Data Ninjas

Machine Learning - A Practical Guide

Organizations often want to leverage machine learning but are afraid because they don't know where to start.

Use this guide and you can get started in days…

Contents

Chapter 1

What is machine learning?

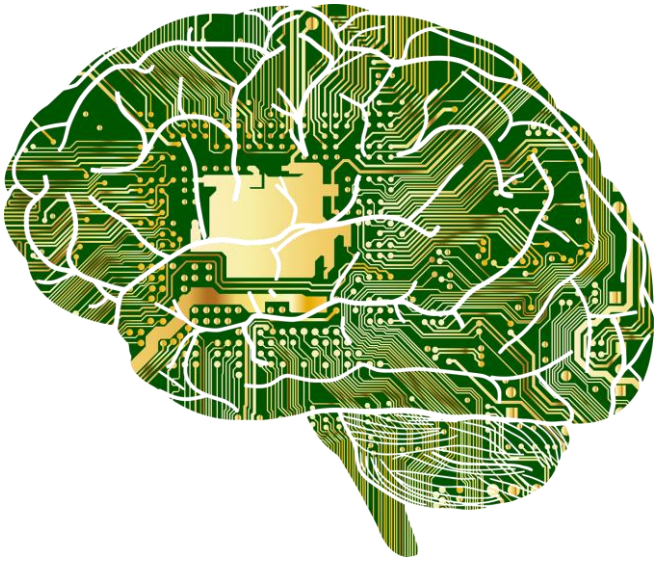
Arthur Samuel (1959) gave us the following definition –

Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.

Now that we have defined what machine learning is – we will explore further some of the applications (or use cases) in following chapters. Depending on the type of observations provided, machine learning can be split into two major subdisciplines (and two minor subdisciplines too!)

These are 1. supervised learning – "right answers" are provided to machines 2. unsupervised learning – "no right answers, unlabeled data" and the minor variations 1. reinforcement learning 2. recommender systems

We will take a look at each type of machine learning in detail. A key aspect of ML that makes it particularly appealing in terms of business value is that it does not require as much explicit programming in advance to gain intelligent insight because of its ability to use learning algorithms that simulate some human learning capabilities.

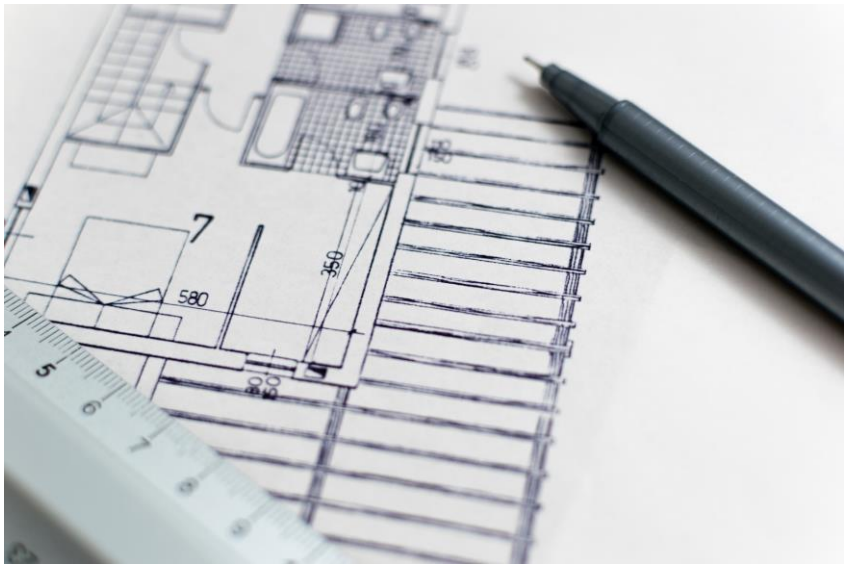


Once data are acquired and prepared for ML, and algorithms are selected, modeled and evaluated, the learning system proceeds through learning iterations on its own to uncover latent business value from data. The bulk of the work however stays within the bounds of acquiring and preparing the data aka features.

Yes, machine learning may identify previously unidentified opportunities or problems to be solved. But the machine is not autonomously creative. The machine will not spontaneously develop new hypotheses from facts (data) not in evidence. Nor can the machine determine a new way to respond to emerging stimuli.
However, machines can learn from and use predictions created by a machine learning pipeline or ensemble learning. We will discuss about that at a later point but first things first – let's review some core use cases where machine learning is applicable.

Chapter 2

Machine Learning Use Cases



Most organizations have use cases they could explore and apply machine learning to get the results. These use cases can be categorized on the functional areas – as described in the right column on this page or by two core segments as follows

Discovery

Business Question: What business problem might we have? Do we know all the parameters that affect a specific outcome?

Example: How are our customers really segmented?

Predictive

Business Question: Which of the current business challenge would we like to predict?

Example: What will our customer buy next? What will be demand for our products for the next month or during the holidays?

Marketing:
- Predicting life time value
- Wallet share estimate
- Churn
- Customer segmentation
- Product Mix

Sales:
- Lead scoring
- Demand forecasting
- Sales Forecasting

Logistics:
- Demand forecasting
- Failures & outliers

Risk:
- Credit risk
- Currency risk
- AP Recovery
- AML

Customer Support:
- Volume predictions
- Optimization

Can you use machine learning for any business problem?

We have to say no! Despite what your best guess is or what you are being told by the hype, machine learning cannot be used to solve every business problem. We would like to explain why that is so.

Most organizations have two types of business problems that can often be mistaken for the other type.

The first type of problem is that you are looking for efficiency gains. These tasks are:

1. Well defined or clear.
2. These tasks follow well defined sequence of steps.
3. Usually are being executed by human(s)
4. Are human error prone.

A lot of organizations are doing this today – examples are manual validation or data entry tasks. Sometimes organizations also rely upon humans to integrate data to generate reports via tools like spreadsheets

The second type of problem exists when simple or complex automation just is not enough. It takes machines to understand hidden relationships or patterns that a human intuition does not fully comprehend. These types are problems are good candidates for machine learning projects. These patterns that we are talking about inherently are centered around predictability.

Focus on Learning

Problems that require learning rather than just intuition should be prime candidates for machine learning.

Ask yourself and validate with business owners if the problem can be solved using automation.

Also spend enough time to ensure that all of the influences (factors) are reasonably well known before deciding on using machine learning.

Automation problems can also use machine learning (but in a pipeline)

You can think of entering a purchase order into an ERP system to be a good candidate for automation, however you can use the historical data for purchase orders to predict your purchase patterns for a given quarter (unless of course you are implementing a new product/service line and making capital investments).

You can also use machine learning to build better data quality

Chapter 3

Machine Learning Life Cycle

Most organizations have use cases they could explore and apply machine learning to get the results. These use cases can be categorized on the functional areas – as described in the right column on this page or by two core segments as follows

Discovery

Business Question: What business problem might we have? Do we know all the parameters that affect a specific outcome?

Example: How are our customers really segmented?

Predictive

Business Question: Which of the current business challenge would we like to predict?

Example: What will our customer buy next? What will be demand for our products for the next month or during the holidays?

Machine Learning Life Cycle

Problem Identification:
- Automation problem?
- Learning problem?

Problem Understanding:
- Four step evaluation
- Explore or Predict?
- Supervised or unsupervised?

Data Acquisition:
- Source of record
- Data Quality at start?

Data Processing:
- Features
- Calculated features

Model the problem:
- ML Model identification
- Rapid model sufficiency

Test & Validation:
- Test with relevant data
- Validate with relevant data

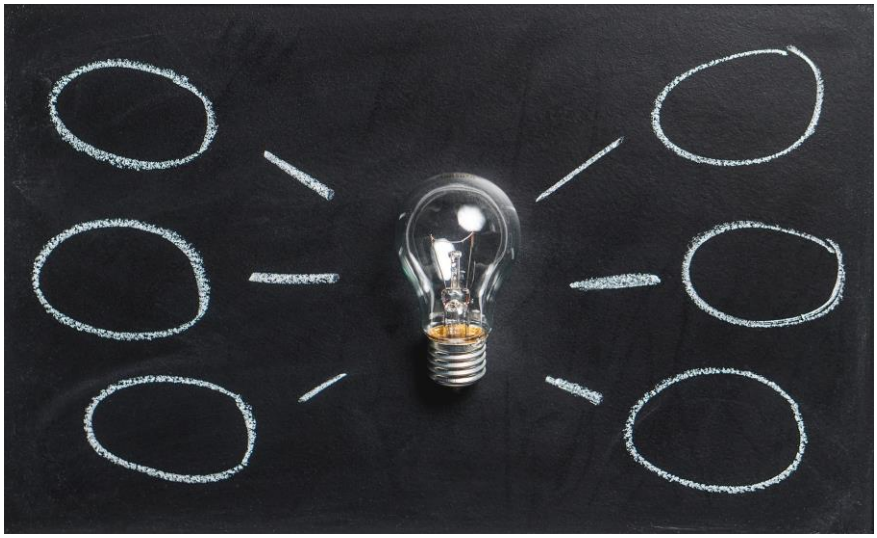
Deploy

Problem identification is the most important step in a machine learning project lifecycle. Once the problem is identified, the next step is to understand the problem in its entirety. This means gaining a thorough understanding of whether the solution will help you explore or predict an outcome.

Based on the assessment, you will be able to identify whether supervised or unsupervised learning is needed.

Data Acquisition
Data scientists spend a lot of time identifying and acquiring the data needed to solve a machine learning problem. This steps often determines how (quickly) successful a particular project would end up. Our recommendation is to take a reasonable amount of time to identify and acquire the data (all the data!) that you may need.

Training machine learning algorithms and generating predictions is not difficult, however the age-old principle of "garbage-in, garbage-out" applies to machine learning too!



Problem identified and classified, data acquired – now what?

Data processing or feature engineering is the all important next step.

Data scientists spend a lot of time in matching the right data to the right model. Data transformation is the key. Data are transformed to serve two purposes. First purposes is to format data in a way so that they can be consumed by the model well. Unstructured, semi-structured and structured data have to be prepped so that the models can consume the data. The second aspect of data processing is to infuse better quality in the data.

We recommend that you perform **rapid prototyping** of machine learning algorithms. Once you have defined the model category – there will be several algorithms at your disposal and the best fit may not be apparent at first.

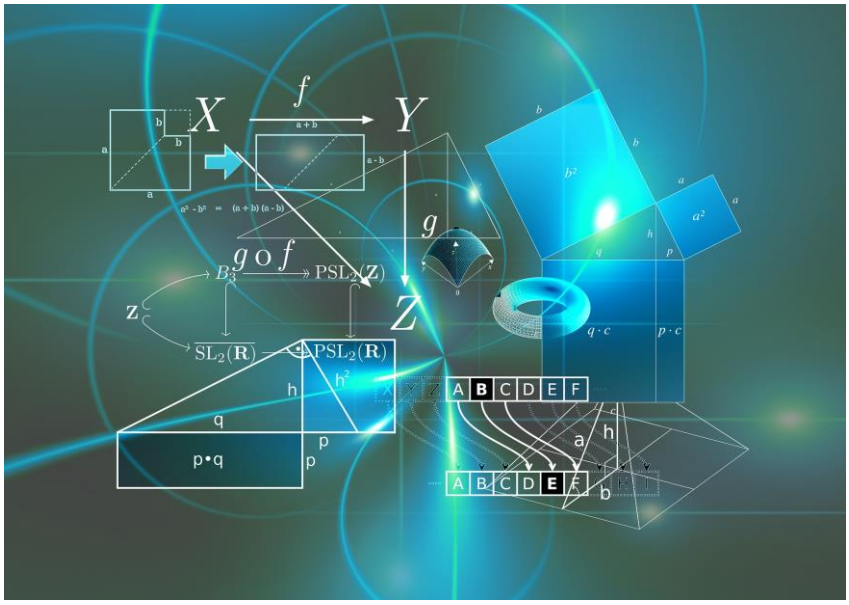
The model choice and rapid prototyping will also help you think and create appropriate features. Most machine learning algorithms can be used for to solve a particular problem. We will talk about the top 10 models in the next chapter.

Test & validation of the model and the result is the next important step. There are several things that happen with models and data. The two important aspects of any machine learning model is how much bias and variance – two words you will hear a lot, factor into the predictions. It is also important to understand the "fit" of the model. We will discuss these in detail in sequel to this e-book.

Deployment of the model is the next logical step. This usually concludes the first iteration of machine learning projects. We would like to reiterate that is it important to treat machine learning as an iterative process.

Chapter 4

Machine Learning Algorithms



Machine learning algorithms are categorized in two broad categories. These are described below. The slightly lesser used categories are discussed in this chapter with a slightly lower emphasis.

Supervised Learning

Supervised learning is commonly used to identify relationship between labeled data. Generally the independent variables are identified and measured and the effects of the values on dependent variable are unknown

The supervised algorithms are further used for two types of problem solving – classification, where labels are discreet values and regression, where labels are continuous values

Unsupervised Learning

Unsupervised learning is employed where labels are not available and thus the problem is free of label influences.

These algorithms are very useful in finding hidden relationships.

Machine Learning Algorithms

Supervised Learning
- Bayesian Statistics
- Decision Trees
- Forecasting
- Neural Networks
- Random Forests
- Regression Analysis
- Support Vector Machines

Unsupervised Learning:
- Affinity Analysis
- Clustering
- K-means
- Nearest Neighbor
- Self organizing maps

Reinforcement Learning:
- Artificial Neural Networks
- Learning Automata

Topics Worth Reviewing:
- Deep Learning
- Cognitive Computing
- Natural Language Processing

Supervised learning

Algorithm: Neural Networks
Description: Uses interconnected groupings of data – similar to neurons of a brain. Leverages the ability of a neuron structure to find complex and deeper relationship within data.
Uses: Financial outcome predictions, fraud detection, image identifications

Algorithm: Classification and/or regression
Description: Uses interconnected groupings of data – similar to neurons of a brain. Leverages the ability of a neuron structure to find complex and deeper relationship within data.
Uses: Financial outcome predictions, fraud detection, image identifications

Algorithm: Trees/Forests/Jungles
Description: Decisions are made based on conditions and logical groupings. Best used for classification problems because of inherent ability to create subsets
Uses: Any large search use case; risk/threat assessments

Algorithm: Support Vector Machines
Description: Primarily a binary classification algorithm with versatile uses.
Uses: Customer segmentation, spam detection, fraud detection

Algorithm: Ensemble
Description: Powerful and almost "all-in-one"
e.g. A regression tree ensemble is a predictive model composed of a weighted combination of multiple regression trees *algorithms. Uses iterative approach to evaluate better model fit*
Uses: Customer segmentation, spam detection, fraud detection

Unsupervised learning

Algorithm: Clustering
Description: Groups similar data based on the evaluation of unlabeled features. Data points in a cluster are similar to each other than others. There are several method of clustering unlabeled data
Uses: Streaming analytics, underwriting, IoT or sensor data

Algorithm: Pattern recognition
Description: Find out trends or patterns in the data. This technique can be used for supervised or unsupervised learning
Uses: Spam or noise detection, fraud, identity management

Algorithm: Principal Component Analysis
Description: Identify linearly correlated features thereby helping a data analyst to reduce number of features to be learned. Data compression is one of the primary benefits of this algorithm
Uses: Machine learning pipeline, closely coupled segment analysis

Algorithm: Support Vector Machines
Description: Primarily a binary classification algorithm with versatile uses.
Uses: Customer segmentation, spam detection, fraud detection

Algorithm: Anomaly Detection
Description: Better suited to identify large number of anomalies in data, especially different types of anomalies.
Uses: Data Center monitoring, Manufacturing

Another class of machine learning algorithms – recommender systems:

- Primarily used for "you may like" or "you may want to buy" type of decisions aka recommendations
- Actual user driven (or based on actual user behavior) data are not available.
- Slightly different than supervised/unsupervised learning algorithms
- Can be collaborative filtering based or content based
- Content based recommenders focus on properties of items
- Collaborative filters focus on users and items

Chapter 5

More on Algorithms

"It's not who has the best algorithm wins, it's who has the most data"

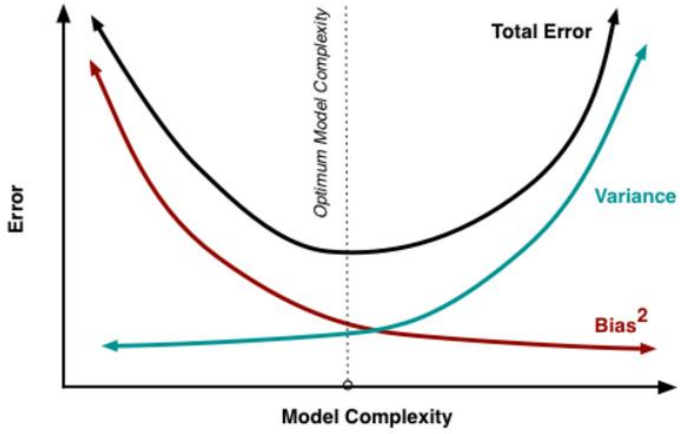
Machine learning algorithms have been around for a number of years and these algorithms continue to get better (even if the core concept may or may not). However most algorithms are limited by what you "feed" into them and that is data.

We recommend that if you follow the guidelines we described earlier in the book, you would have a reasonably self contained problem to solve and possibly the data that you need to solve the problem.

You should look at the size, quality and nature of the data first and also look at the systems limits – this includes physical computational power at your disposal as well as amount of time available to solve (and may be in a recurring way) the problem.

Most experienced data scientists never start with a means in mind. They start with a goal in mind and often times end up experimenting with the machine learning "design" until they are satisfied that the results are as good as they can be.

We also recommend that instead of jumping right into machine learning and building predictions, you perform a thorough descriptive analysis of the data. Sometimes the biggest ideas are revealed by simply looking at some charts – especially boxplots, heatmaps and correlation plots.



We definite recommend using the PCA algorithm as the first step to your machine learning pipeline. This will help you not only remove redundant features from the data, it also helps with compressing the data, thereby reducing the time to run/process the data through the algorithms.

Another key aspect is to reduce bias and variance simultaneously so that you get the least amount of error from the model. We will discuss bias and variance in detail in a future blog.

Chapter 6

1. Educate business or LOB owners in your organizations on the benefits of machine learning. Keep them focused on the goals and not the intricacies of machine learning
2. Do not treat machine learning as an opaque object in your analytical technology domain. Instead, learn the basics before you jump into machine learning projects
3. Learn along with the algorithms. Each iteration of you training the algorithms and testing/validating the results will be just as educating as the final result
4. Don't shorten the machine learning cycle just to advertise flashy results early in the lifecycle. Focus and apply scientific rigor – after all these projects are data research projects
5. Once you start dealing with feature engineering and working (in spite of bias-variance) with volumes of data, the models will become complicated
6. Testing and validation of models are very important. Your best buddies for this task should be business users
7. Once the models are tested and validated, use a storyboard approach to describe the problem and solution to all the stakeholders
8. Machine learning won't replace your gut feel (or the business SME's gut feel) but it will guide you in the right direction
9. Deploy the algorithms in production and continuously measure performance. Don't be afraid to re-train the algorithms after a reasonable time period too!
10. In-production algorithms should result in new, repurposed or discarded business processes. Ensure that change is managed accordingly

Chapter 7

Summary

Machine learning can be an incredibly powerful tool in the arsenal of COO, CFO and CI/TOs. The power has to be wielded properly!

It is incredibly satisfying for a mature organization to reap benefits from machine learning, however it is our belief that even organizations who are trying to build a solid foundation of analytics can benefit from machine learning.

It is important to follow a clear and concise plan to execute on the machine learning projects. This invariably starts with the problem identification and definition. It is important not to get carried away or even lost in the descriptive analysis phase because it has the potential to throw you off track.

Stick to clearly outlined benefits and costs right from the beginning. Let the machine learning capabilities evolve organically and seek help from outside (think your data ninjas!) as and when needed!

Thank you for downloading this book. We wish you happy trails in your journey to use machine learning to its full potential!

We are your Data Ninjas

Data Ninjas have helped customer across many industry verticals to help achieve their goals quickly.

Whether you are looking to implement basic business intelligence or take the next step in machine learning, you will find quality, agility and transparency in Data Ninjas Solutions

Our Focus Areas are
Machine Learning with R/Python/H2O and KNIME

BI with Microsoft, Jaspersoft, Amazon

Data Management with Hadoop, MPP and other leading data management platforms

Get to know us better
www.yourdataninjas.com