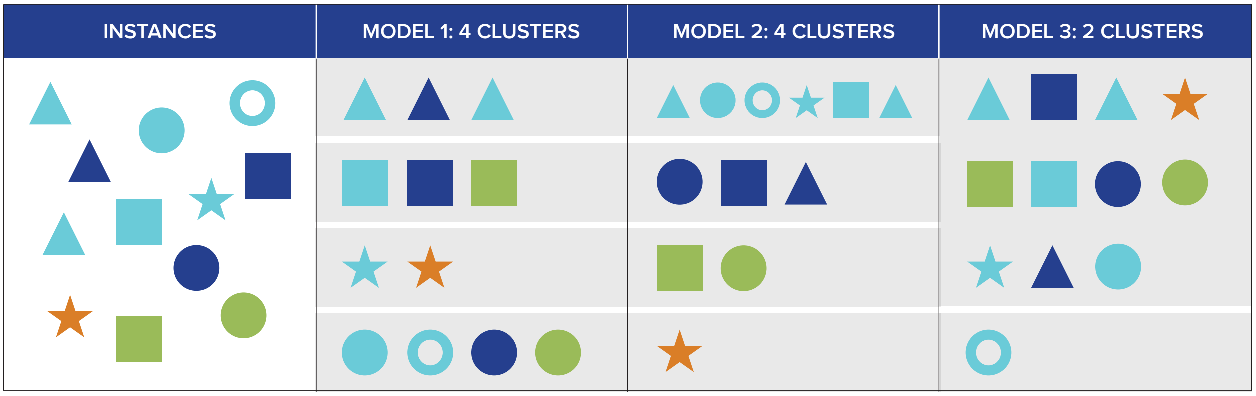

FIGURE 3. Example of how an unsupervised learning algorithm could sort training instances.

The Difference a Good Teacher Makes

When possible, using labels in the learning process makes understanding the results of the learning easier and provides the user with a better experience. Accuracy can be measured, something not possible in unsupervised learning. Since accuracy can be measured, the algorithms themselves can be designed to maximize various competing aspects of accuracy, such as precision and recall or false positive and false negative rates. Not only can accuracy be measured, confidence in the accuracy of any predictions can also be quantified. The model answers a **clearly** formulated question, typically "How should this sample be labeled, given the set of labels used for the training instances?" or "For this sample, what is the value of a given unmeasured feature?" There is no doubt as to what the results mean; the learner is simply trying to model the behavior of the subject matter experts in labeling unknown samples.

BluVector: Cybersecurity and Machine Learning

BluVector is a leader in applying machine learning techniques to the hardest challenges in cybersecurity. BluVector's supervised machine learning technology is based on over a decade of research and development in understanding how to learn from vast collections of both benign and malicious software. Our supervised machine learning models leverage both benign and malicious characteristics to predict whether software and application files pose a threat to an enterprise. By including expertly labeled, benign and malicious samples in our training we have a robust capability even against previously undiscovered or zero-day malware.

Furthermore, BluVector never stops learning. As an enterprise's analysts work with the system BluVector leverages their input to tailor and improve performance.

Whether supervised or unsupervised, machine learning approaches to cybersecurity challenge users to find the unknown and previously unseen. BluVector's ability to interoperate and enhance existing security tools and assist users in hunting these difficult threats allow users to save time, money and reduce risk of potentially costly cyber breaches.

BLUVECTOR®

CYBER THREAT HUNTING PLATFORM

www.BluVectorcyber.com • info@bluvectorcyber.com • 1.855.672.4258