Frameworks for Approaching the Machine Learning Process

R. S. Sawant*
Department of Computer Sciences, College of Engineering, Pune, Maharashtra, India
Corresponding Author
Email Id: sawant.rs02@gmail.com

ABSTRACT

When choosing a machine learning framework, it's important to understand what type of data you have and what type of applications you want to build. "The reason why people talk about deep learning and unstructured data is that it is the only form of machine learning that can do that today. Deep learning frameworks offer building blocks for designing, training and validating deep neural networks, through a high level programming interface. This eliminates the need to manage packages and dependencies or build deep learning frameworks from source.

Keywords*: Machine Learning Process, Framework, Approaches, Deep Learning.*

INTRODUCTION

Machine learning is simply the scientific study of Algorithms and statistical models that Computer systems use to effectively perform a specific task without explicit instructions, relying on patterns and inference instead [1]. It is a subset of Artificial Intelligence. Each time you do a web search on Google it works so well because their machine learning software has figured out how to rank what pages. When Facebook or Apple recognizes your friend in your friend in your photos or suggest to you your friends on your contact list that's also Machine Learning. Each time you read your Email and your Spam filter saves you from having a wade through tons of spam, that's simply because your Computer has learned to distinguish spam from non-spam email [2]. It is the general concept of getting Computers to learn without being explicitly programmed. Building Intelligent Machines which can do anything you or I can do. Machine Learning has granted Computers systems entirely new abilities. But how does it really work? How does it learn? This i will be giving a walk-through with a basic example, and use it as an excuse talk about the process of getting answers from your data using Machine Learning [3].

IMPORTANT STEPS OF MACHINE LEARNING [4]

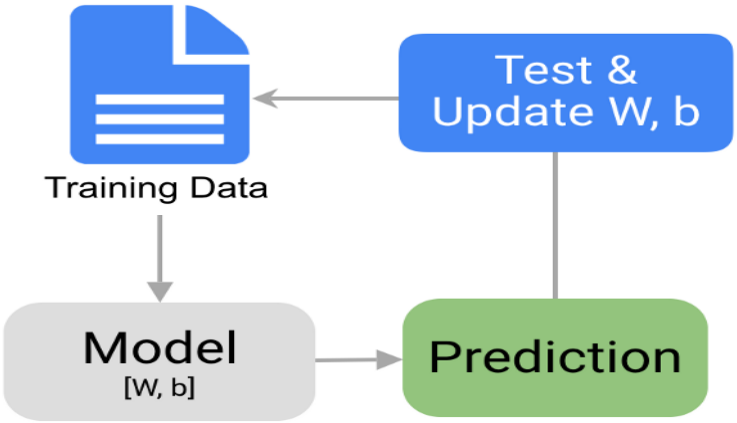

Fig. 1. Steps of Machine Learning [4]

1) Data Collection

- The quantity & quality of your data dictate how accurate our model is
- The outcome of this step is generally a representation of data (Guo simplifies to specifying a table) which we will use for training
- Using pre-collected data, by way of datasets from Kaggle, UCI, etc., still fits into this step [5].

2) Data Preparation

- Wrangle data and prepare it for training
- Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.)
- Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data
- Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis
- Split into training and evaluation sets [6].

3) Choose a Model

- Different algorithms are for different tasks; choose the right one [7].

4) Train the Model

- The goal of training is to answer a question or make a prediction correctly as often as possible
- Linear regression example: algorithm would need to learn values for m (or W) and b (x is input, y is output)
- Each iteration of process is a training step [8]

5) Evaluate the Model

- Uses some metric or combination of metrics to "measure" objective performance of model
- Test the model against previously unseen data
- This unseen data is meant to be somewhat representative of model performance in the real world, but still helps tune the model (as opposed to test data, which does not)
- Good train/eval split? 80/20, 70/30, or similar, depending on domain, data availability, dataset particulars, etc [9].

6) Parameter Tuning

- This step refers to *hyperparameter* tuning, which is an "artform" as opposed to a science
- Tune model parameters for improved performance
- Simple model hyperparameters may include: number of training steps, learning rate, initialization values and distribution, etc [10].

7) Make Predictions

- Using further (test set) data which have, until this point, been withheld from the model (and for which class labels are known), are used to test the model; a better approximation of how the model will perform in the real world [11].

Universal Workflow of Machine Learning [12]
A universal workflow of machine learning, which describes as a blueprint for solving machine learning problems. The blueprint ties together the concepts have learned about in this paper: problem definition, evaluation, feature engineering, and fighting over fitting.

1) Defining the problem and assembling a dataset
2) Choosing a measure of success
3) Deciding on an evaluation protocol
4) Preparing your data
5) Developing a model that does better than a baseline
6) Scaling up: developing a model that overfits
7) Regularizing your model and tuning your parameters

Drafting a Simplified Framework
We can reasonably conclude that Guo's framework outlines a "beginner" approach to the machine learning process, more explicitly defining early steps, while Chollet's is a more advanced approach, emphasizing both the explicit decisions regarding model evaluation and the tweaking of machine learning models [13]. Both approaches are equally valid, and do not prescribe anything fundamentally different from one another; you could superimpose Chollet's on top of Guo's and find that, while the steps of the 2 models would not line up, they would end up covering the same tasks in sum [14].

1) Data collection
   → Defining the problem and assembling a dataset (1)
2) Data preparation
   → Preparing your data (4)
3) Choose model
4) Train model
   → Developing a model that does better than a baseline (5)
5) Evaluate model
   → Choosing a measure of success (2)
   → Deciding on an evaluation protocol (3)
6) Parameter tuning
   → Scaling up: developing a model that over fits (6)
   → Regularizing your model and tuning your parameters (7)
7) Predict
   → It's not perfect, but I stand by it.

CONCLUSION

This presents something important: both frameworks agree, and together place emphasis, on particular points of the framework. It should be clear that model evaluation and parameter tuning are important aspects of machine learning. Addition agreed-upon areas of importance are the assembly/preparation of data and original model selection/training. Let's use the above to put together a simplified framework to machine learning, the main areas of the machine learning process**:**

1) **Data collection and preparation**: everything from choosing where to get the data, up to the point it is clean and ready for feature selection/engineering

2) **Feature selection and feature engineering**: this includes all changes to the data from once it has been cleaned up to when it is ingested into the machine learning model
3) **Choosing the machine learning algorithm and training our first model**: getting a "better than baseline" result upon which we can (hopefully) improve
4) **Evaluating our model**: this includes the selection of the measure as well as the actual evaluation; seemingly a smaller step than others, but important to our end result
5) **Model tweaking, regularization, and hyperparameter tuning**: this is where we iteratively go from a "good enough" model to our best effort

REFERENCES

1) A Framework for Approaching Textual Data Science Tasks
2) A General Approach to Preprocessing Text Data
3) The Data Science Process, Rediscovered
4) Bishop, C. M. (2006), Pattern Recognition and Machine Learning, Springer, ISBN 978-0-387-31073-2
5) Machine learning and pattern recognition "can be viewed as two facets of the same field
6) Friedman, Jerome H. (1998). "Data Mining and Statistics: What's the connection?". Computing Science and Statistics. **29** (1): 3–9.
7) "What is Machine Learning?". www.ibm.com. Retrieved 2021-08-15.
8) Zhou, Victor (2019-12-20). "Machine Learning for Beginners: An Introduction to Neural Networks". Medium. Retrieved 2021-08-15.
9) Domingos 2015, Chapter 6, Chapter 7.
10) Ethem Alpaydin (2020). Introduction to Machine Learning (Fourth ed.). MIT. pp. xix, 1–3, 13–18. ISBN 978-0262043793.
11) Samuel, Arthur (1959). "Some Studies in Machine Learning Using the Game of Checkers". IBM Journal of Research and Development. **3** (3): 210–229. CiteSeerX 10.1.1.368.2254. doi:10.1147/rd.33.0210.
12) R. Kohavi and F. Provost, "Glossary of terms," Machine Learning, vol. 30, no. 2–3, pp. 271–274, 1998.
13) Gerovitch, Slava (9 April 2015). "How the Computer Got Its Revenge on the Soviet Union". Nautilus. Retrieved 19 September 2021.