The Basics of Machine Learning

Alexander B. Chinn

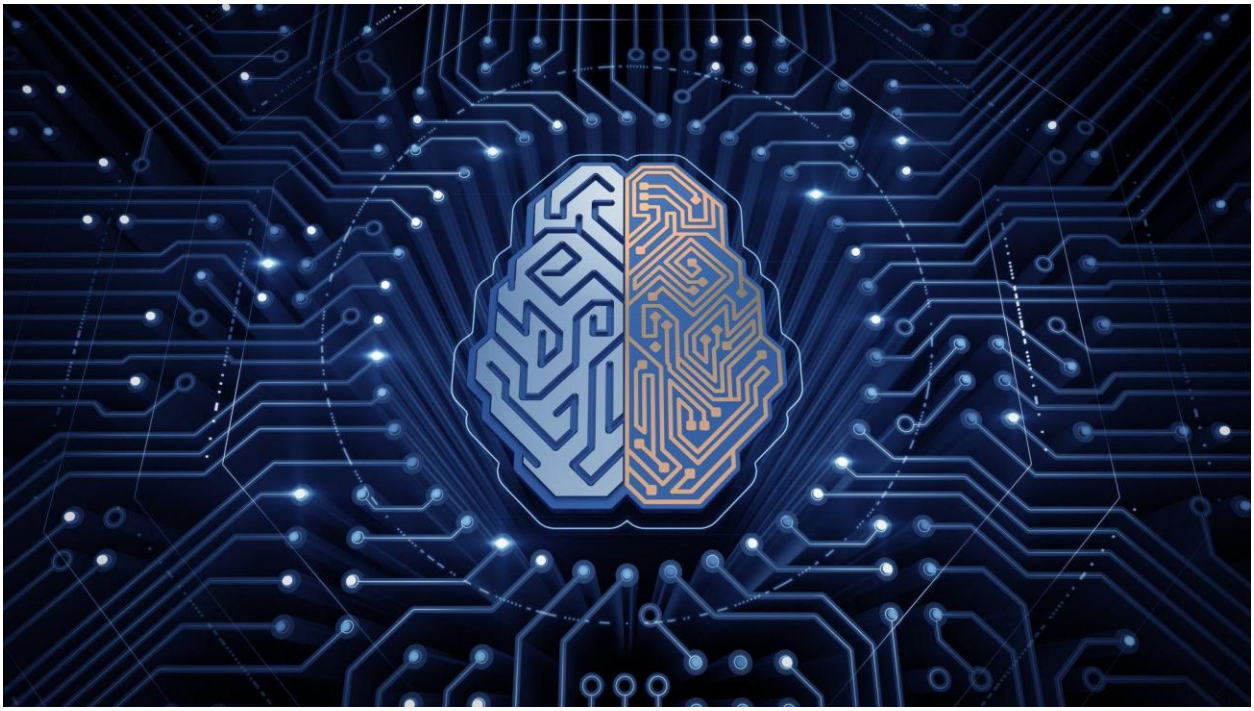
Missouri University of Science and Technology

1120: Exposition of Argumentation

Suzanne Withem

10/15/2023

The Basics of Machine Learning



In the world of today there are now countless applications that will all fall under the term of AI. You may not even realize quite how prevalent it is in our world. AI can be found in the fairly obvious places, chat bots, Image generators, and voice deepfakes. If you take the time to look more closely you can find examples of AI in far more niece places such as spell check, advertisement choices, and browsers. While it may be used in so many different places for so many different things you may be left wondering what exactly is AI, how does it work, and can I make my own? Today I will go over a subset of AI, Machine learning. Machine learning or ML is the branch of AI that focuses on developing algorithms and models that allow computers to learn from data and improve from previous experiences. ML is generally only used to make prediction models for any number of things. There are four main types of machine learning, each with their own special characteristics and applications, Supervised Machine Learning, Unsupervised Machine Learning, Semi-Supervised Machine Learning, and Reinforcement Learning.

Supervised learning is the process in which a model gets trained on what is called a "Labelled Dataset". Labeled datasets contain both input and output values. Supervised Learning algorithms will learn how to map points between an input and its correct output. This algorithm is very good at things such as image recognition, data extraction, speech recognition, and predictive analytics. For example, if you want to determine if a picture contains a car inside of it you could feed a dataset of car pictures into a Supervised Learning Algorithm. The algorithm will begin to learn how to tell if a picture contains a car. Eventually you would want to try your

algorithm against pictures it has never seen again to test that it works. Supervised Learning has two main advantages, it can be incredibly accurate, and they are very easy to setup. However, because of how this algorithm is trained it can take a very large amount of time to create the labeled data sets you need to train on. This algorithm will also likely struggle when introduced to patterns it has never seen.

Unsupervised Machine Learning is reasonably similar to supervised learning in that it is still solely based on patterns but with one major difference. The data set it trains on does not have labels. This algorithm is very good at discovering hidden patterns within data. One major application of this machine learning is in customer predictions. Stores such as Walmart will track what you purchase and use this type of algorithm to figure out patterns in your shopping habits and will use this data to attempt to sell you more stuff. Using such an algorithm, stores can find things such as when something sells best, who's buying that product, and what people normally buy with that product. Two major advantages this model holds are its ability to find very unusual patterns and the fact that it doesn't need labeled outputs in its data set. This can make the process of gathering data much faster. However, due to the lack of labels there is a high chance of the quality of the models' output being poor.

Semi-Supervised Machine Learning is essentially the child of an Unsupervised and Supervised algorithm. It can take a data set in which some things are labeled, and others are not. For instance, the inputs can be labeled, and the outputs not labeled. This is very useful due to the fact that obtaining labeled data can often be very expensive, money-wise or time intensive. Consider the following example. You are building a translator, and creating labeled translation pairs for sentences can be very difficult to make. You can use a Semi-Supervised algorithm to use a database with only a few labeled sentence pairs and still come out with a very competent translator. A semi-Supervised algorithm is a lot more versatile than other models once trained and doesn't have to be used on very specific inputs. Two major disadvantages of Semi Supervised are that these models are quite complicated to implement compared to the other major methods, and these still require at least partially labeled data sets which can be very difficult to acquire.

Reinforcement machine learning is a learning method in which an algorithm is learning by trial and error. Using this method your machine learns to determine the ideal behavior to use to maximize its performance of a certain task. This particular type of machine learning is crucial in areas where there can be lots of unusual environments to make decisions in. Consider the following example. You're making an Ai for your favorite game, checkers. Should you make this checkers AI a type of Reinforcement model, its steps toward playing the game go as follows. The AI will look at the state of the game board and then simply try every move combination possible out to a certain point. It will then play whatever move will start it down the path that will achieve the best outcome according to whatever parameters you've set it. There are two major advantages of Reinforcement machine learning. These algorithms have autonomous decision-making abilities that are well adapted for tasks that require a sequence of decisions. They are also very good at achieving very unconventional solutions to problems that have no conventional

solutions. However, these algorithms can be very expensive computationally and time wise. These algorithms have very long training times and require immense amounts of data to train effectively. Reinforcement learning is not very useful for solving very simple problems.

Now that you know about four different types of machine learning algorithms it may be useful to learn the basic steps on how to train one.

Gathering Data:

For the first step of this process, you will want to identify at least one source of data to train your new AI on. You will want this to be high quality and quantity data. The more data you get the more accurate your AI can be. Your project will never get off the ground if you cannot complete this first crucial step as the type of data you may find will heavily influence your decisions here on out.

Data Preparation:

If you have found multiple data sets, you will now want to push them all together into one data set and then randomize the order of the data. You will now have to convert all this data into one specific format. This will normally depend on what type of algorithm you plan on using or it can also depend on what types of data you have acquired and the limitation on transforming that dataset. Having converted your data, you will now want to clean up any duplicate pieces of data, any missing values in the data, and removing all invalid data. This will ensure that no unnecessary data is left in your data set to unnecessarily increase your AI's training time.

Data Analysis:

You will now have to select what type of analytical technique you will use in your AI. Then the fun step, building the actual AI. If your interested in how this working from a coding perspective I recommend checking out Microsoft Learn which will help you more from a coding standpoint. The real goal of this step is to make a model that can analyze your data using whatever method of data analysis you've selected. W3schools has a great page on various types of data analysis if you're interested in more specific details on this point.

Train the model:

Having now made an AI model and having the data to train it on the next logical step is in fact to train your AI. This step is typically by far the longest of these steps. The process of training can easily take upwards of several days depending on how accurate you need your AI to be and how large of a data set you've chosen.

Test the model:

Having now trained your model you must now test it. Testing is always done on a different dataset than the one you trained you Ai on. This is done to see how well it does on something it has never seen before and to test how well it is likely to do in the real-world solution you are developing it for. Should the model not perform adequately you will have to go back to either the training step and retrain the AI or you may have to go back even further to the original gathering of data and get more or higher quality data.

Deployment:

If you are lucky enough to have you model preform well enough you are now ready to deploy it. You can now use it to solve whatever problem you may have once had. Hopefully it works in whatever application you try to use it in. Troubleshooting issues after deployment can be a nightmare due to there being so many places for a problem to have gone unnoticed. Which is why it is important at all stages of this to make sure you put the maximum amount of effort into your AI so that you do not have to spend twice that trying to solve a problem with it.

And there you have it. I hope now your interest in this subject is at least a little satisfied, but I hope a enough of your curiosity remains to go out and learn more about this topic. There is so much to talk about even on this specific part of AI that I could never hope to even scratch the surface of this topic let alone delve deep into its most complicated topics. Should you wish to continue your journey into AI there are a multitude of websites, blogs, and books that are completely free out there that can take a deeper dive into whatever you've still got questions on.

References

Theobald, O. (2021). *Machine Learning for Absolute Beginners*.

Ethem Alpaydin. (2021). *Machine learning*. The Mit Press.

IBM. (2023). *What is Machine Learning?* Www.ibm.com; IBM. https://www.ibm.com/topics/machine-learning