Supervised Learning

By: IBM Cloud Education

19 August 2020

Artificial intelligence

Supervised Learning

Learn how supervised learning works and how it can be used to build

Feature product

Watson Studio

SPSS Statistics

IBM Cloud Pak for Data

Site feedback

highly accurate machine learning models.

What is supervised learning?

Supervised learning, also known as supervised machine learning, is a subcategory of machine learning and artificial intelligence. It is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately, which occurs as part of the cross validation process. Supervised learning helps organizations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox.

Related link

Linear regression

Logistic regression

Data Science

How supervised learning works

Supervised learning uses a training set to teach models to yield the desired output. This training dataset includes inputs and correct outputs, which allow the model to learn over time. The algorithm measures its accuracy through the loss function, adjusting until the error has been sufficiently minimized.

Supervised learning can be separated into two types of problems when data mining—classification and regression:

- **Classification** uses an algorithm to accurately assign test data into specific categories. It recognizes specific entities within the dataset and attempts to draw some conclusions on how those entities should be

labeled or defined. Common classification algorithms are linear classifiers, support vector machines (SVM), decision trees, k-nearest neighbor, and random forest, which are described in more detail below.

- **Regression** is used to understand the relationship between dependent and independent variables. It is commonly used to make projections, such as for sales revenue for a given business. Linear regression, logistical regression, and polynomial regression are popular regression algorithms.

Supervised learning algorithms

Various algorithms and computation techniques are used in supervised machine

learning processes. Below are brief explanations of some of the most commonly used learning methods, typically calculated through use of programs like R or Python:

Neural networks

Primarily leveraged for deep learning algorithms, neural networks process training data by mimicking the interconnectivity of the human brain through layers of nodes. Each node is made up of inputs, weights, a bias (or threshold), and an output. If that output value exceeds a given threshold, it “fires” or activates the node, passing data to the next layer in the network. Neural networks learn this mapping function through supervised learning, adjusting based on the loss function through the process of gradient descent. When the cost function is at or near zero, we can be confident in the model’s accuracy to yield the correct answer.

Naive Bayes

Naive Bayes is classification approach that adopts the principle of class conditional independence

from the Bayes Theorem. This means that the presence of one feature does not impact the presence of another in the probability of a given outcome, and each predictor has an equal effect on that result. There are three types of Naïve Bayes classifiers: Multinomial Naïve Bayes, Bernoulli Naïve Bayes, and Gaussian Naïve Bayes. This technique is primarily used in text classification, spam identification, and recommendation systems.

Linear regression

Linear regression is used to identify the relationship between a dependent variable and one or more independent variables and is typically leveraged to make predictions about future outcomes. When there is only one independent variable and one dependent variable, it is known as simple linear regression. As the number of independent variables increases, it is referred to as multiple linear regression. For each type of linear regression, it seeks to plot a line of best fit, which is calculated through the method of least squares. However, unlike other regression models, this line is straight when plotted on a graph.

Logistic regression

While linear regression is leveraged when dependent variables are continuous, logistical regression is selected when the dependent variable is categorical, meaning they have binary outputs, such as "true" and "false" or "yes" and "no." While both regression models seek to understand relationships between data inputs, logistic regression is mainly used to solve binary classification problems, such as spam identification.

Support vector machine (SVM)

A support vector machine is a popular supervised learning model developed by Vladimir Vapnik, used for both data classification and regression. That said, it is typically leveraged for classification problems, constructing a hyperplane where the distance between two classes of data points is at its maximum. This hyperplane is known as the decision boundary, separating the classes of data points (e.g., oranges vs. apples) on either side of the plane.

K-nearest neighbor

K-nearest neighbor, also known as the KNN algorithm, is a non-parametric algorithm that classifies

data points based on their proximity and association to other available data. This algorithm assumes that similar data points can be found near each other. As a result, it seeks to calculate the distance between data points, usually through Euclidean distance, and then it assigns a category based on the most frequent category or average.

Its ease of use and low calculation time make it a preferred algorithm by data scientists, but as the test dataset grows, the processing time lengthens, making it less appealing for classification tasks. KNN is typically used for recommendation engines and image recognition.

Random forest

Random forest is another flexible supervised machine learning algorithm used for both classification and regression purposes. The "forest" references a collection of uncorrelated decision trees, which are then merged together to reduce variance and create more accurate data predictions.

Unsupervised vs. supervised vs. semi-supervised learning

Unsupervised machine learning and supervised machine learning are frequently discussed together. Unlike supervised learning, unsupervised learning uses unlabeled data. From that data, it discovers patterns that help solve for clustering or association problems. This is particularly useful when subject matter experts are unsure of common properties within a data set. Common clustering algorithms are hierarchical, k-means, and Gaussian mixture models.

Semi-supervised learning occurs when only part of the given input

data has been labeled. Unsupervised and semi-supervised learning can be more appealing alternatives as it can be time-consuming and costly to rely on domain expertise to label data appropriately for supervised learning.

For a deep dive into the differences between these approaches, check out "Supervised vs. Unsupervised Learning: What's the Difference?"

Supervised learning examples

Supervised learning models can be used to build and advance a number of business applications, including the following:

- **Image- and object-recognition:** Supervised learning algorithms can be used to locate, isolate, and categorize objects out of videos or images, making them

useful when applied to various computer vision techniques and imagery analysis.

- **Predictive analytics:** A widespread use case for supervised learning models is in creating predictive analytics systems to provide deep insights into various business data points. This allows enterprises to anticipate certain results based on a given output variable, helping business leaders justify decisions or pivot for the benefit of the organization.

- **Customer sentiment analysis:** Using supervised machine learning algorithms, organizations can extract and classify important pieces of information from large volumes of data—including context, emotion, and intent—with very little human intervention. This can be incredibly useful when gaining a better understanding of customer interactions and can be used to improve brand engagement efforts.

- **Spam detection:** Spam detection is another example of a supervised learning model. Using supervised classification algorithms, organizations can

train databases to recognize patterns or anomalies in new data to organize spam and non-spam-related correspondences effectively.

Challenges of supervised learning

Although supervised learning can offer businesses advantages, such as deep data insights and improved automation, there are some challenges when building sustainable supervised learning models. The following are some of these challenges:

- Supervised learning models can require certain levels of expertise to structure accurately.

- Training supervised learning models can be very time

intensive.

- Datasets can have a higher likelihood of human error, resulting in algorithms learning incorrectly.

- Unlike unsupervised learning models, supervised learning cannot cluster or classify data on its own.

Supervised learning and IBM

Supervised learning models can be a valuable solution for eliminating manual classification work and for making future predictions based on labeled data. However, formatting your machine learning algorithms requires human knowledge and expertise to avoid overfitting data models.

IBM and its data science and AI teams have spent years perfecting the development and deployment

of supervised learning models with numerous business use cases. With the help of powerful tools such as IBM Watson Studio on IBM Cloud Pak for Data, organizations can create highly scalable machine learning models regardless of where their data lives, all while being supported by IBM's robust hybrid multicloud environment.

For more information on how IBM can help you create your own supervised machine learning models, explore IBM Watson Studio.

Sign up for an IBMid and create your IBM Cloud account.

Why IBM Cloud

Why IBM Cloud

Hybrid Cloud approach

Trust and security

Open Cloud

Data centers

Case studies

Products and Solutions

Cloud Paks

Cloud pricing

View all products

View all solutions

Learn about

What is Hybrid Cloud?

What is Cloud Computing?

What is Confidential Computing?

What is a Data Lake?

What is a Data Warehouse?

What is Artificial Intelligence (AI)?

What is Machine Learning?

What is DevOps?

Resources

Get started

Docs

Architectures

IBM Garage

Training and Certifications

Partners

Cloud blog

Hybrid Cloud careers

My Cloud account

Let’s talk