Data science terminology

Application Programming Interface (API)

Collection of software routines, protocols, and tools which provide a programmer with all the building blocks for developing an application program for a specific platform (environment). An API also provides an interface that allows a program to communicate with other programs, running in the same environment. (BusinessDictionary.com)

Artificial Intelligence (AI)

Artificial intelligence is a field of computer science dedicated to solving cognitive problems commonly associated with human intelligence such as learning, problem solving, visual perception and speech and pattern recognition.

Artificial Intelligence System

A technological system that uses a model to make inferences to generate output, including predictions, recommendations or decisions.

Corpus

In linguistics, corpus is referred to as a large and structured set of texts. In the context of topic modelling, a corpus is a set of documents and each document is viewed as a mixture of topics that are present in the corpus. (wikipedia.org)

Data Science

Data Science is an interdisciplinary field that uses scientific methods and algorithms to extract information and insights from diverse data types. It combines domain expertise, programming skills and knowledge of mathematics and statistics to solve analytically complex problems.

Deep Learning

Subset of machine learning that imitates the workings of the human brain in processing data and improves performance. Typically, a multi-level algorithm that gradually identifies things at higher levels of abstraction. For example, the first level may identify certain lines, then the next level identifies combinations of lines as shapes, and then the next level identifies combinations of shapes as specific objects. Deep learning is popular for image classification. (www.datascienceglossary.org)

Event

An event in the Unified Modeling Language (UML) is a notable occurrence at a particular point in time. Events can, but do not necessarily, cause state transitions from one state to another in state machines represented. (wikipedia.org)

Latent variables

Latent variables are variables that are not directly observed but are rather inferred (through a mathematical model) from other variables that are observed (directly measured). Mathematical models that aim to explain observed variables in terms of latent variables are called latent variable models. (datascienceglossary.org)

Machine Learning (ML)

"Machine learning is the science of getting computers to automatically learn from experience instead of relying on explicitly programmed rules, and generalize the acquired knowledge to new settings."

United Nations Economic Commission for Europe's Machine Learning Team (2018 report)

The use of machine learning in official statistics.

In essence, Machine Learning automates analytical model building through optimization algorithms and parameters that can be modified and fine-tuned.

Machine Learning Algorithms

Machine learning algorithms use computational methods to "learn" information directly from data without relying on a predetermined equation as a model. The algorithms adaptively improve their performance as the number of samples available for learning increases. (Mathworks.com)

Machine Learning Model

A digital representation of patterns identified in data through automated processing using an algorithm designed to enable the recognition or replication of those patterns.

Natural Language Processing (NLP)

Natural language processing (NLP) is a method to translate between computer and human languages. It is a method of getting a computer to understandably read a line of text without the computer being fed some sort of clue or calculation. In other words, NLP automates the translation process between computers and humans. (techopedia.com)

One-hot vector

In NLP, a one-hot vector is a 1 x N matrix (vector), made of 0 and 1, used to distinguish each word in a vocabulary from every other word in the vocabulary. One-hot encoding ensures that machine learning does not assume that higher numbers are more important. For example, 'laughter' is not more important than 'laugh' when both words are represented in the vector. (wikipedia.org)

Parsing

Breaking a data block into smaller chunks by following a set of rules, so that it can be more easily interpreted, managed, or transmitted by a computer. Spreadsheet programs, for example, parse a data to fit it into a cell of certain size. (businessdictionary.com). ML algorithms can also be used to parse data.

Poisson process

A Poisson Process is a model for a series of discrete events where the average time between events is known, but the exact timing of events is random. A Poisson Process meets the following criteria: (towardsdatascience.com)

• Events are independent of each other. The occurrence of one event does not affect the probability another event will occur.

• The average rate (events per time period) is constant.

• Two events cannot occur at the same time

Python

A programming language available since 1994 that is popular with people doing data science. Python is noted for ease of use among beginners, and great power when used by advanced users, especially when taking advantage of specialized libraries such as those designed for machine learning and graph generation. (datascienceglossary.org)

R

An open-source programming language and environment for statistical computing and graph generation available for Linux, Windows, and Mac. (datascienceglossary.org)

Reinforcement Learning (RL)

Reinforcement Learning (RL) is a sub-field of Machine Learning involving a controller (termed an agent) capable of taking actions in the form of decisions within a system. After each decision is made by the controller, the system evolves to a new state and the controller receives a measure of utility. By trial and error, the controller learns from its experience to optimize an action selection strategy that maximizes the expected cumulative utility within the system. RL is typically used to solve problems that can be modelled as sequential decision processes.

Robotic Process Automation (RPA)

Robotic process automation (RPA) is the term used for software tools that partially or fully automate human activities that are manual, rule-based, and repetitive. They work by replicating the actions of an actual human interacting with one or more software applications to perform tasks such as data entry, process standard transactions, or respond to simple customer service queries. (aiim.org)

Semantic

Semantics can address meaning at the levels of words, phrases, sentences, or larger units of discourse. In machine learning, semantic analysis of a corpus is the task of building structures that approximate concepts from a large set of documents. It generally does not involve prior semantic understanding of the documents. (wikipedia.org)

Stochastic optimization

Stochastic optimization methods are optimization methods that generate and use random variables. For stochastic problems, the random variables appear in the formulation of the optimization problem itself, which involves random objective functions or random constraints. Stochastic optimization methods also include methods with random iterates. (wikipedia.org)

Supervised Learning

A type of machine learning algorithm in which a system is taught via examples. For instance, a supervised learning algorithm can be taught to classify input into specific, known classes. The classic example is sorting email into spam versus non-spam. (datascienceglossary.org)

Unsupervised Learning

A class of machine learning algorithms designed to identify groupings of data without knowing in advance what the groups will be. (datascienceglossary.org)

Web scraping

Web scraping is a term for various methods used to collect information from across the Internet. Generally, this is done with software that simulates human Web surfing to collect specified bits of information from different websites. (techopedia)