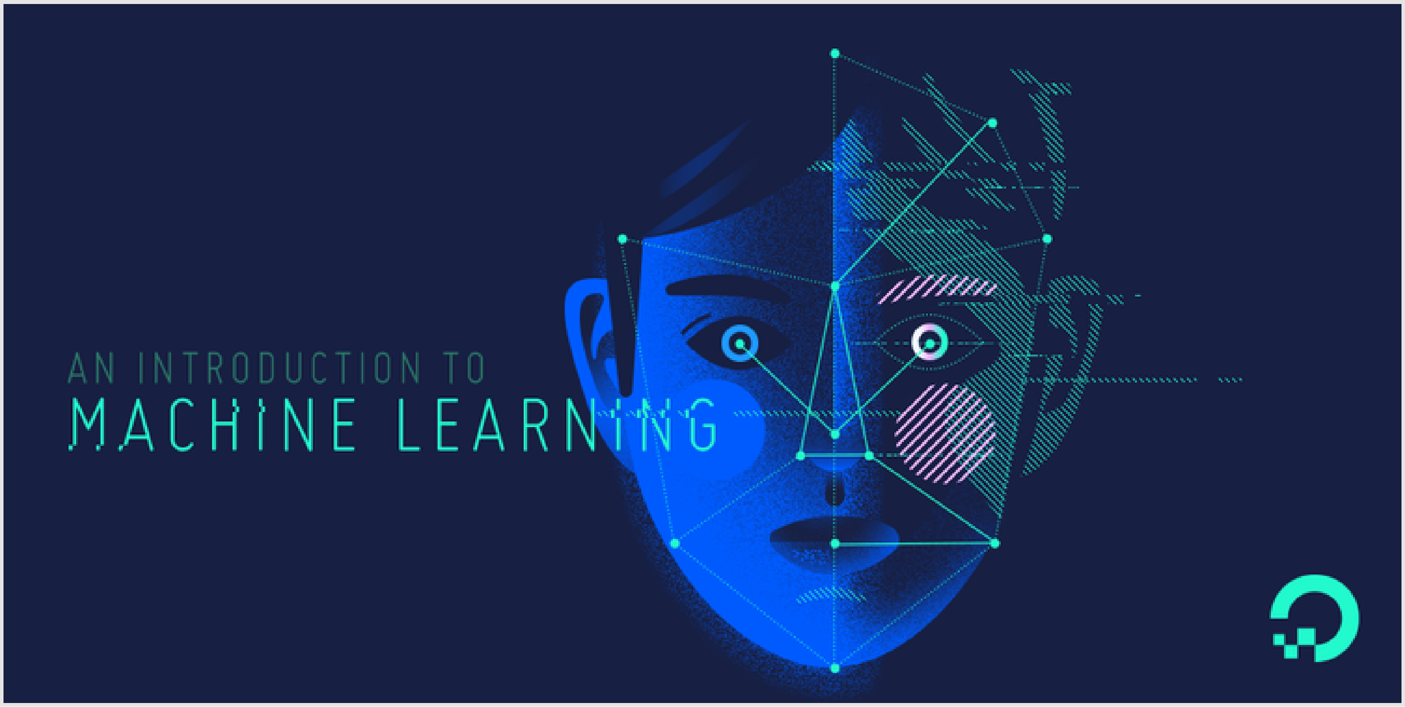
Subscribe Share Contents



An Introduction to Machine Learning

Posted September 28, 2017 112.6k MACHINE LEARNING DEVELOPMENT CONCEPTUAL

52

By: Lisa Tagliaferri

Introduction

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people.

Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead computers to train on data inputs and use statistical analysis in order to output values that fall within

SCROLL TO TOP

a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

Any technology user today has benefitted from machine learning. Facial recognition technology allows social media platforms to help users tag and share photos of friends. Optical character recognition (OCR) technology converts images of text into movable type. Recommendation engines, powered by machine learning, suggest what movies or television shows to watch next based on user preferences. Self-driving cars that rely on machine learning to navigate may soon be available to consumers.

Machine learning is a continuously developing field. Because of this, there are some considerations to keep in mind as you work with machine learning methodologies, or analyze the impact of machine learning processes.

In this tutorial, we'll look into the common machine learning methods of supervised and unsupervised learning, and common algorithmic approaches in machine learning, including the k-nearest neighbor algorithm, decision tree learning, and deep learning. We'll explore which programming languages are most used in machine learning, providing you with some of the positive and negative attributes of each. Additionally, we'll discuss biases that are perpetuated by machine learning algorithms, and consider what can be kept in mind to prevent these biases when building algorithms.

Machine Learning Methods

In machine learning, tasks are generally classified into broad categories. These categories are based on how learning is received or how feedback on the learning is given to the system developed.

Two of the most widely adopted machine learning methods are **supervised learning** which trains algorithms based on example input and output data that is labeled by humans, and **unsupervised learning** which provides the algorithm with no labeled data in order to allow it to find structure within its input data. Let's explore these methods in more detail.

Supervised Learning

In supervised learning, the computer is provided with example inputs that are labeled with their desired outputs. The purpose of this method is for the algorithm to be able to "learn" by comparing its actual output with the "taught" outputs to find errors, and modify the model accordingly. Supervised learning therefore uses patterns to predict label values on additional unlabeled data.

SCROLL TO TOP

For example, with supervised learning, an algorithm may be fed data with images of sharks labeled as `fish` and images of oceans labeled as `water`. By being trained on this data, the supervised learning algorithm should be able to later identify unlabeled shark images as `fish` and unlabeled ocean images as `water`.

A common use case of supervised learning is to use historical data to predict statistically likely future events. It may use historical stock market information to anticipate upcoming fluctuations, or be employed to filter out spam emails. In supervised learning, tagged photos of dogs can be used as input data to classify untagged photos of dogs.

Unsupervised Learning

In unsupervised learning, data is unlabeled, so the learning algorithm is left to find commonalities among its input data. As unlabeled data are more abundant than labeled data, machine learning methods that facilitate unsupervised learning are particularly valuable.

The goal of unsupervised learning may be as straightforward as discovering hidden patterns within a dataset, but it may also have a goal of feature learning, which allows the computational machine to automatically discover the representations that are needed to classify raw data.

Unsupervised learning is commonly used for transactional data. You may have a large dataset of customers and their purchases, but as a human you will likely not be able to make sense of what similar attributes can be drawn from customer profiles and their types of purchases. With this data fed into an unsupervised learning algorithm, it may be determined that women of a certain age range who buy unscented soaps are likely to be pregnant, and therefore a marketing campaign related to pregnancy and baby products can be targeted to this audience in order to increase their number of purchases.

Without being told a “correct” answer, unsupervised learning methods can look at complex data that is more expansive and seemingly unrelated in order to organize it in potentially meaningful ways. Unsupervised learning is often used for anomaly detection including for fraudulent credit card purchases, and recommender systems that recommend what products to buy next. In unsupervised learning, untagged photos of dogs can be used as input data for the algorithm to find likenesses and classify dog photos together.

Approaches

As a field, machine learning is closely related to computational statistics, so having a background knowledge in statistics is useful for understanding and leveraging machine learning algorithms.

SCROLL TO TOP

For those who may not have studied statistics, it can be helpful to first define correlation and regression, as they are commonly used techniques for investigating the relationship among quantitative variables. **Correlation** is a measure of association between two variables that are not designated as either dependent or independent. **Regression** at a basic level is used to examine the relationship between one dependent and one independent variable. Because regression statistics can be used to anticipate the dependent variable when the independent variable is known, regression enables prediction capabilities.

Approaches to machine learning are continuously being developed. For our purposes, we'll go through a few of the popular approaches that are being used in machine learning at the time of writing.

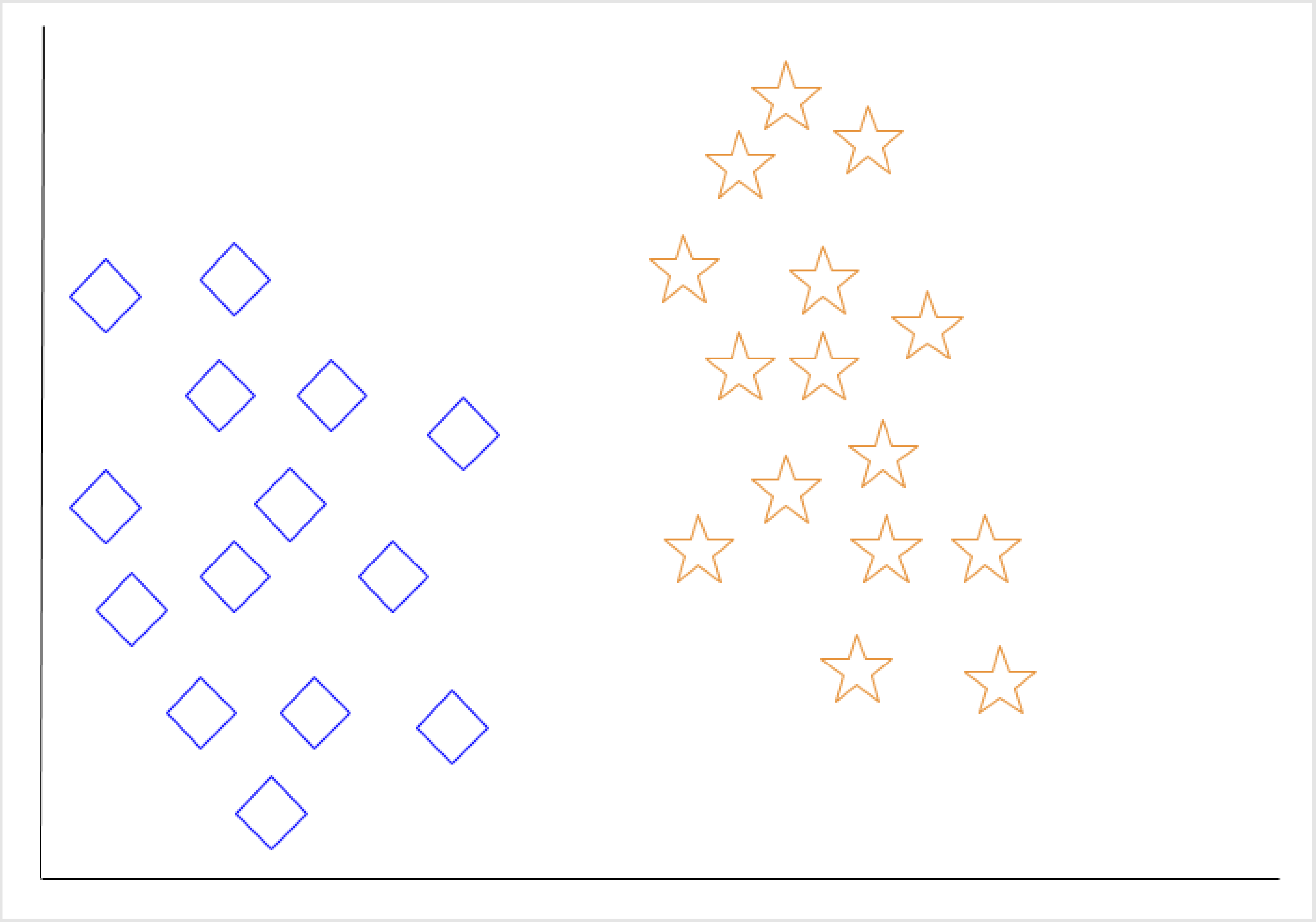
k-nearest neighbor

The k-nearest neighbor algorithm is a pattern recognition model that can be used for classification as well as regression. Often abbreviated as k-NN, the **k** in k-nearest neighbor is a positive integer, which is typically small. In either classification or regression, the input will consist of the k closest training examples within a space.

We will focus on k-NN classification. In this method, the output is class membership. This will assign a new object to the class most common among its k nearest neighbors. In the case of k = 1, the object is assigned to the class of the single nearest neighbor.

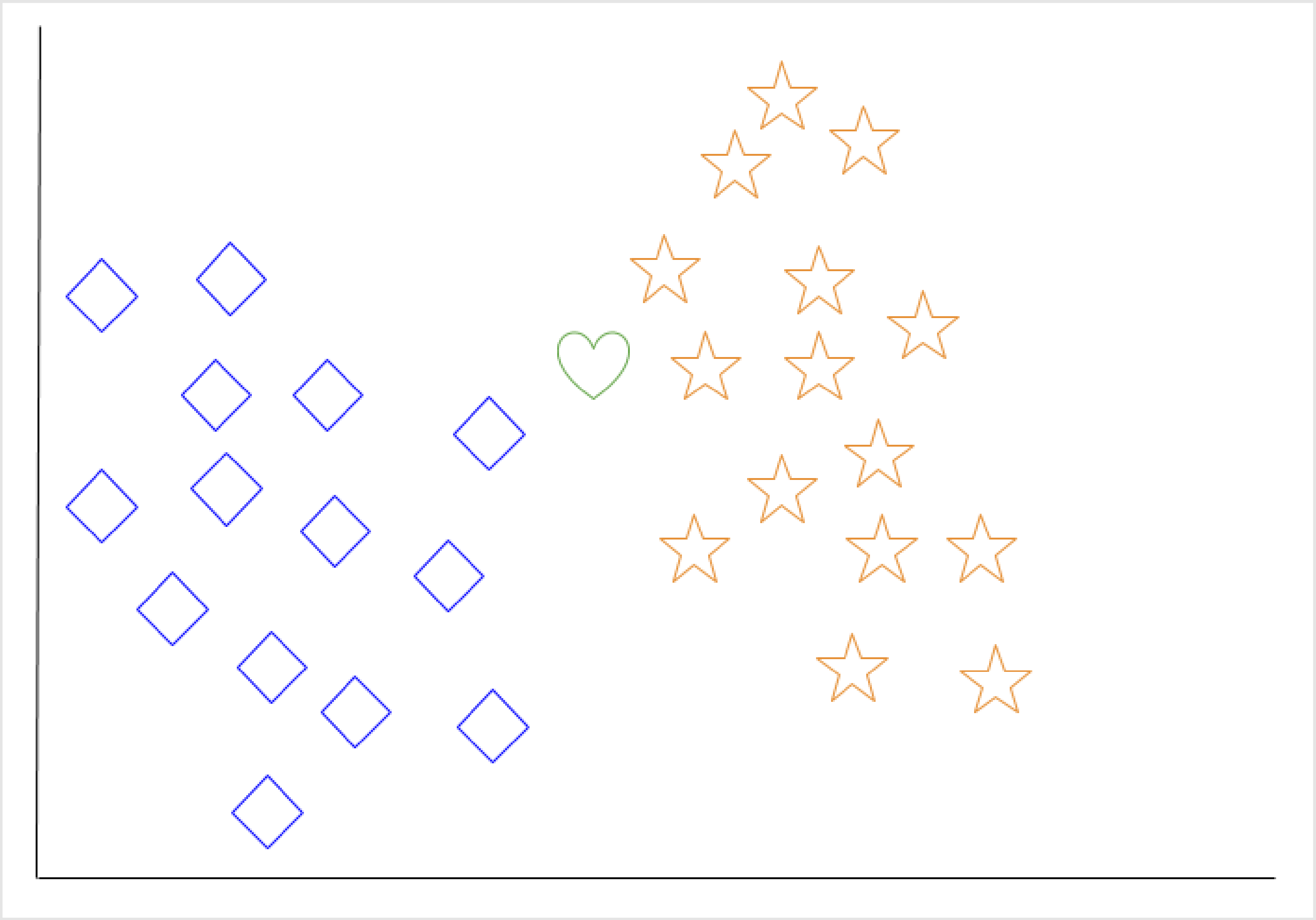
Let's look at an example of k-nearest neighbor. In the diagram below, there are blue diamond objects and orange star objects. These belong to two separate classes: the diamond class and the star class.

SCROLL TO TOP



When a new object is added to the space — in this case a green heart — we will want the machine learning algorithm to classify the heart to a certain class.

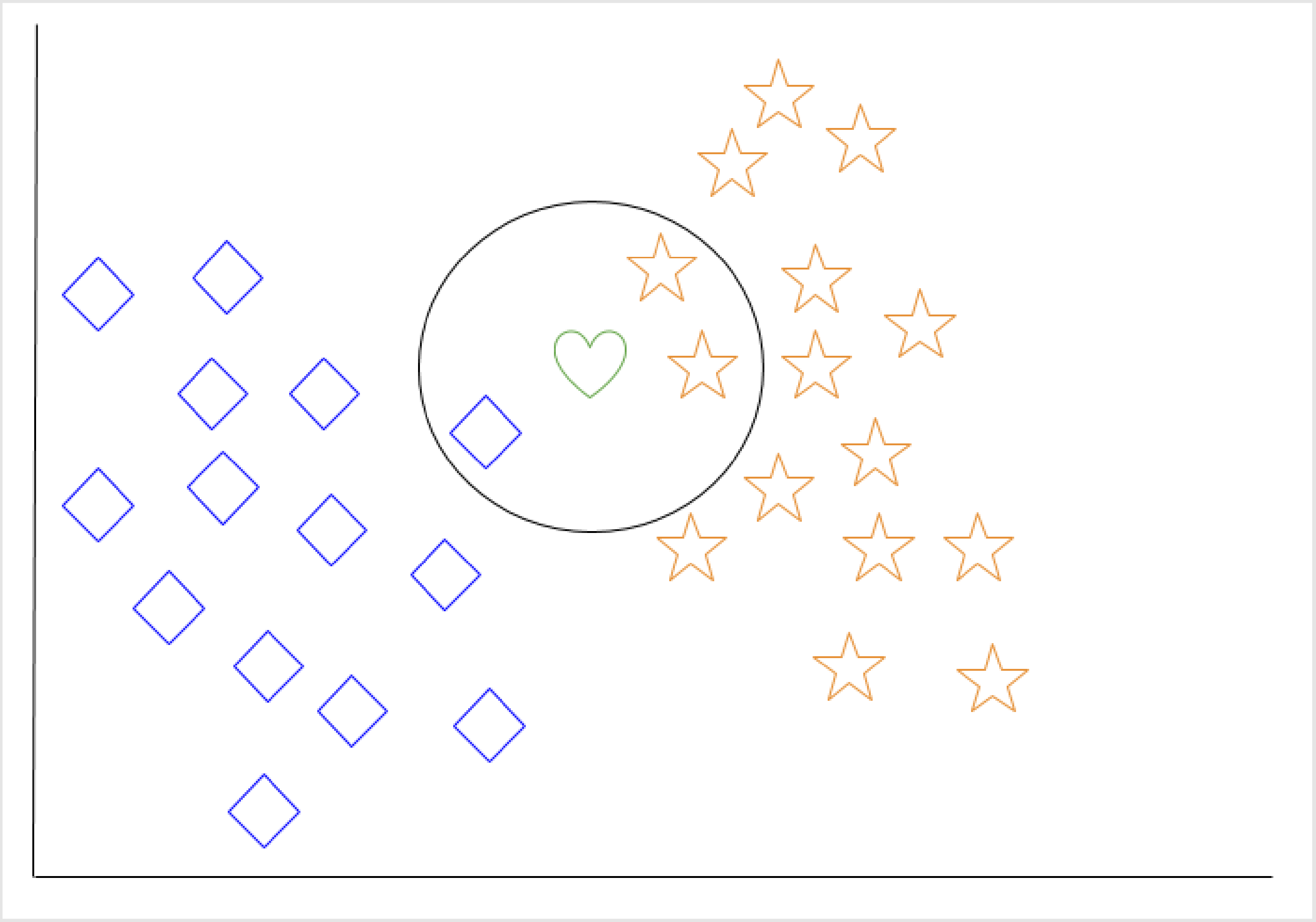
SCROLL TO TOP



When we choose k = 3, the algorithm will find the three nearest neighbors of the green heart in order to classify it to either the diamond class or the star class.

In our diagram, the three nearest neighbors of the green heart are one diamond and two stars. Therefore, the algorithm will classify the heart with the star class.

SCROLL TO TOP



Among the most basic of machine learning algorithms, k-nearest neighbor is considered to be a type of “lazy learning” as generalization beyond the training data does not occur until a query is made to the system.

Decision Tree Learning

For general use, decision trees are employed to visually represent decisions and show or inform decision making. When working with machine learning and data mining, decision trees are used as a predictive model. These models map observations about data to conclusions about the data’s target value.

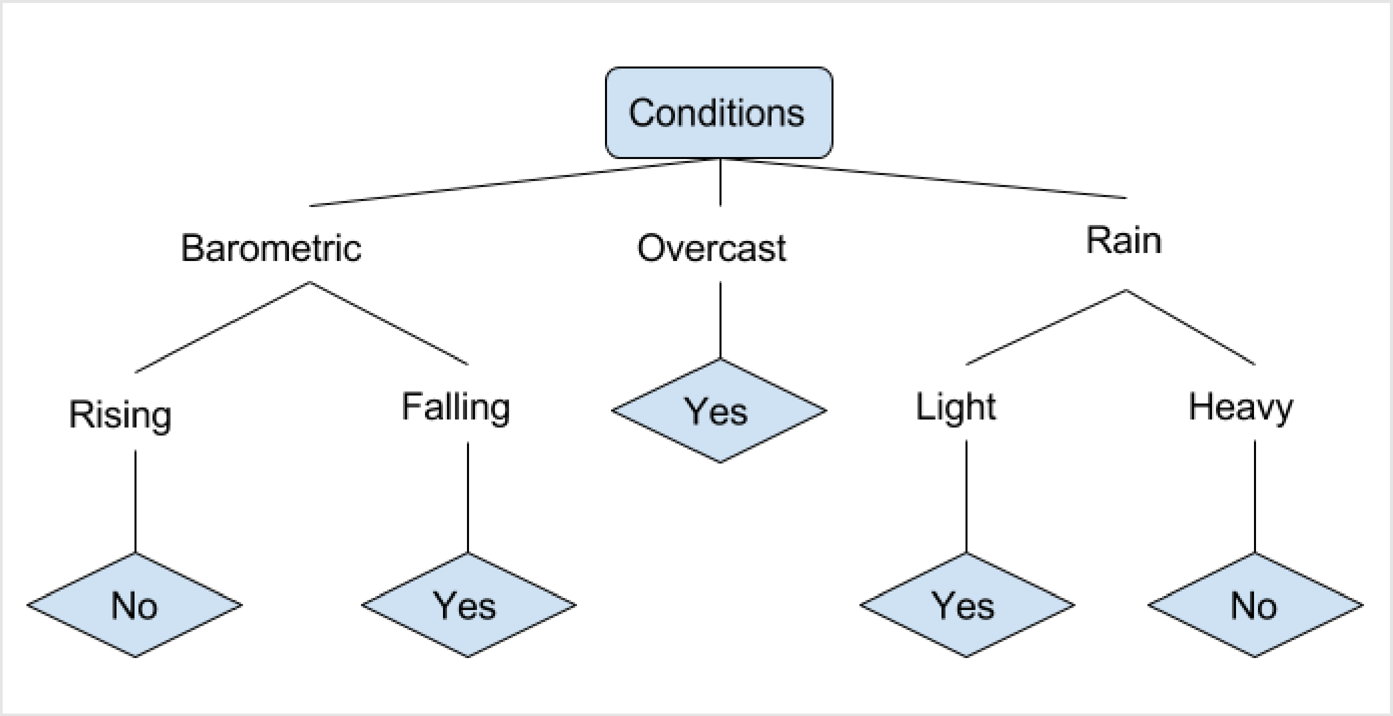
The goal of decision tree learning is to create a model that will predict the value of a target based on input variables.

In the predictive model, the data’s attributes that are determined through observation are represented by the branches, while the conclusions about the data’s target value are represented in the leaves.

SCROLL TO TOP

When “learning” a tree, the source data is divided into subsets based on an attribute value test, which is repeated on each of the derived subsets recursively. Once the subset at a node has the equivalent value as its target value has, the recursion process will be complete.

Let’s look at an example of various conditions that can determine whether or not someone should go fishing. This includes weather conditions as well as barometric pressure conditions.



In the simplified decision tree above, an example is classified by sorting it through the tree to the appropriate leaf node. This then returns the classification associated with the particular leaf, which in this case is either a `Yes` or a `No`. The tree classifies a day’s conditions based on whether or not it is suitable for going fishing.

A true classification tree data set would have a lot more features than what is outlined above, but relationships should be straightforward to determine. When working with decision tree learning, several determinations need to be made, including what features to choose, what conditions to use for splitting, and understanding when the decision tree has reached a clear ending.

Deep Learning

Deep learning attempts to imitate how the human brain can process light and sound stimuli into vision and hearing. A deep learning architecture is inspired by biological neural networks and consists of multiple layers in an artificial neural network made up of hardware and GPUs.

Deep learning uses a cascade of nonlinear processing unit layers in order to extra
features (or representations) of the data. The output of one layer serves as the inp

SCROLL TO TOP

successive layer. In deep learning, algorithms can be either supervised and serve to classify data, or unsupervised and perform pattern analysis.

Among the machine learning algorithms that are currently being used and developed, deep learning absorbs the most data and has been able to beat humans in some cognitive tasks. Because of these attributes, deep learning has become the approach with significant potential in the artificial intelligence space

Computer vision and speech recognition have both realized significant advances from deep learning approaches. IBM Watson is a well-known example of a system that leverages deep learning.

Programming Languages

When choosing a language to specialize in with machine learning, you may want to consider the skills listed on current job advertisements as well as libraries available in various languages that can be used for machine learning processes.

From data taken from job ads on indeed.com in December 2016, it can be inferred that Python is the most sought-for programming language in the machine learning professional field. Python is followed by Java, then R, then C++.

Python's popularity may be due to the increased development of deep learning frameworks available for this language recently, including TensorFlow, PyTorch, and Keras. As a language that has readable syntax and the ability to be used as a scripting language, Python proves to be powerful and straightforward both for preprocessing data and working with data directly. The scikit-learn machine learning library is built on top of several existing Python packages that Python developers may already be familiar with, namely NumPy, SciPy, and Matplotlib.

To get started with Python, you can read our tutorial series on "How To Code in Python 3," or read specifically on "How To Build a Machine Learning Classifier in Python with scikit-learn" or "How To Perform Neural Style Transfer with Python 3 and PyTorch."

Java is widely used in enterprise programming, and is generally used by front-end desktop application developers who are also working on machine learning at the enterprise level. Usually it is not the first choice for those new to programming who want to learn about machine learning, but is favored by those with a background in Java development to apply to machine learning. In terms of machine learning applications in industry, Java tends to be used more than Python for network security, including in cyber attack and fraud detection use cases.

Among machine learning libraries for Java are Deeplearning4j, an open-source a

SCROLL TO TOP

deep-learning library written for both Java and Scala; MALLET (**MA**chine **L**earning for Languag**E**

Toolkit) allows for machine learning applications on text, including natural language processing, topic modeling, document classification, and clustering; and Weka, a collection of machine learning algorithms to use for data mining tasks.

R is an open-source programming language used primarily for statistical computing. It has grown in popularity over recent years, and is favored by many in academia. R is not typically used in industry production environments, but has risen in industrial applications due to increased interest in data science. Popular packages for machine learning in R include caret (short for **C**lassification **A**nd **RE**gression **T**raining) for creating predictive models, randomForest for classification and regression, and e1071 which includes functions for statistics and probability theory.

C++ is the language of choice for machine learning and artificial intelligence in game or robot applications (including robot locomotion). Embedded computing hardware developers and electronics engineers are more likely to favor C++ or C in machine learning applications due to their proficiency and level of control in the language. Some machine learning libraries you can use with C++ include the scalable mlpack, Dlib offering wide-ranging machine learning algorithms, and the modular and open-source Shark.

Human Biases

Although data and computational analysis may make us think that we are receiving objective information, this is not the case; being based on data does not mean that machine learning outputs are neutral. Human bias plays a role in how data is collected, organized, and ultimately in the algorithms that determine how machine learning will interact with that data.

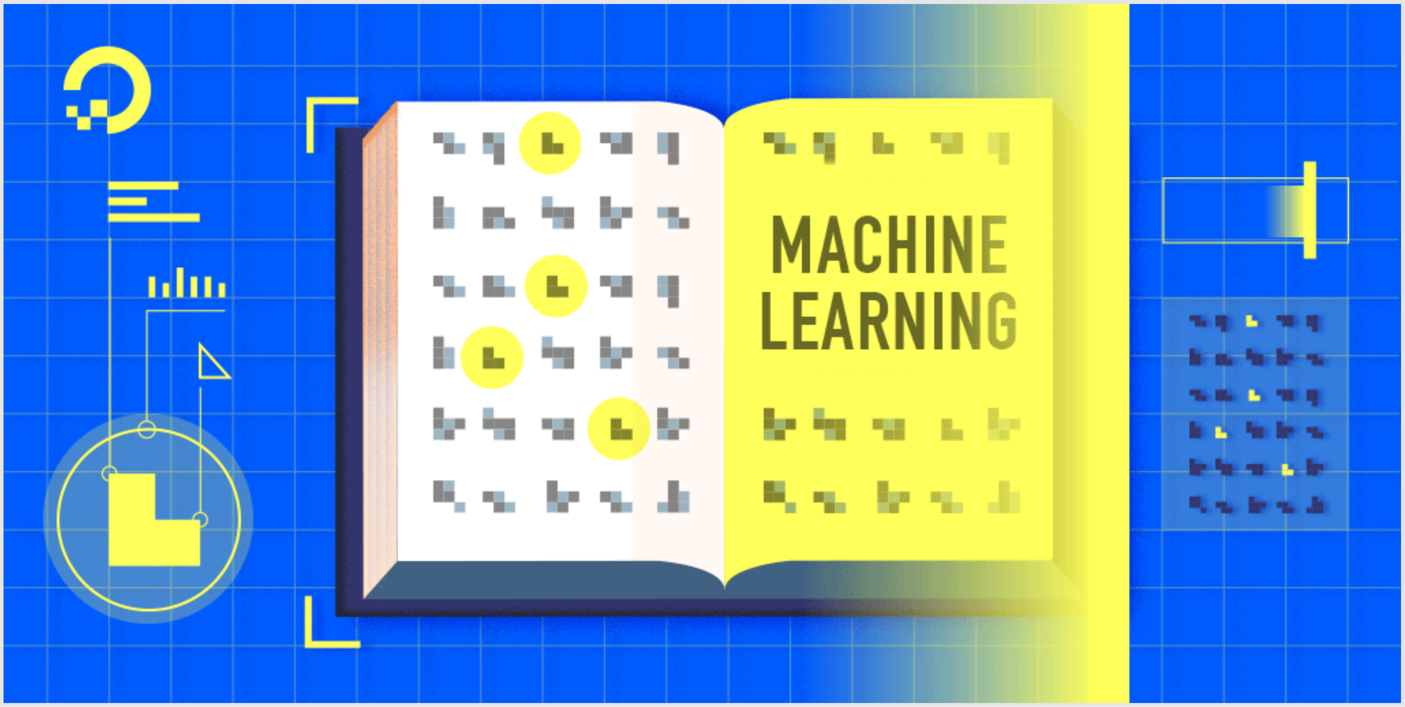
If, for example, people are providing images for “fish” as data to train an algorithm, and these people overwhelmingly select images of goldfish, a computer may not classify a shark as a fish. This would create a bias against sharks as fish, and sharks would not be counted as fish.

When using historical photographs of scientists as training data, a computer may not properly classify scientists who are also people of color or women. In fact, recent peer-reviewed research has indicated that AI and machine learning programs exhibit human-like biases that include race and gender prejudices. See, for example “Semantics derived automatically from language corpora contain human-like biases” and “Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints” [PDF].

As machine learning is increasingly leveraged in business, uncaught biases can perpetuate systemic issues that may prevent people from qualifying for loans, from being shown ads for high-paying job opportunities, or from receiving same-day delivery options.

SCROLL TO TOP

Because human bias can negatively impact others, it is extremely important to be aware of it, and to also work towards eliminating it as much as possible. One way to work towards achieving this is by ensuring that there are diverse people working on a project and that diverse people are testing and reviewing it. Others have called for regulatory third parties to monitor and audit algorithms, building alternative systems that can detect biases, and ethics reviews as part of data science project planning. Raising awareness about biases, being mindful of our own unconscious biases, and structuring equity in our machine learning projects and pipelines can work to combat bias in this field.



Conclusion

This tutorial reviewed some of the use cases of machine learning, common methods and popular approaches used in the field, suitable machine learning programming languages, and also covered some things to keep in mind in terms of unconscious biases being replicated in algorithms.

Because machine learning is a field that is continuously being innovated, it is important to keep in mind that algorithms, methods, and approaches will continue to change.

In addition to reading our tutorials on “How To Build a Machine Learning Classifier in Python with scikit-learn” or “How To Perform Neural Style Transfer with Python 3 and PyTorch,” you can learn more about working with data in the technology industry by reading our Data Analysis tutorials.

By: Lisa Tagliaferri

♡ Upvote (52) Sub SCROLL TO TOP

Related Tutorials

How To Install the Anaconda Python Distribution on Ubuntu 18.04

How To Modify Attributes, Classes, and Styles in the DOM

How to Deploy Elixir-Phoenix Applications with MySQL on Ubuntu 16.04

Understanding Classes in JavaScript

How To Create Django Views

26 Comments

Leave a comment...

SCROLL TO TOP

Log In to Comment

1ikb3zz *October 1, 2017*

1 thanks a lot for the article!

review *October 12, 2017*

1 Thanks Lisa for the article!

emmaezenwere *October 14, 2017*

0 Does a digital ocean Linux server have the Dlib library preinstalled? If not, does it support Dlib?

kubisztal *October 16, 2017*

1 Hey! Thanks for such good article! Really useful :)

trevorweir *October 17, 2017*

0 Well, Lisa, this is a fabulous topic and though your writing is exemplary, you somehow lost me somewhere between paragraph 6 and 9 (not a difficult thing to do apparently).

I struggled on for a few more paragraphs but had to give it up. Guess, maybe my last 3 girlfriends were right, perhaps I really only do read 'pictures', lol

Anyway, quite stupidly, I decided that maybe you put the pictures at the end. So, I rapidly started reading from the last paragraph coming backwards, ostensibly looking for the pictures.

Ha ha, that eased the boredom and I quickly got back to the same paragraph where i got bogged down like molasses - but alas, no pictures ...

So, now I see that according to indeed.com (2016 ,Dec) Python, Java, R and c++ are amongst the most used ML languages, perhaps you could redo this article when you have time in the future and sprinkle it with quicky examples that we noobies can quickly throw up in a droplet and really try to dive into in this exciting field in five plus minutes. Yeah, you can see my ADD talking here.

If it can't do something with incredible speed and brilliance in 5 minutes, it thinks its probably not worth doing, lol.

SCROLL TO TOP

In truth, perhaps you could actually put those same examples directly into the Machine Learning AI image, so that starting the machine instance will give us some of those same examples to quickly look at.

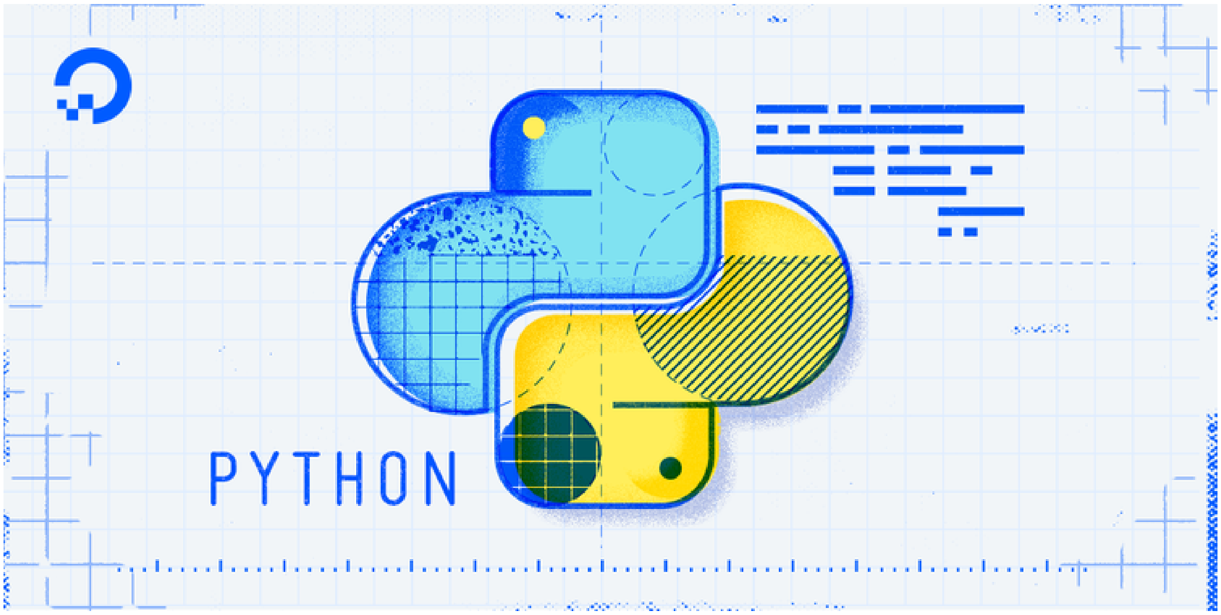
And yes, I booted up a droplet to see what it (the empty Jupyter Notebook) would looked like and for the noobies amongst us (which includes me) it's like looking at a blank page with nothing on it.

Edit: There are pictures here. Have no idea how I missed them the first time thru.

ltagliaferri MOD *October 17, 2017*

0 Hi Trevor, you may be interested in our step-by-step guides that address machine learning topics. These include the following tutorials:

- "How To Use the Machine Learning One-Click Install Image on DigitalOcean"
- "How To Set Up Jupyter Notebook for Python 3"
- "How To Build a Machine Learning Classifier in Python with Scikit-learn"
- "How To Perform Neural Style Transfer with Python 3 and PyTorch"



How To Set Up Jupyter Notebook for Python 3

This tutorial will walk you through setting up Jupyter Notebook to run either locally or from an Ubuntu 16.04 server, as well as teach you how to connect to and use the notebook. Jupyter notebooks (or simply notebooks) are documents produced by the Jupyter

SCROLL TO TOP

pranitpatil0503 *November 20, 2017*

1 Great introduction to java machine learning! I am java beginner and article like this which is explaining the concept so deeply which is incredibly helpful for us as beginners. Thanks for providing a great insight, now i can say java learning is so important. thanks again!

---

leviya *November 27, 2017*

1 hi Lisa Tagliaferri,

It is really a great work and the way in which u r sharing the knowledge is excellent. Thanks for helping me to understand basic concepts and more. As a beginner in Machine Learning your post help me a lot. Thanks for your informative article.

---

printerr45 *December 6, 2017*

0 Good way of presentation hp printer dubai of the subject.Thank you for added up my knowledge.

---

tysontom93 *December 6, 2017*

0 This article is excellent.Payroll software Dubai really loved this.Your article nailed the essence of machine learning.

---

yuvanasav *December 12, 2017*

0 This article is very useful for beginner,machine learning is one of the fastest growing technology in computer science world, with far-reaching applications.I would like to know, is python skill mandatory to learn machine learning?

---

misan128 *December 13, 2017*

1 Awesome article, I'm willing to learn about Machine Learning and this is a great start point to know what this is about. Thanks!

---

ltagliaferri MOD *December 13, 2017*

0 Awesome, glad to hear this was useful as you begin to investigate machine learning!

---

bernicestockstill *January 31, 2018*

0 Cool) and when is the beginning?

SCROLL TO TOP

MadhuSK *February 21, 2018*

0 Hi Lisa,
I must say that this is very informative blog. I am also working on some applications that use machine learning algorithm. I would like to add as well.

Following are some of the easy-to-understand & publicly used examples of machine learning & AI-
-Chat bots
-Voice recognition and responsive system.
The above two, learn to refine the response time to time. The older these system gets the better are the responses.

---

shawwesley *February 21, 2018*

0 Perfect for beginners as for me)

---

regianefolter *February 21, 2018*

0 Hey, that's a great content!! Thanks for sharing your knowledge. If you are up to this kind of resources, you may find this machine learning tutorial interesting: http://bit.ly/2BHHeYB - I'd appreciate your feedback!

FredVasco *February 26, 2018*

0 I like this one)

---

ictseoexpert *March 6, 2018*

0 Really helpfull article for machine learning. Payroll software UAE appreciate your work.

---

angelpokeronline *March 11, 2018*

0 deep dive understanding about machine learning,
will keep reading your article Lisa!

agen bola
jasa seo

SCROLL TO TOP

solodeji *March 25, 2018*

0 Great and outstanding article, I'm willing to learn about Machine Learning and this is a great starting point for me to know what this is about. Thanks!

Load More Comments

Community Tutorials Questions Projects Tags Newsletter RSS

Distros & One-Click Apps Terms, Privacy, & Copyright Security Report a Bug Write for DOnations Shop

SCROLL TO TOP