TOP 5 MACHINE LEARNING ALGORITHMS FOR BEGINNERS

Machine learning is a major component in the race towards artificial intelligence. Whether you're seeking true artificial intelligence or simply trying to gain insight from all the data you've been collecting, machine learning is a major step forward. But where to get started?

If you're a beginner, machine learning can feel overwhelming – how to choose which algorithms to use, from the seemingly infinite options, and how to know just which one will provide the right predictions (data outputs). These top 5 machine learning algorithms for beginners offer a fine balance of ease, lower computational power, and immediate, accurate results.

How machine learning works

At its most basic, machine learning is a way for computers to run various algorithms without direct human oversight in order to learn from data. Because we don't know the function – the look or form that caused the data – we have to learn it from these algorithms.

Machine learning can include running any variety of tasks in order for the machine to determine a high-probability outcome for various information, such as the functions between input and output or the hidden structures in unlabeled data. In instance-based learning, the machine can produce class labels by comparing previous instances.

These ways of learning are summed up in the three types of machine learning algorithms:

Supervised learning. These methods rely on labeled training data sets to learn a function

between input variables (X) and output variables (Y). The most common types include:

- Classification methods, which predict the output of a given data sample when the output variable is categorical, for instance dead or alive, sick or healthy.
- Regression methods, which predict output variables that are real values, such as the age of a person or the amount of snowfall.
- Ensemble methods, which combine predictions from weaker algorithmic output to predict new output.

Unsupervised learning. These methods use only input variables (X), not output variables, and rely on unlabeled training data sets to map the underlying structure of the data. Common examples include:

- Association methods, which uncover probability of items in a collection, as in market-basket analysis.
- Clustering methods, which group samples of objects based on similarity.

Reinforcement learning. These methods allow the user or other designated agent to decide the best next action, based on the current state and learned behaviors that maximize the rewards. This approach is often used in robotics.

Explore Supervised vs Unsupervised Machine Learning for AIOps.

Machine learning algorithms for beginners

Deciding just which algorithm to use is tricky, more often an art than a science. That's because your data set can vary widely, in size, quality, and nature. Choosing an algorithm is also limited to your access to computational power, the urgency of your task, and the ultimate goal.

Experts caution that the more options exist, the harder it is to choose. Even top data scientists don't know ahead of time exactly which algorithm will provide exactly the predictions needed. Instead, machine learning is about digging in, experimenting, and seeing what works for the problem you need to solve.

Without further ado and in no particular order, here are the top 5 machine learning algorithms for those just getting started:

Linear regression

Despite its name, linear regression is a classification method, not a regression method. This predictive modeling approach is very well understood, as statistics has been using this tool for decades before the invention of the modern computer.

The goal of linear regression is to make to most accurate predictions possible by finding the values for two coefficients that weight each input variable. These techniques can include linear algebra, gradient descent optimization, and more.

Employing linear regression is easy and generally provides very accurate results. More experienced users know to remove variables from your training data set that is closely correlated and to remove as much noise (unrelated output variables) as possible.

Logical regression

Similar to linear regression, logical regression is another statistical, well-understood method for classification which finds the values for two coefficients that weight each input variable. The difference is that this solves problems of binary classification, relying on a logical, non-linear function instead. Therefore, logical regression determines whether a data instance belongs to one class or another and can also provide the reason behind the prediction, unlike linear regression.

When using this algorithm, limiting correlating data and removing noise is also important.

Classification and regression trees

Sometimes known as CART, classification and regression trees, are a simple form of decision trees, wherein the modeled tree is binary, using only algorithms and data structures.

There are two types of nodes on this tree:

- Branch nodes, which represent a single input variable and offer a single split point on the variable (assuming its numeric).
- Leaf nodes, which represent the two output variables.

When the machine runs the algorithm, the prediction plays out by following the branch node splits until reaching a leaf node, which is the prediction, or class value output.

Classification and regression trees are easy to learn and use, and accurate for a whole range of problems. These are especially speedy to implement, as the data requires no special preparation.

K-nearest neighbor (KNN)

KNN is short for the K-nearest neighbor method, in which the user specifies the value of K. Unlike previous algorithms, this one trains on the entire dataset.

The goal of KNN is to predict an outcome for a new data instance. The algorithm trains the machine to check the entire dataset to find the k-nearest instances to this new data instance or to find the k-number of instances that are most similar to the new instance. The prediction, or output, is one of two things:

- The mode or most frequent class, in a classification problem
- The mean of the outcomes, in a regression problem

This algorithm usually employs methods to determine proximity such as Euclidean distance and Hamming distance.

The pros of KNN are that its simplicity and ease of use. Though it can require a lot of memory to store large datasets, it only calculates (learns) at the moment a prediction is needed.

When using a high number of input variables, the machine-learned understanding of “closeness” can be compromised. This situation, known as the curse of dimensionality, can be avoided by limiting your input variables to only those that are most relevant to predicting the output.

Naïve Bayes

Like other beginner algorithms, the Naïve Bayes algorithm is a classifier that uses training data in

simple manner with powerful outputs.

Naïve Bayes employs the Bayes Theorem of Probability to classify content. The Bayes Theorem calculates probability of an event occurring or a hypothesis being true based on prior knowledge, making the model able to handle two types of probabilities:

- Determining class
- Determining a conditional probability of each class, provided X value

Importantly, the "naïve" part of the title is based on the algorithm's assumption that the variables are independent of each other, which is often unrealistic in real-world examples.

The content best suited for this Naïve Bayes is often language-based, such as web pages and articles, plus smaller bodies of text, such as tweets or metadata from blogs. This algorithm is a go-to option when trying to rank content or classify data based on categories (content themes). This algorithm has also been used effectively in predicting disease development and location, as well as analyzing human sentiment.

Employing these five machine learning algorithms may not be complicated, but they do take time to master. The time for beginners is time well spent, as these are important building blocks for further machine learning experimentation.