Yufeng
G
Applying machine learning to the world. Developer and Advocate for @googlecloud. Runner, chef, musician. Opinions are solely my own.
Aug 31, 2017 · 9 min read

The 7 Steps of Machine Learning

From detecting skin cancer, to sorting cucumbers, to detecting escalators in need of repairs, machine learning has granted computer systems entirely new abilities.

But how does it really work under the hood? Let's walk through a basic example, and use it as an excuse talk about the process of getting answers from your data using machine learning.

Let's pretend that we've been asked to create a system that answers the question of whether a drink is wine or beer. This question answering system that we build is called a "model", and this model is created via a process called "training". The goal of training is to create an accurate model that answers our questions correctly most of the time. But in order to train a model, we need to collect data to train on. This is where we begin.

If you are new to machine learning and want a quick overview first, check out this article before continuing:

This is just the beginning

Wine or Beer?

Our data will be collected from glasses of wine and beer. There are many aspects of the drinks that we *could* collect data on, everything from the amount of foam, to the shape of the glass.

For our purposes, we'll pick just two simple ones: The color (as a wavelength of light) and the alcohol content (as a percentage). The hope is that we can split our two types of drinks along these two factors alone. We'll call these our "*features*" from now on: color, and alcohol.

The first step to our process will be to run out to the local grocery store and buy up a bunch of different beers and wine, as well as get some equipment to do our measurements—a spectrometer for measuring the color, and a hydrometer to measure the alcohol content. Our grocery store has an electronics hardware section :)

Gathering Data

Once we have our equipment and booze, it's time for our first real step of machine learning: **gathering data**. This step is very important because the quality and quantity of data that you gather will directly determine how good your predictive model can be. In this case, the data we collect will be the color and the alcohol content of each drink.

Color (nm)	Alcohol %	Beer or Wine?
610	5	Beer
599	13	Wine
693	14	Wine

This will yield a table of color, alcohol%, and whether it's beer or wine. This will be our **training data.**

Data preparation

A few hours of measurements later, we have gathered our training data. Now it's time for the next step of machine learning: **Data preparation**, where we load our data into a suitable place and prepare it for use in our machine learning training.

We'll first put all our data together, and then randomize the ordering. We don't want the order of our data to affect what we learn, since that's not part of determining whether a drink is beer or wine. In other words, we make a determination of what a drink is, independent of what drink came before or after it.

This is also a good time to do any pertinent visualizations of your data, to help you see if there are any relevant relationships between different variables you can take advantage of, as well as show you if there are any data imbalances. For example, if we collected way more data points about beer than wine, the model we train will be biased toward guessing that virtually everything that it sees is beer, since it would be right most of the time. However, in the real-world, the model may see beer and wine an equal amount, which would mean that guessing "beer" would be wrong half the time.

We'll also need to split the data in two parts. The first part, used in training our model, will be the majority of the dataset. The second part will be used for evaluating our trained model's performance. We don't want to use the same data that the model was trained on for evaluation, since it could then just memorize the "questions", just as you wouldn't use the same questions from your math homework on the exam.

Sometimes the data we collect needs other forms of adjusting and manipulation. Things like de-duping, normalization, error correction, and more. These would all happen at the data preparation step. In our case, we don't have any further data preparation needs, so let's move forward.

Choosing a model

The next step in our workflow is choosing a model. There are many models that researchers and data scientists have created over the years. Some are very well suited for image data, others for sequences (like text, or music), some for numerical data, others for text-based data. In our case, since we only have 2 features, color and alcohol%, we can use a small linear model, which is a fairly simple one that should get the job done.

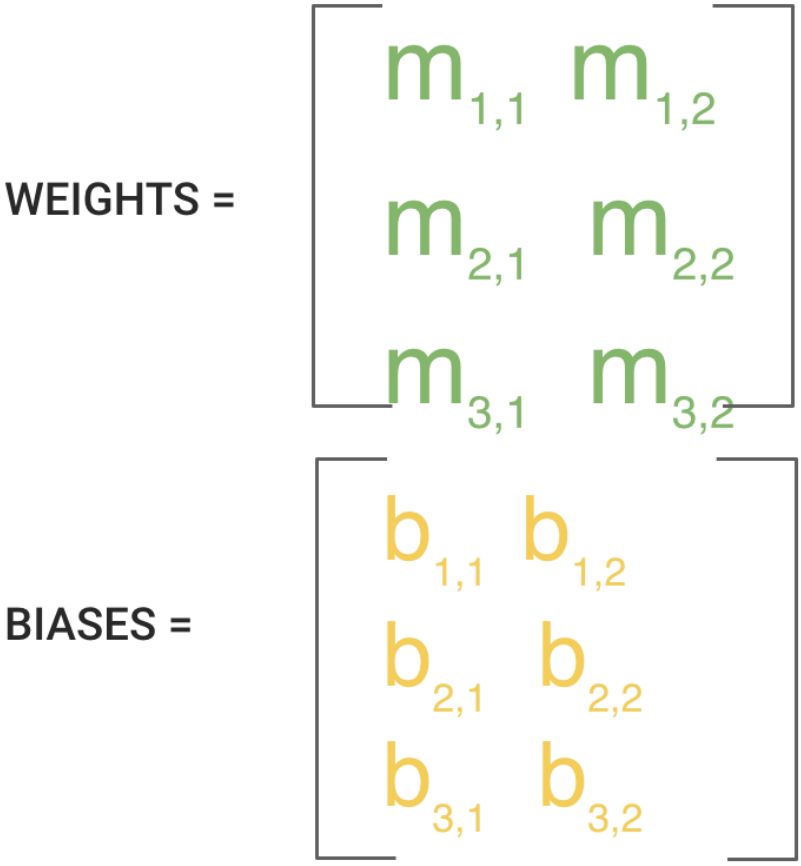
Training

Now we move onto what is often considered the bulk of machine learning—the **training**. In this step, we will use our data to incrementally improve our model's ability to predict whether a given drink is wine or beer.

In some ways, this is similar to someone first learning to drive. At first, they don't know how any of the pedals, knobs, and switches work, or when any of them should be used. However, after lots of practice and correcting for their mistakes, a licensed driver emerges. Moreover, after a year of driving, they've become quite adept. The act of driving and reacting to real-world data has adapted their driving abilities, honing their skills.

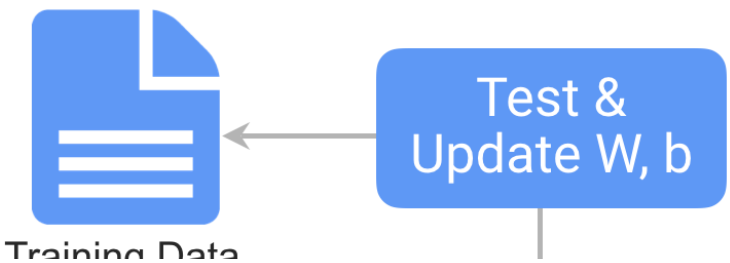
$$y = m * x + b$$

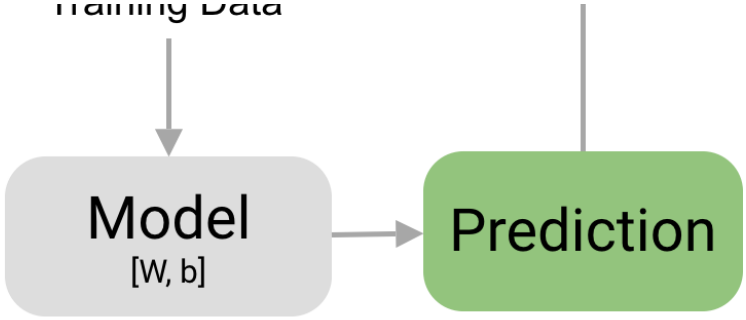
We will do this on a much smaller scale with our drinks. In particular, the formula for a straight line is y=m*x+b, where x is the input, m is the slope of that line, b is the y-intercept, and y is the value of the line at the position x. The values we have available to us for adjusting, or "training", are m and b. There is no other way to affect the position of the line, since the only other variables are x, our input, and y, our output.



In machine learning, there are many m's since there may be many features. The collection of these m values is usually formed into a matrix, that we will denote W, for the "weights" matrix. Similarly for b, we arrange them together and call that the biases.

The training process involves initializing some random values for W and b and attempting to predict the output with those values. As you might imagine, it does pretty poorly. But we can compare our model's predictions with the output that it should produced, and adjust the values in W and b such that we will have more correct predictions.





This process then repeats. Each iteration or cycle of updating the weights and biases is called one training “step”.

Let’s look at what that means in this case, more concretely, for our dataset. When we first start the training, it’s like we drew a random line through the data. Then as each step of the training progresses, the line moves, step by step, closer to an ideal separation of the wine and beer.

Evaluation

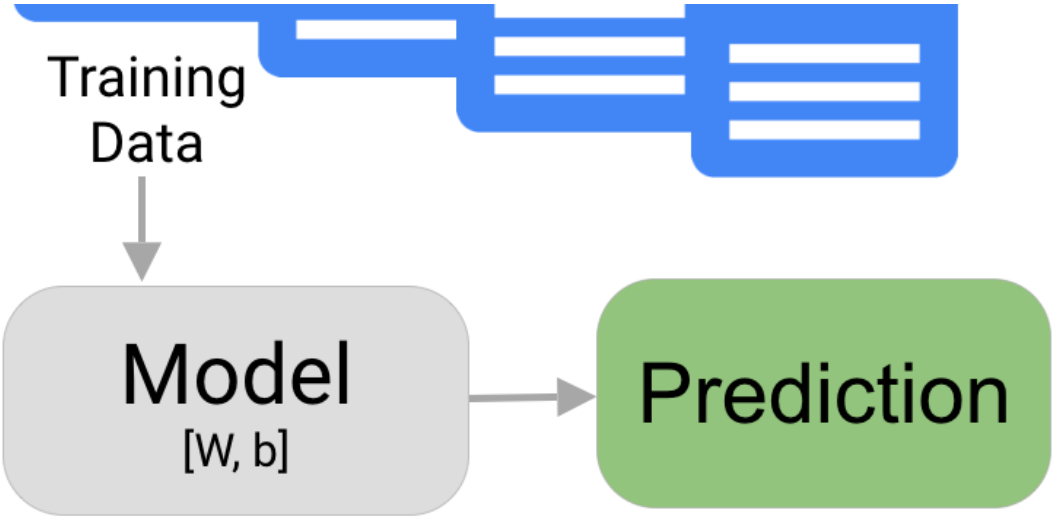
Once training is complete, it’s time to see if the model is any good, using **Evaluation**. This is where that dataset that we set aside earlier comes into play. Evaluation allows us to test our model against data that has never been used for training. This metric allows us to see how the model might perform against data that it has not yet seen. This is meant to be representative of how the model might perform in the real world.

A good rule of thumb I use for a training-evaluation split somewhere on the order of 80/20 or 70/30. Much of this depends on the size of the original source dataset. If you have a lot of data, perhaps you don’t need as big of a fraction for the evaluation dataset.

Parameter Tuning

Once you’ve done evaluation, it’s possible that you want to see if you can further improve your training in any way. We can do this by **tuning our parameters**. There were a few parameters we implicitly assumed when we did our training, and now is a good time to go back and test those assumptions and try other values.

One example is how many times we run through the training dataset during training. What I mean by that is we can “show” the model our full dataset multiple times, rather than just once. This can sometimes lead to higher accuracies.



Another parameter is "*learning rate*". This defines how far we shift the line during each step, based on the information from the previous training step. These values all play a role in how accurate our model can become, and how long the training takes.

For more complex models, initial conditions can play a significant role in determining the outcome of training. Differences can be seen depending on whether a model starts off training with values initialized to zeroes versus some distribution of values, which leads to the question of which distribution to use.

The potentially long journey of parameter tuning

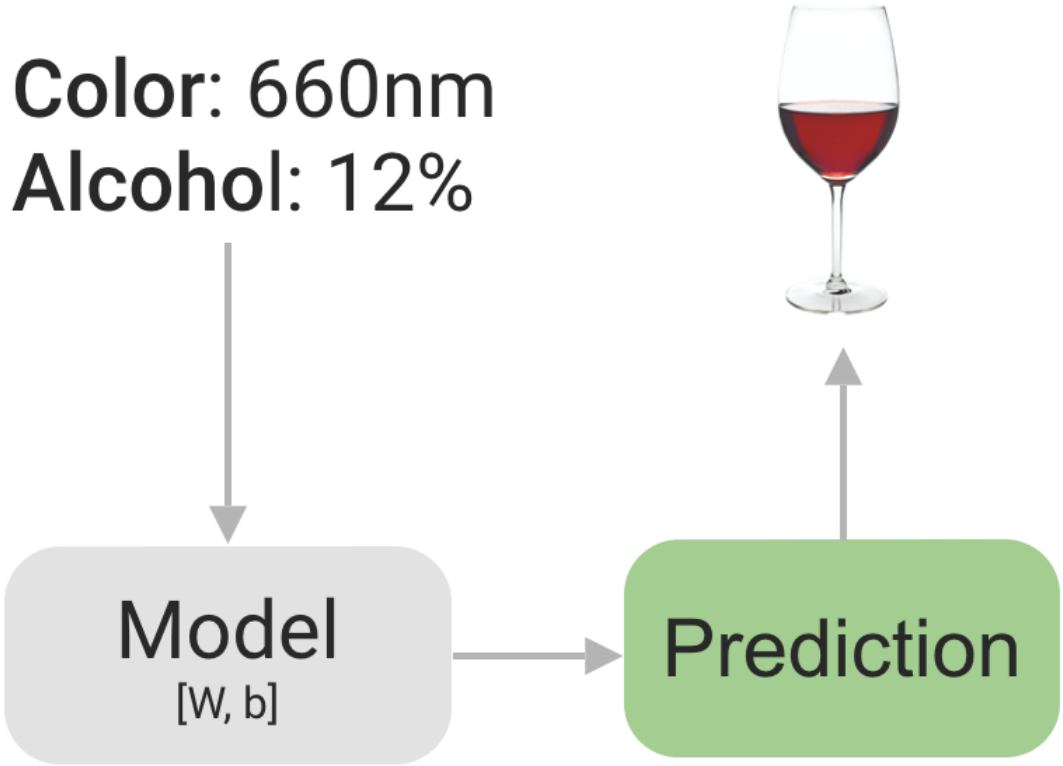
As you can see there are many considerations at this phase of training, and it's important that you define what makes a model "good enough", otherwise you might find yourself tweaking parameters for a very long time.

These parameters are typically referred to as "*hyperparameters*". The adjustment, or tuning, of these hyperparameters, remains a bit of an art, and is more of an experimental process that heavily depends on the specifics of your dataset, model, and training process.

Once you're happy with your training and hyperparameters, guided by the evaluation step, it's time to finally use your model to do something useful!

Prediction

Machine learning is using data to answer questions. So **Prediction**, or inference, is the step where we get to answer some questions. This is the point of all this work, where the value of machine learning is realized.



We can finally use our model to predict whether a given drink is wine or beer, given its color and alcohol percentage.

The big picture

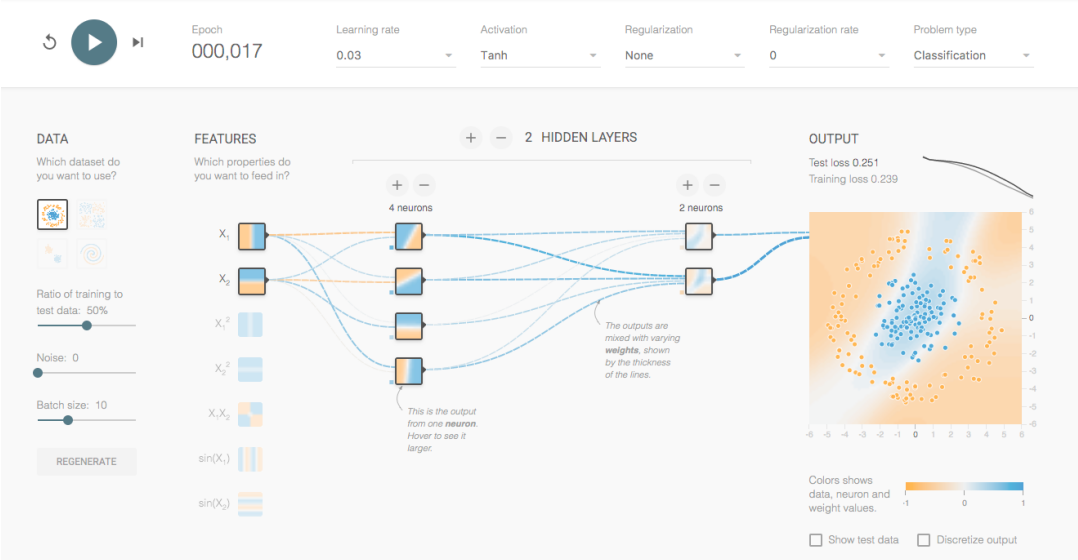
The power of machine learning is that we were able to determine how to differentiate between wine and beer using our model, rather than using human judgement and manual rules. You can extrapolate the ideas presented today to other problem domains

as well, where the same principles apply:

- Gathering data
- Preparing that data
- Choosing a model
- Training
- Evaluation
- Hyperparameter tuning
- Prediction.

TensorFlow Playground

For more ways to play with training and parameters, check out the TensorFlow Playground. It's a completely browser-based machine learning sandbox where you can try different parameters and run training against mock datasets.



What's next?

While we will encounter more steps and nuances in the future, this serves as a good foundational framework to help think through the problem, giving us a common language to talk about each step, and go deeper in the future.

Next time, we will build our first "real" machine learning model, using code. No more drawing lines and going over algebra!

Machine Learning | TensorFlow | Big Data | Artificial Intelligence | Towards Data Science