Preparing for a Data Science/Machine Learning program

Author: Kanja Saha

Date: August 25, 2019

LinkedIn post

If you are reading this article, there is a good chance you are considering taking a Machine Learning(ML) or Data Science(DS) program soon and do not know where to start. Though it has a steep learning curve, I would highly recommend and encourage you to take this step. Machine Learning is fascinating and offers tremendous predictive power. If ML researcher continues with the innovations that are happening today, ML is going to be an integral part of every business domain in the near future.

Many a time, I hear, "Where do I begin?". Watching videos or reading articles is not enough to acquire hands-on experience and people become quickly overwhelmed with many mathematical/statistical concepts and python libraries. When I started my first Machine Learning program, I was in the same boat. I used to Google for every unknown term and add "for dummies" at the end :-). Over time, I realized that my learning process would have been significantly smoother had I spent 2 to 3 months on the prerequisites (7 to 10 hours a week) for these programs. My goal in this post is to share my experience and the resources I have consulted to complete these programs.

One question you may have is whether you will be ready to work in the ML domain after program completion. In my opinion, it depends on the number of years of experience that you have. If you are in school, just graduated or have a couple of years of experience, you will likely find an internship or entry-level position in the ML domain. For others with more experience, the best approach will be to implement the projects from your boot camp at your current workplace on your own and then take on new projects in a couple of years. I also highly recommend participating in Kaggle competitions and related discussions. It goes without saying that one needs to stay updated with recent advancements in ML, as the area is continuously evolving. For example, automated feature engineering is growing traction and will significantly simplify a Data Scientist's work in this area.

This list of boot camp prerequisite resources is thorough and hence, long :-). My intention is NOT to overwhelm or discourage you but to prepare you for an ML boot camp. You may already be familiar with

some of the areas and can skip those sections. On the other hand, if you are in high school, I would recommend completing high school algebra and calculus before moving forward with these resources.

As you may already know, Machine Learning (or Data Science) is a multidisciplinary study. The study involves an introductory college-level understanding of Statistics, Calculus, Linear Algebra, Object-Oriented Programming(OOP) basics, SQL and Python, and viable domain knowledge. Domain knowledge comes with working in a specific industry and can be improved consciously over time. For the rest, here are the books and online resources I have found useful along with the estimated time it took me to cover each of these areas.

Before I begin with the list, a single piece of advice that most find useful for these boot camps is avoid going down the rabbit hole. First, learn *how* without fully knowing *why*. This may be counter-intuitive but it will help you learn all the bits and pieces that work together in Machine Learning. Try to stay within the estimated hours(maybe 25% more) I have suggested. Once you have a good handle on the how, you will be in a better position to deep five into each of the areas that make ML possible.

Machine Learning:

1. Machine Learning Basics - Principles of Data Science: Sinan Ozdemir does a great job of introducing us to the world of machine learning. It is easy to understand without prior programming or mathematical knowledge. (Estimated time: 5 hours)
2. Applications of Machine Learning - A-Z by Udemy: This course cost less than $20 and gives an overview of what business problems/challenges are solved with machine learning and how. This keeps you excited and motivated if and when you are wondering why on earth you are suddenly learning second-order partial derivatives or eigenvalues and eigenvectors. Just watching the videos and reading through the solutions will suffice at this point. Your priority is code and

ML domain familiarity. (Estimated time: 2-3 hours/week until completion. If you do not understand fully, that is ok at this time).

3. Reference Book - ORielly: Read this book after you are comfortable with Python and other ML concepts that are mentioned here but not necessary to start a program.
4. ML Glossary - readthedocs.org: Learning these(most of them) terms just by themselves perhaps will not make much sense, but once you are in the program, it is quite helpful to quickly refresh you memory every week.

SQL:

1. SQL Basics - HackerRank: You will not need to write SQL as most ML programs provide you with CSV files to work with. However, knowing SQL will help you to get up to speed with pandas, Python's data manipulation library. Not to mention it is a necessary skill for Data Scientists. HackerRank expects some basic understanding of joins, aggregation function etc. If you are just starting out with SQL, my previous posts on databases may help before you start with HackerRank. (Estimated time: Couple of hours/week until you are comfortable with advanced analytic queries. SQL is very simple, all you need is practice!)

OOP:

1. OOP Basics - OOP in Python : Though OOP is widespread in machine learning engineering and data engineering domain, Data Scientists need

not have deep knowledge of OOP. However, we benefit from knowing the basics of OOP. Besides, ML libraries in Python make heavy use of OOP and being able to understand OOP code and the errors it throws will make you self sufficient and expedite your learning. (Estimated time: 10 hours)

Python:

1. Python Basics - learnpython: If you are new to programming, start with the basics: data types, data structures, string operations, functions, and classes. (Estimated time: 10 hours)
2. Intermediate Python - datacamp: If you are already a beginner python programmer, devote a couple of weeks to this. Python is one of the simplest languages and you can continue to pick up more Python as you undergo your ML program. (Estimated hours: 3-5 hours/week until you are comfortable creating a class for your code and instantiating it whenever you need it. For example, creating a data exploration class and call it for every data set for analysis.
3. Data Manipulation - 10 minutes to Pandas: 10 minutes perhaps is not enough but 10 hours with Pandas will be super helpful in working with data frames: joining, slicing, aggregating, filtering etc. (Estimated time: 2-3 hours/week for a month)
4. Data Visualization - matplotlob: All of the hard work that goes into preparing data and building models will be of no use unless we share the model output in a way that is visually appealing and interpretable to your audience. Spend a few hours understanding line plots, bar charts, box plots, scatter plots and time-series that is generally used to present the output. Seaborne is another powerful visualization library but you can look into that later. (Estimated time: 5 hours)

5. Community help - stackoverflow: Python's popularity in the engineering and data science communities makes it easy for anyone to get started. If you have a question on how to do something in Python, you will most probably find an answer on StackOverflow.
6. Time-Space Complexity - Wisconsin-Madison: This is completely optional for preparing for an ML program. However, as a Data Scientist, I believe it is essential to understand the performance of data structure operations, sorting algorithms and the ML algorithms. Data Science (Feature Engineering/ Implementation/ Validation/ Testing) is an iterative process and accuracy/performance is given priority over the time it takes. However, many a time a robust model fail when deployed in production due to run time issues. And that is where the knowledge of Big O notation is valuable. A simpler example and definition is available in geekforgeeks along with some exercise to get comfortable with the concept. (Estimated time: 10 hours and ongoing)

Probability & Statistics:

1. Summary Statistics - statisticshowto: A couple of hours will be sufficient to understand the basic theories: mean, median, range, quartile, interquartile range.
2. Probability Distributions - analyticsvidya: Understanding data distribution is the most important step before choosing a machine learning algorithm. As you get familiar with the algorithms, you will learn that each one of them makes certain assumptions on the data, and feeding data to a model that does not satisfy the model's assumptions will deliver the wrong results. (Estimated time: 10 hours)
3. Conditional Probability - Khanacademy : Conditional Probability is the basis of Bayes Theorem, and one must understand Bayes theorem

because it provides a rule for moving from a prior probability to a posterior probability. It is even used in parameter optimization techniques. A few hands-on exercises will help develop a concrete understanding. (Estimated Time: 5 hours)

4. Hypothesis Testing - PennState: Hypothesis Testing is the basis of Confusion Matrix and Confusion Matrix is the basis for most model diagnostics. It is an important concept you will come across very frequently. (Estimated Time: 10 hours)
5. Simple Linear Regression - Yale & Columbia Business School: The first concept most ML programs will teach you is linear regression and prediction on a data set with a linear relationship. Over time, you will be introduced to models that work with non-linear data but the basic concept of prediction stays the same. (Estimated time: 10 hours)
6. Reference book - Introductory Statistics: If and when you want a break from the computer screen, this book by Robert Gould and Colleen Ryan explains topics ranging from "What are Data" to "Linear Regression Model".

Calculus:

1. Basic Derivative Rules - KhanAcademy: In machine learning, we use optimization algorithms to minimize loss functions (different between actual and predicted output). These optimization algorithms (such as gradient descent) uses derivatives to minimize the loss function. At this point, do not try to understand loss function or how the algorithm works. When the time comes, knowing the basic derivative rules will make understanding loss function comparatively easy. Now, if you are 4 years past college, chances are you have a blurred memory of calculus (unless, of course, math is your superpower). Read the basics

to refresh your calculus knowledge and attempt the unit test at the end. (Estimated Time: 10 hours)

2. Partial derivative - Columbia: In the real world, there is rarely a scenario where there is a function of only one variable. (For example, a seedling grows depending on how sun, water, minerals it gets. Most data sets are multidimensional. Hence the need to know partial derivatives. These two articles are excellent and provide the math behind the Gradient Descent. Rules of calculus - multivariate and Economic Interpretation of multivariate Calculus. (Estimated Time: 10 hours)

Linear Algebra:

1. Brief refresher - Udacity: Datasets used for Machine learning models are often high dimensional data and represented as a matrix. Many ML concepts are tied to Linear Algebra and it is important to have the basics covered. This may be a refresher course, but at their cores, it is equally useful for those who are just getting to know Linear Algebra. (Estimated Time: 5 hours)
2. Matrices, eigenvalues, and eigenvectors - Stata: This post has explained matrices intuitively and will help you visualize them. Continue to the next post on eigenvalues and vectors as well. Many a time, we are dealing with a data set with a large number of variables and many of them are strongly correlated. To reduce dimensionality, we use Principal Component Analysis (PCA), at the core of which is Eigenvalues and Eigenvectors. (Estimated Time: 5 hours)
3. PCA, eigenvalues, and eigenvectors - StackExchange: This comment/answer does a wonderful job in intuitively explaining PCA and how it relates to eigenvalues and eigenvectors. Read the answer

with the highest number of votes (the one with Grandmother, Mother, Spouse, Daughter sequence). Read it multiple times if it does not make sense in the first take. (Estimated Time: 2-3 hours)

4. Reference Book - Linear Algebra Done Right: For further reading, Sheldon Axler's book is a great reference but completely optional for the ML coursework.

These are the math and programming basics that are needed to get started with Machine Learning. You may not understand everything at this point (and that is ok) but some degree of familiarity and having an additional resource handy will make the learning process enjoyable. This is an exciting path and I hope sharing my experience with you helps in your next step. If you have further questions, feel free to email me or comment here!