MACHINE LEARNING IN 5 DAYS

DATA SCIENCE CURRICULUM PART 2

COMPILED BY STUDY GROUP AFRICA ON GITHUB

Table of Contents

DAY 1 DATA SCIENCE & ML

WHAT IS DATA SCIENCE?

https://youtu.be/vN5uZZ1h7VE

ML BY GOOGLE CLOUD TECH ON YOUTUBE

1. **Gathering Data 1:49 |**
2. **Preparing Data 2:21 |**
3. **Model Selection 4:03 |**
4. **Training 4:30 |**
5. **Evaluation 6:46 |**
6. **Parameter Tuning 7:24 |**
7. **Prediction 8:55 |**

TRANSCRIPT
Intro
From detecting skin cancer to sorting cucumbers to detecting escalators in need of repair, 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 to talk about the process of getting answers from your data using machine learning.

Welcome to Cloud AI Adventures.

My name is Yufeng Guo.

On this show, we'll explore the art, science, and tools of 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.

- In machine learning, the goal of training is to create an accurate model that answers our questions correctly most of the time. But in order to train the model, we need to collect data to train on. This is where we will begin.

  - Our data will be collected from glasses of wine and beer. There are many aspects of drinks that we could collect data on everything from the amount of foam to the shape of the glass.
  - But for our purposes, we'll just pick two simple ones (i)the color as a wavelength of light and (ii)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, buy up a bunch of different drinks,and get some equipment to do our measurements-- a spectrometer for measuring the color and a hydrometer to measure the alcohol content. It appears that our grocery store has an electronics hardware section as well.
  Once our equipment and then booze-- we got it all set up--

STEP 1 Gathering Data

it's time for our first real step of machinelearning-- gathering that 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 alcohol content of each drink.This will yield us a table of color, alcohol content, and whether it's beer or wine. This will be our training data.

So a few hours of measurements later, we've gathered our training data and had a few drinks, perhaps.And now it's time for our next step of machine learning--data preparation--

STEP 2 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 then randomize the ordering.

- We wouldn't want the order of our data to affect how we learn since that's not part of determining whether a drink is beer or wine.
  In other words, we want to make a determination of what a drink is independent of what drink came before or after it in the sequence.
  - This is also a good time to do any pertinent visualizations of your data, helping you see if thereis any relevant relationships between different variables as well as show you if there are any data imbalances.

For instance, if we collected way more data points about beer than wine, the model we train will be heavily 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 in equal amount, which would mean that it would be guessing beer wrong half the time.

We also need to split the data into two parts.

    - The first part used in training our model will be the majority of our dataset.
- The second part will be used for evaluating our train model's performance.

We don't want to use the same data that the model was trained on for evaluation since then it would just be able to memorize the questions,
just as you wouldn't want to use the questions from your math homework on the math exam.

    - Sometimes the data we collected needs other forms of adjusting and manipulation—things like duplication, normalization, error correction, and others.

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 on forward.

The next step in our workflow is choosing a model.

STEP 3 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, such as text or music, some for numerical data, and others for text-based data.

In our case, we have just two features-- color and alcohol percentage.

We can use a small linear model, which is a fairly simple one that will get the job done.
Now we move on to what is often considered the bulk of machine learning--
the training.In this step, we'll 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 they should be pressed or 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 at driving. The act of driving and reacting to real-world data has adapted their driving abilities, honing their skills.
We will do this on a much smaller scale with our drinks.

    - In particular, the formula for a straight line is y equals mx plus b, where x is the input,m is the slope of the line, b is the y-intercept, and y is the value of the line at that position x.

The values we have available to us to adjust or train are just m and b, where the m is that slope and b is the y-intercept. 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 values is usually formed into a matrix that is denoted w for the weights matrix.

  Similarly, for b, we arranged them together,and that's called the biases.

STEP 4 TRAINING

The training process involves initializing some random values
for w and b and attempting to predict the outputs with those values.

As you might imagine, it does pretty poorly at first,but we can compare our model's predictions with the output that it should have produced and adjust the values in w and b such that we will have more accurate predictions on the next time around.

- So this process then repeats.
  Each iteration or cycle of updating the weights and biases is called one training step.

  So let's look at what that means 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 the ideal separation of the wine and beer.

STEP 5 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 is 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.

STEP 6 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 some of our parameters. There were a few that we implicitly assumed when we did our training, and now is a good time to go back and test those assumptions, try other values.

One example of a parameter we can tune is
(i) how many times we run through the training set during training. We can actually show the data multiple times. So by doing that, we will potentially lead to higher accuracies.
(ii) 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 as well in determining the outcome of training. Differences can be seen depending on whether a model starts off training with values initialized at zeros versus some distribution of the values and what that distribution is.

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 for you. Otherwise, we might find ourselves tweaking parameters for a very long time.

Now, these parameters are typically referred to as hyperparameters.

The adjustment or tuning of these hyperparameters still remains a bit more of an art than a science, and it's 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,

STEP 7 Prediction

guided by the evaluation step, it'sfinally time to use your model to do something useful. Machine learning is using data to answer questions,so prediction or inference is that step where we finally get to answer some questions.

  - This is the point of all of 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 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 judgment 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 it and evaluating it, doing your hyperparameter training, and finally, prediction.

If you're looking 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. And don't worry, you can't break the site. Of course, we will encounter more steps and nuances in future episodes, but this serves as a good foundational framework to help us think through the problem, giving us a common

language to think about each step and go deeper in the future.Next time on AI Adventures, we'll build our first real machine learning model, using code--no more drawing lines and going over algebra.

DAY 2 APPROACHES

- **APPROACHING MACHINE LEARNING**

https://lnkd.in/dpy_uPFP

3 Types of Machine Learning You Should Know

Written by Coursera • Updated on Jun 16, 2023

https://coursera.org/share/e93159a6dd2ed098badb10bcd0973b93

- **ML MODELS EXPLAINED**
  https://www.youtube.com/watch?v=yN7ypxC7838

- **WHAT ARE THE DIFFERENCES BETWEEN ML ALGORITHM AND ML MODEL?**

https://youtu.be/eYlbg0TqcNk

Types of machine learning models

There are two types of problems that dominate machine learning: classification and prediction. These problems are approached using models derived from algorithms designed for either classification or regression (a method used for predictive modeling). Occasionally, the same algorithm can be used to create either classification or regression models, depending on how it is trained.

Below you will find a list of common algorithms used to create classification and regression models.

Classification models

- **Logistic regression**
- **Naive Bayes**
- **Decision trees**
- **Random forest**

- **K-nearest neighbor (KNN)**
- **Support vector machine**

Regression models

- **Linear regression**
- **Ridge regression**
- **Decision trees**
- **Random forest**
- **K-nearest neighbor (KNN)**
- **Neural network regression**

Source: https://www.coursera.org/articles/machine-learning-models

DAY 3 MACHINE LEARNING PROJECTS

- **CHOOSING THE MODEL**

https://youtu.be/SF0YdBXjr_A

- **End To End Machine Learning Project Implementation With Dockers,Github Actions And Deployment**
  https://youtu.be/MJ1vWb1rGwM

- **Leak Detection System using Machine Learning Techniques - Daniele Kappes by InfoQ Brasil**

https://youtu.be/nmZ4tjAE-EY

- **BIOINFORMATICS FIELD**
  Build your first machine learning model in Python by Data Professor

https://youtu.be/29ZQ3TDGgRQ

DAY 4 MACHINE LEARNING PROJECTS continued

GOOGLE CLOUD TECH

We look at AI in solving problems today. Playlist:

https://youtube.com/playlist?list=PLIivdWyY5sqJ1YuMdGjRwJ3fFYZ_vWQ62&si=5uvhIQDHFAGR0qbw

GOOGLE BIGQUERY (SQL NO PYTHON)

Querying 100 Billion Rows using SQL, 7TB in a single table
https://youtu.be/Eo4cB7IvLhg

PROBLEM SOLVING GOOGLE CLOUD TECH

https://developers.google.com/machine-learning/problem-framing/problem

DAY 5 REVISION, QUIZZES AND DISCUSSION

REVISION QUESTIONS

1. **What are the 7 step of ML?**
2. **What are the 3 types(approaches) of ML?**
3. **What are the 2 dominant classifications of ML models?**
4. **Which tools or programming language(s) would you use at work in ML?**
5. **If Python doesn't scale which programming language(s) would you learn to use?**

QUIZ

1. **What are the differences between Data Science, ML and AI?**
2. **What is NLP?**
3. **What is Python?**
4. **What is GoLang?**
5. **Name any 3 programming languages or tools used in Data Science & ML.**

DISCUSSION

In groups of 2 or more discuss which career path you would like to follow?