A Few Useful Things To Know About Machine Learning

Mahdi Bostanabad
Sharif University of Technology

Principles of Machine Learning
Monday 3rd March, 2025

Contents

Introduction

The paper "A Few Useful Things To Know About Machine Learning", published in the Communications of the ACM, 2012, discusses some important notes on practical aspect of ML.

The author, Pedro Domingos, is a well-known data scientist and ML researcher known for Markov logic network enabling uncertain inference.

In this presentation, we will review few but crucial notes on how to use ML like an experienced scientist and be aware about its practical implementations in the real world.

Learning

Learning = Representation + Evaluation + Optimization

Representation

- Choosing a representation for a learner (ML model) is tantamount to choosing the set of models that it can possibly learn. This set is called the *hypothesis space* of the learner.

Evaluation

- An evaluation function is needed to distinguish good learners from bad ones.

Optimization

- The choice of optimization technique is key to the efficiency of the learner, and also helps determine the classifier produced if the evaluation function has more than one optimum.

Learning

Table 1. The three components of learning algorithms.

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search
Hyperplanes	Likelihood	Branch-and-bound
Naive Bayes	Posterior probability	Continuous optimization
Logistic regression	Information gain	Unconstrained
Decision trees	K-L divergence	Gradient descent
Sets of rules	Cost/Utility	Conjugate gradient
Propositional rules	Margin	Quasi-Newton methods
Logic programs		Constrained
Neural networks		Linear programming
Graphical models		Quadratic programming
Bayesian networks		
Conditional random fields		

Learning – It's Generalization that Counts!

The fundamental goal of ML is to generalize beyond the examples in the training set.

- Doing well on solely training set means nothing! That is just memorizing and not learning.
- The matter of train/test split highly relies on data. Do not perform a simple 80/20 or 70/30 without inspecting your data!
- Better representation $\xrightarrow{\text{often}}$ Better generalization

Learning

Learn many models, not just one!

- Search for the *best* ML model given the data and task (how?)
- How to train and test multiple models while being computationally efficient? (will come back to this later)
- Ensemble methods:
  - Bagging
  - Boosting
  - Stacking

Learning – Representation

Representable Does Not Imply Learnable!

Just because a function can be represented does not mean it can be learned.

- Standard decision tree learners cannot learn trees with more leaves than there are training examples
- In continuous spaces, representing even simple functions using a fixed set of primitives often requires an infinite number of components
- Some representations are exponentially more compact than others for some functions $\longrightarrow$ Require exponentially less data to learn those functions.
- Finding methods to learn these deeper representations is one of the major research frontiers in ML (This is not actually true in 2025!)
- Machine Learning $\rightarrow$ Representation Learning $\rightarrow$ Deep Learning

Data

Data Alone Is Not Enough!

- Data alone is not enough, no matter how much of it you have.
- Every learner must embody some knowledge or assumptions beyond the data it is given in order to generalize beyond it.
- One key criterion for choosing a representation is which kinds of knowledge are easily expressed in it
  - We have enough knowledge about what makes examples similar in our domain $\rightarrow$ Instance-based methods
  - We have knowledge about probabilistic dependencies $\rightarrow$ Graphical models
- ML is not magic; it cannot get something from nothing. What it does is get more from less!

Data

Correlation Does Not Imply Causation!

- ML is usually applied to *observational* data, as opposed to *experimental* data.
- By designing experiments, we can *control* the inherent causalities $\longrightarrow$ Drug Design, etc.
- Good to note that correlation is a sign of a potential causal connection, and we can use it as a guide to further investigation.

- Propensity Score Matching
  - Regression
  - Hypothesis Testing
  - ...

Data – High Dimensions

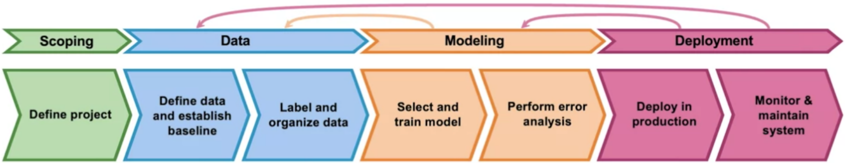
Intuition Fails In High Dimensions

- After *overfitting*, the biggest problem in ML is *the curse of dimensionality*.
- Breakdown of many ML algorithms that rely on similarity-based reasoning
- More features $\longrightarrow$ More data $\longrightarrow$ Higher dimensions
- This is why *dimensionality reduction* methods matter very much!

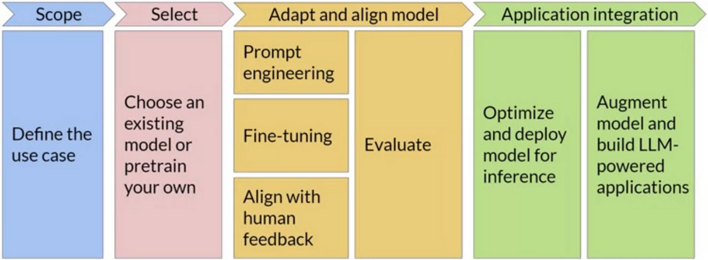
Note That

- A top-notch ML Engineer and/or Data Scientist is also somewhat a Software Engineer (AI-Powered Products, Agentic AI, . . .).
- To earn $$ is to be the *fastest* and not necessarily the most efficient!
  - AutoML frameworks (TPOT, H2O, . . .)
  - Neural Architecture Search (NAS) with RL
- To sell an ML model is to design an *ML system* and make the *pipleline* into a product.
  - Scope definition
  - . . .
  - MLOps
  - . . .
  - Deployment
  - . . .
  - Monitoring

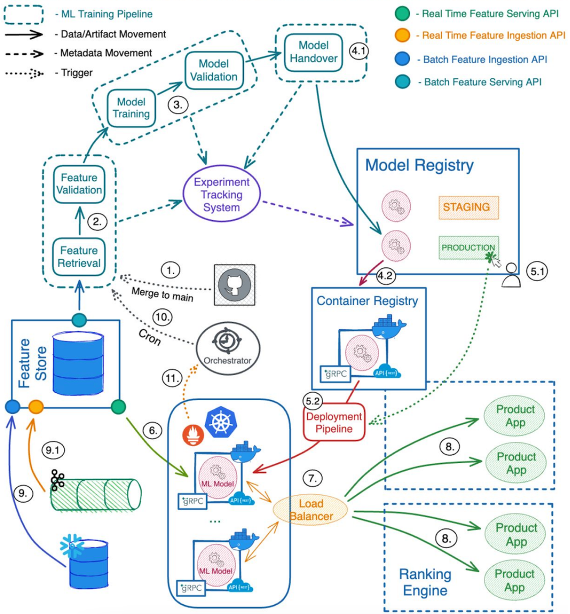
ML Project Lifecycle



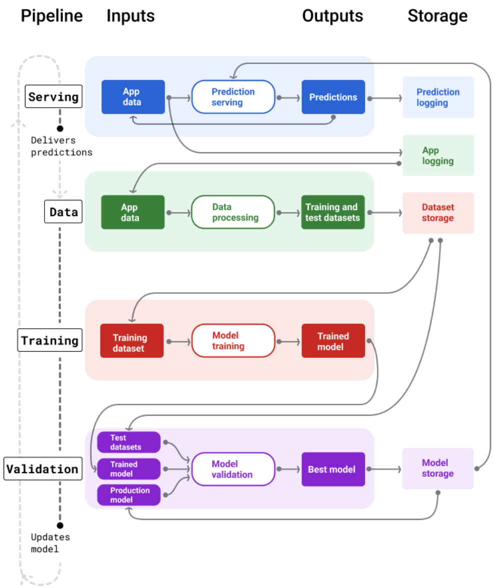
Generative AI LLM Project Lifecycle



An Example of A Pipeline



An Example of A Pipeline



References

[1] Domingos, Pedro M.. “A few useful things to know about machine learning.” *Communications of the ACM* 55 (2012): 78 - 87.

[2] Wagstaff, Kiri. “Machine learning that matters.” *arXiv preprint arXiv:1206.4656* (2012).