Understanding Machine Learning

by Charles A. R. MVB · Jan. 04, 17 · Big Data Zone

Need to build an application around your data? Learn more about dataflow programming for rapid development and greater creativity.

What exactly is machine learning?

The simplest definition I came across:

> ***Machine learning is "[…] the branch of AI that explores ways to get computers to improve their performance based on experience".***
>
> ***Source: Berkeley***

Let's break that down to set some foundations on which to build our machine learning knowledge.

Branch of AI: Artificial intelligence is the study and development by which a computer and its systems are given the ability to successfully accomplish tasks that would typically require a human's intelligent behavior. Machine learning is a part of that process. It's the technology and process by which we train the computer to accomplish the said task.

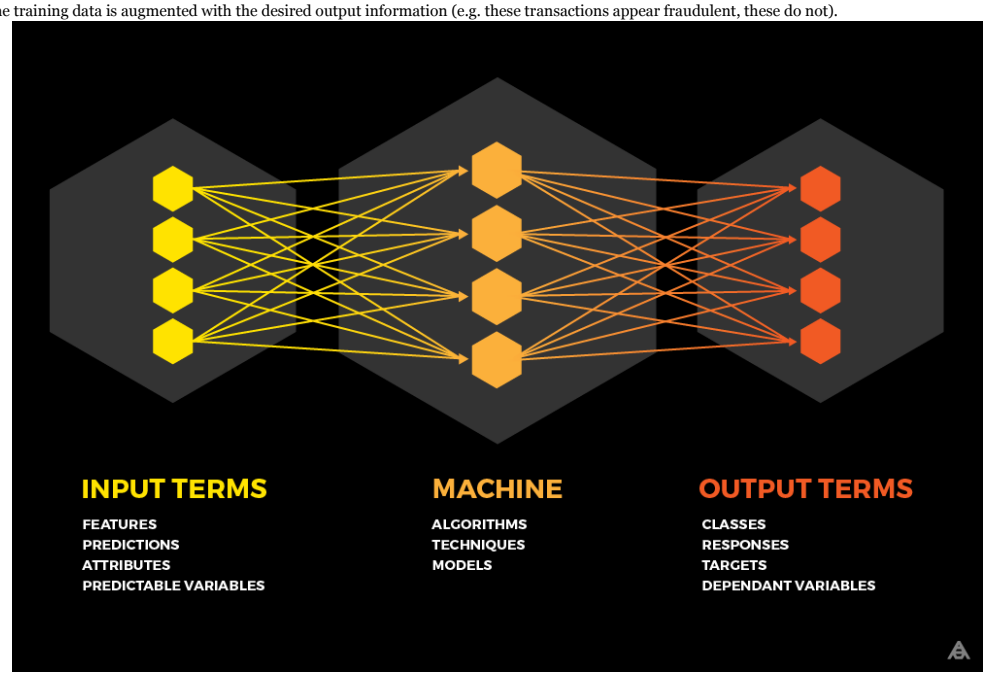
Explores ways: Machine learning techniques are still emerging. Some models for training a computer are already recognized and used (as we will see below), but it is expected that more will be developed with time. The idea to be remembered here is that different models can be used when training a computer. Different business problems require different models.

Get computers to improve their performance: For a computer to accomplish a task with AI, it needs practice and adaptation. A machine learning model needs to be trained using data and in most cases, a little human help.

Based on experience: providing an AI with experience is another way of saying – to provide it with data. As more data is fed into the system, the more accurately the computer can respond to it and to future data that it will encounter. More accuracy in understanding the data means a better chance to successfully accomplish its given task or to increase its degree of confidence when providing predictive insight.

Quick example:

1. Entry data is chosen and prepared along with input conditions (e.g. credit card transactions).
2. The machine learning algorithm is built and trained to accomplish a specific task (e.g.detect fraudulent transactions).
3. The training data is augmented with the desired output information (e.g. these transactions appear fraudulent, these do not).



How Does Machine Learning Work?

Machine learning is often referred to as magical or a black box:

Insert data → magic black box→ Mission accomplished.

Let's take a look at the training process itself to better understand how machine learning can create value with data.

- **Collect**: Machine learning is dependent on data. The first step is to make sure you have the right data as dictated by the problem you are trying to solve. Consider your ability to collect it, its source, the required format, and so on.
- **Clean:** Data can be generated by different sources, contained in different file formats, and expressed in different languages. It might be required to add or remove information from your data set, as some instances might be missing information while others might contain undesired or irrelevant entries. Its preparation will impact its usability and the reliability of the outcome.

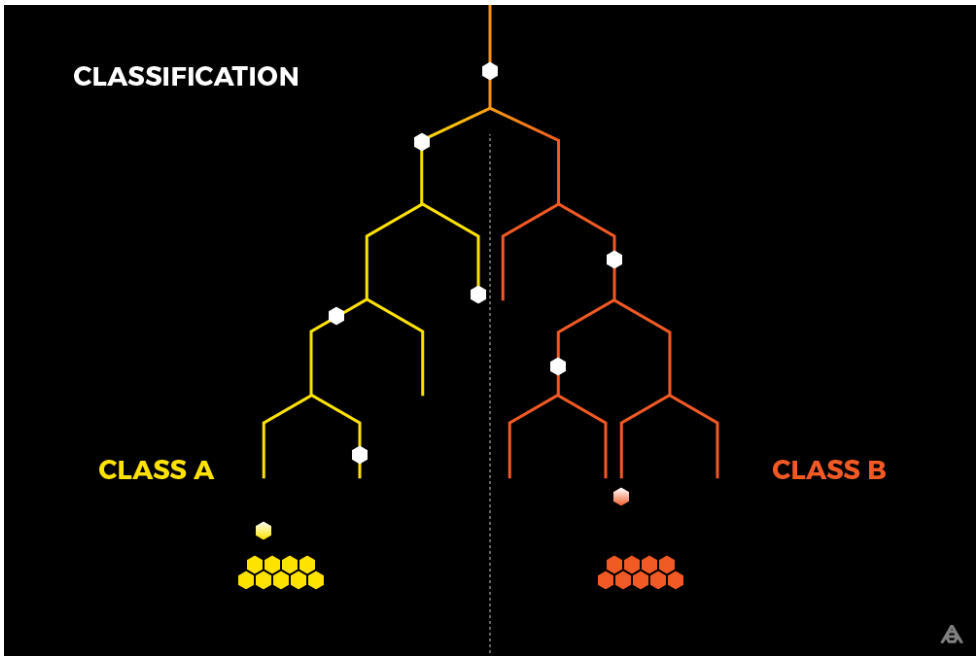
- **Split:** Depending on the size of your data set, only a portion might be required. This is usually referred to as sampling. From the chosen sample, your data should be split into two groups: one to train the algorithm and the other to evaluate it.
- **Train:** This stage essentially aims at finding the mathematical function that will accurately accomplish the chosen goal. Training takes on different forms depending on the type of model used. Fitting a line in a simple linear regression model can be seen as training; generating the decision trees for a Random Forest Algorithm is also training; changing the questions in a decision tree is effectively adjusting the parameters of the model.To keep things simple, let's focus on neural networks. Basically, using a portion of your data set, the algorithm will attempt to process the data, measure its own performance and auto-adjust its parameters (also called backpropagation) until it can consistently produce the desired outcome with sufficient reliability.
- **Evaluate:** Once the algorithm performs well on the training data, its performance is measured again with data that it has not yet seen. Additional adjustments are made when needed. This process allows you to prevent overfitting, which happens when the learning algorithm performs well but only with your training data.
- **Optimize:** The model is optimized for integration within the destined application to ensure it is as lightweight and as fast as possible.

Are There Different Types of Machine Learning?

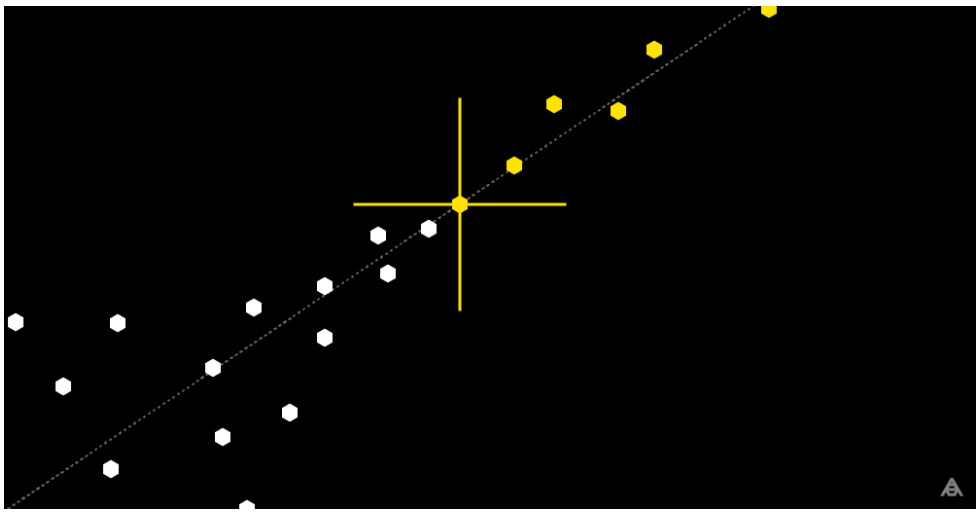
There are many different models that can be used in machine learning but they are typically grouped into three different types of learning: supervised, unsupervised, and reinforcement. Depending on the task to complete, some models are more appropriate and better performing than others.

Supervised learning: in this type of learning, the correct outcome for each data point is explicitly labeled when training the model. This means the learning algorithm is already given the answer when reading the data. Rather than finding the answer, it aims to find the relationship so that when unassigned data points are introduced, it can correctly classify or predict them.

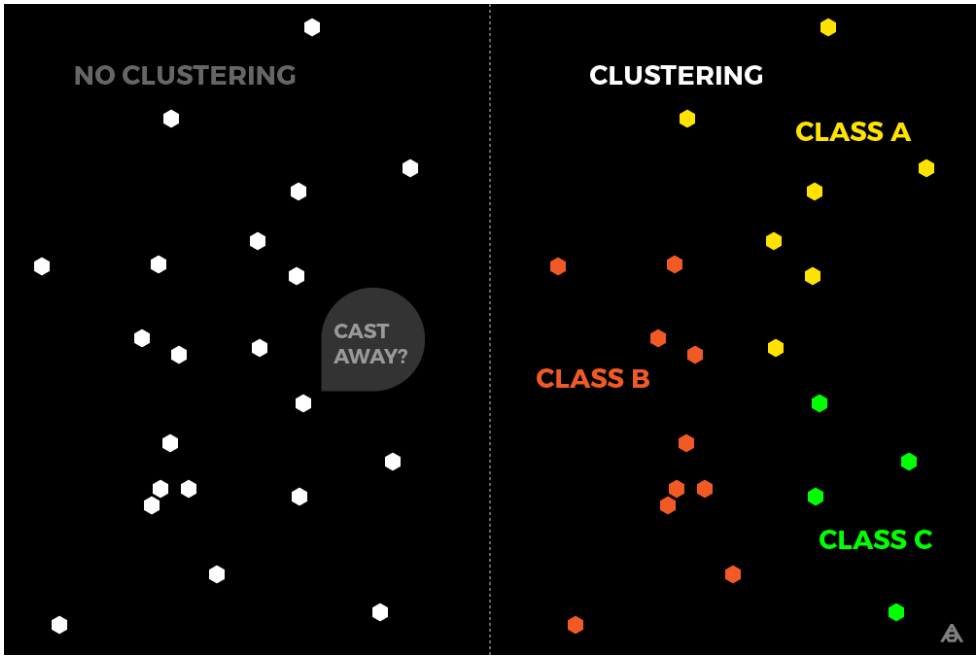
In a classification context, the learning algorithm could be, for example, fed with historic credit card transactions each labeled as *safe* or *suspicious*. It would learn the relationship between these two classifications and could then label new transactions appropriately, according to the classification parameters (e.g. purchase location, time between transactions, etc.).



In a context where data points are continuous in relation to one another, like a stock's price through time, a regression learning algorithm can be used to predict the following data point.



Unsupervised learning: In this case, the learning algorithm is not given the answer during training. Its objective is to find meaningful relationships between the data points. Its value lies in discovering patterns and correlations. For example, clustering is a common use of unsupervised learning in recommender systems (e.g. people who liked this bottle of wine, also enjoyed this one).



Reinforcement learning: this type of learning is a blend between supervised and unsupervised learning. It is usually used to solve more complex problems and requires interaction with an environment. Data is provided by the environment and allows the agent to respond and learn. In practice, this ranges from controlling robotic arms to find the most efficient motor combination, to robot navigation where collision avoidance behavior can be learned by negative feedback from bumping into obstacles. Logic games are also well-suited to reinforcement learning, as they are traditionally defined as a sequence of decisions: games such as poker, backgammon and more recently Go with the success of AlphaGo from Google. Other applications of reinforcement learning are common in logistics, scheduling, and tactical planning of tasks.

What Can Machine Learning Be Used For?

Three stages of machine learning development and their application within a business are to be considered: descriptive, predictive, and prescriptive.

The descriptive stage refers to the recording and analysis of historical data for increased business intelligence. Managers are provided with descriptive information and a better understanding of the results and consequences of past actions and decisions. This process is now routine for most large businesses around the world- for example, reviewing sales records and matching promotional efforts to understand their impact and ROI.

The second stage of applied machine learning is prediction. Gathering data and using it to predict a specific outcome allows for increased reactivity and to make decisions faster and with more accuracy. For example, predicting churn can allow for its prevention. This stage of application is currently being embraced by most businesses.

Yet, the third and most advanced stage of machine learning is already being adopted by existing businesses and pushed forward by newly founded endeavors. Predicting a behavior or outcome is not sufficient when aiming for effective and efficient business practices. Understanding the cause, motive, and context is a prerequisite to optimal decision-making. Concretely, this stage is possible when human and machine combine efforts. Machine learning is used to find meaningful relations and to predict outcomes while data experts serve as translators to make sense of why the relation exists. As such, it becomes possible to prescribe actions with greater precision.

Furthermore, I would add another application of machine learning other than predictive insight: process automation. I've provided a more detailed overview and comparison these two concepts here.

Here are some examples of what problems machine learning can solve.

Logistics and production

- Rethink Robotics uses machine learning to train their robotic arms and improve production speeds;
- JaybridgeRobotics automates industrial grade vehicles for more efficient operations;

- Nanotronics automates optical microscopes for improved inspections;
- Netflix and Amazon optimize resource distribution according to user demand;
- Other examples include: predicting ERP/ERM needs; predicting asset failure & maintenance, improving quality assurance, and increasing production line performance.

Sales and marketing

- 6sense predicts which lead is more susceptible to buy and at what time;
- Salesforce Einstein helps anticipate sales opportunities and automate tasks;
- Fusemachines automates sales tasks with an AI assistant;
- AirPR provides insight to increase PR performance;
- Retention Science suggests cross-channel actions to drive engagement;
- Other examples include: predicting a customer's lifetime value, increasing customer segmentation accuracy, detecting customer shopping patterns, and optimizing a user's in-app experience.

Human resources

- Entelo helps recruiters identify and qualify candidates;
- hiQ assists managers with talent management.

Finance

- Cerebellum Capital and Sentient augment investment management decisions with machine learning powered software;
- Dataminr can assist with real-time financial decisions by providing early alerts on social trends and breaking news;
- Other examples include: detecting fraudulent behavior and predicting stock prices.

Healthcare

- Atomwise uses predictive models to reduce medicine production time;
- Deep6 Analytics identifies eligible patients for clinical trials
- Other examples include: diagnosing diseases more accurately, improving personalized care, and assessing health risks.

You can find even more examples of machine learning and artificial intelligence and other related resources in an awesome list put together by Sam DeBrule.

Before you go.

Remember that collaboration is key. AI and machine learning are fascinating but can be tricky at times. If you are dabbling in AI, you should talk to your local AI expert. If there is one thing I learned about the AI field, is that its members are strongly passionate and are more than willing to help out. You can also comment or ask questions below, it would be my pleasure to help if I can.

Like This Article? Read More From DZone


Machine Learning and the Hunt for Dementia


New AI Turns to Google to Help When It Realizes It Isn't Smart Enough


Hawking: Machine Learning AI Our "Biggest Existential Threat"

Free DZone Refcard
R Essentials

Topics: MACHINE LEARNING , BIG DATA , AI

Published at DZone with permission of Charles A. R., DZone MVB. See the original article here.
Opinions expressed by DZone contributors are their own.

Hybrid Data Solutions Power Digital Enterprises

by Tom Smith · Apr 19, 17 · Big Data Zone

It was great talking to **Jeff Veis**, S.V.P., CMO, **Dave Postle**, Global Head of Services, and **John Bard**, Director of Product Marketing at Actian. They are focused on providing an easy-to-integrate solution for developers to use in data management, integration, and analytics with APIs and open interfaces. According to Jeff, "Analytics are eating software and shifting how people think about computing."

Actian has just introduced Actian X to unify diverse data across the enterprise. Actian X is the first native and hybrid database that combines the power of the proven Actian Ingres OLTP database with Actian's industry leading Vector analytics query engine to deliver high-speed performance and scalability critical to powering next generation digital enterprises.

While the industry talks about hybrid deployment models, today's transactional, operational, and analytic data is typically managed in silos, limiting performance and actionable insights. Traditional solutions from established providers usually feature monolithic platforms that increase complexity, cost, and lock-in without delivering the promised performance or insight. Actian X and Actian DataConnect 11 integration offerings deliver solutions specifically designed to meet the demands of today's hybrid data-driven enterprise.

Delivering Industry-Leading Scale and Speed

Actian embraces the hybrid data ecosystems that are emerging with cloud computing, combining best-fit tools to bridge on-premise and cloud environments while powering modern data-driven applications and services. With query times up to 10x faster than the competition, Actian X brings the

record-breaking performance of Actian Vector analytics into the OLTP database to process transactional, analytical, and hybrid workloads from a single database running on a single compute node.

Actian X's common SQL language interface and management framework seamlessly deliver operational analytics and enable a new class of applications that can interleave OLTP and analytics queries on the fly. Unlike alternative solutions that are tied to specific applications or are limited by available system memory, Actian X delivers the capabilities, scale and speed not available in a relational database.

"Actian has made strategic investments in its technology and tools to address the demands of today's enterprise data ecosystem," says Rohit De Souza, CEO of Actian. "The need for rapid insight to make real-time customer offers, detect fraud quickly or optimize supply chains requires enterprises to fundamentally rethink how they employ their data. Hybrid data management integrates high-performance analytics within an enterprise's mission-critical transactional data systems resulting in a system that can analyze and act at the speed of business. This is just the start of a multi-phase plan to unify and transform the world of data analytics."

"The inclusion of the Actian vector-based X-100 query engine is a game changer for us," says Geraint Jones, Database Administrator at the Clinical Trials Service Unit at the University of Oxford. "We've seen the levels of performance this engine brings where we can quickly make a snapshot of our relational data and run analytic queries on billions of data points without needing to export everything out of one installation into another. It opens an entirely new world of possibilities."

Actian X

The first natively integrated hybrid database is designed to manage transactional, analytic, and hybrid data workloads from a single database. Key features of this new release are the following.

Delivers Operational Analytics

Expands capabilities of existing Ingres applications, making it possible to analyze clickstream data and historical customer information to identify targeted add-ons, as well as to cross-sell and upsell opportunities in near real-time, thus increasing retail sales.

Advanced In-Database Functionality

Adds new OLTP features and geospatial algorithm support for enhanced transactional performance and location-based applications.

Improves Outcomes With Integration

Includes new integration capabilities to extend connectivity to a broad spectrum of data sources both inside and outside of the enterprise.

Keeps Business Critical Systems Healthy

Includes the Actian Enterprise Monitoring Appliance to keep track of the health of the database and host system by monitoring and setting alerts for key system functions like disk usage, I/O performance, transaction log files, and network latency.

Improves Disaster Recovery

Actian's new cloud-based DataCloud Backup delivers native managed cloud service for backup built specifically for Actian's relational database offerings with unprecedented scale, security, and economics.

Actian DataConnect 11

Delivering on Actian's hybrid data vision, Actian also launced Actian DataConnect 11. With its zero migration design this flexible, enterprise-class integration solution delivers the foundation necessary to power a next generation cloud architecture. Featuring intuitive workflow design, simplified administration and support for the popular Eclipse open source framework, DataConnect 11 brings new levels of integration and ease of use across enterprise data centers and public cloud environments.

Topics: BIG DATA, OLTP, HYBRID DATA MANAGEMENT, ACTIAN X, ENTERPRISE

Opinions expressed by DZone contributors are their own.

The 7 Types of Data Scientists

by Muktabh Srivastava · Apr 19, 17 · Big Data Zone

Need to build an application around your data? Learn more about dataflow programming for rapid development and greater creativity.

I recently got in question on Quora asking something on lines of what exact skills companies look for when they are recruiting a Data Scientist and whether there is a definition of what exactly a Data Scientist is. As is pretty obvious, there is no one definition, as every company is solving its own set of problems. But I tried to make a few generic job profiles that can somewhat fit the job descriptions of different companies.

There's way more variety, but I've narrowed it down to a couple of general profiles.

1. The R-Using Number-Cruncher

This type of Data Scientist can run quick group counts in R and Python. He or she is the coding version of a Data Analyst from the earlier days. This type of Data Scientist is mostly involved in automated report generation in more analytical organizations.

Tools used: R (dataframes) and SQL.

2. The Modeller

This type of Data Scientist has a deeply mathematical mind and can apply Bayesian/Frequentist inferences/hierarchal models. I'm probably grouping too many people into a single group here, but the common theme here is that mathematics forms the base of the work.

Tools used: R, Fortran, C++, and sometimes, functional languages.

3. The Data Engineer Who's an Occasional Data Scientist

Take a library from here, take some code from there, and make something good enough while you manage the data pipeline. This is a very common type of Data Scientist. Tasks include writing programs to automate report generation in Pandas, trying out simple Machine Learning models, and (nowadays) running a pre-trained neural network on the data.

Tools used: Python toolchain, Pandas, NLTK, and Keras.

4. The Tabular ML'er (AKA the XGBoost Specialist)

This type of Data Scientist can train multiple algorithms and stack models and optimize the heck out of them. These guys have deep expertise with running and optimizing standard algorithms like XGBoost, Ridge Regression, and (nowadays) Keras models.

Tools used: Python, R, XGB, and Keras.

5. The Old School ML-er

He or she is similar to the one above, but they're not limited to categorical models. He or she is very good at feature engineering. This was the only Machine Learning expertise until the newer Deep Learning stuff came up.

Tools used: C++, Python, and Scikit Learn.

6. The Deep Learning Guy

This type of Data Scientist needs a GPU system and a well-tagged dataset, needs to try out architectures, and does no feature engineering. They'll spend a lot of time trying architectures and minimal in feature engineering — but the accuracy will be insane!

Tools used: Python, Theano, Tensorflow, and high-level libraries like Keras.

7. The Domain Specialist

He or she knows a lot about the domain and knows some things about linear models. This Data Scientist codes the domain information and trains a linear algorithm on top. This description includes mechanical engineers, analysts at different firms, and scientists in pure and applied sciences.

Tools used: Matlab, C++/Fortran, and R/Python — but different specialists will use very different tools.

8. The Newbie

The intern. Will evolve into whichever of the seven categories his or her mentor belongs.

Most data companies will have all of these types of Data Scientists. Which category do you fall into?

Check out the Exaptive data application Studio. Technology agnostic. No glue code. Use what you know and rely on the community for what you don't. Try the community version.

Topics: DATA SCIENTISTS, BIG DATA, CAREER

Published at DZone with permission of Muktabh Srivastava. See the original article here.
Opinions expressed by DZone contributors are their own.