1 Basic Concepts in Machine Learning

Machine Learning (ML) is a method of data analysis that automates analytical model building. ML is closely coupled to data science, as shown in figure 1.1, which is a multidisciplinary field that utilizes scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. It is based on the idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention. This section introduces some foundational concepts in machine learning and its applications.

- **ML is not creating robots**: A common misconception is that machine learning is about building robots. In reality, ML focuses on developing algorithms that can learn from and make predictions or decisions based on data.
- **The SPAM filter is one of the initial ML uses**: One of the earliest and most common applications of ML is the spam filter, which classifies emails as spam or not spam. This has been followed by numerous other applications such as:
  - **Speech to text technology**: Converting spoken language into written text, which is used in virtual assistants and transcription services.
  - **Medical diagnostics**: Assisting doctors by predicting diseases from medical images and patient data.
- **ML has lots of fundamental concepts (jargon)**: To effectively understand and apply machine learning, it's essential to grasp several key concepts and terminologies, including:
  - **Supervised vs unsupervised learning**: Supervised learning involves training a model on labeled data, while unsupervised learning deals with data that has no labels and tries to find hidden patterns.

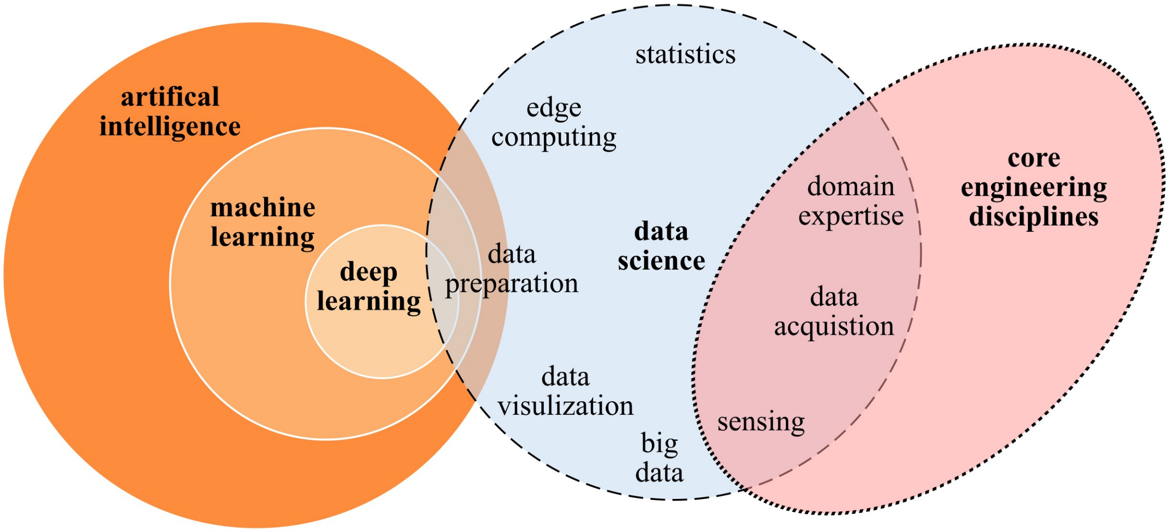


Figure 1.1: Overlap between the fields of Artificial Intelligence (AI), Data science (DS), and the core engineering disciplines of Civil, Mechanical, Electrical, and Chemical Engineering.

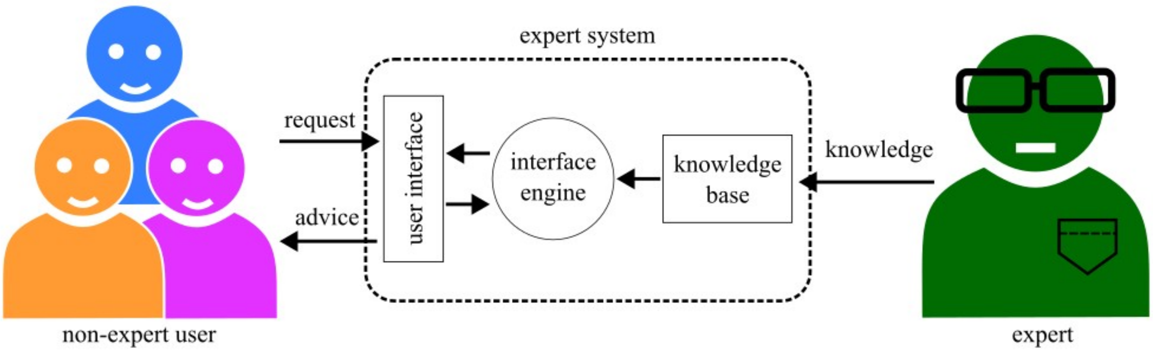


Figure 1.2: Diagram of an "expert system" in AI, which is a computer program that simulates the decision-making ability of a human expert by using a knowledge base and inference rules.

- **Online versus batch learning**: Online learning algorithms update the model incrementally as new data arrives, whereas batch learning algorithms train the model using the entire dataset at once.
- **Instance-based vs model-based learning**: Instance-based learning algorithms, such as k-nearest neighbors, use specific instances to make predictions, whereas model-based algorithms, like linear regression, build a model from the training data and use it to make predictions.

1.1 Examples of Artificial Intelligence

The field of Artificial Intelligence (AI) encompasses a diverse array of technologies and methodologies aimed at enabling machines to perform tasks that typically require human intelligence. This section explores various examples of AI, Machine Learning, and Deep Learning technologies, highlighting their distinct characteristics and applications.

1.1.1 Examples of Artificial Intelligence

Artificial Intelligence (AI) encompasses a wide range of technologies aimed at making machines simulate human intelligence. Some examples include:

- **Expert systems**: Computer programs that simulate the decision-making ability of a human expert by using a knowledge base and inference rules, as diagrammed in figure 1.2.
- **Chatbots**: Programs designed to simulate conversation with human users, especially over the internet.

Figure 1.3: A Symbolics Lisp Machine, a specialized hardware platform designed to run expert systems which are a version of AI focusing on answering questions to challenging problems.[a]

1.1.2 Examples of Machine Learning

Machine Learning, as a subset of AI, involves algorithms that improve automatically through experience. Some common examples include:

- **Linear regression**: A statistical method for modeling the relationship between a dependent variable and one or more independent variables.
- **Classification**: Techniques such as decision trees and support vector machines (SVMs) that categorize data into predefined classes.
- **Simple image/speech recognition**: Algorithms that identify objects in images or convert spoken language into text, fundamental in applications like facial recognition and virtual assistants. An example is the Symbolics Lisp Machine shown in figure 1.3.

[a] Michael L. Umbricht and Carl R. Friend (Retro-Computing Society of RI), CC BY-SA 3.0 <https://creativecommons.org/licenses/by-sa/3.0>, via Wikimedia Commons

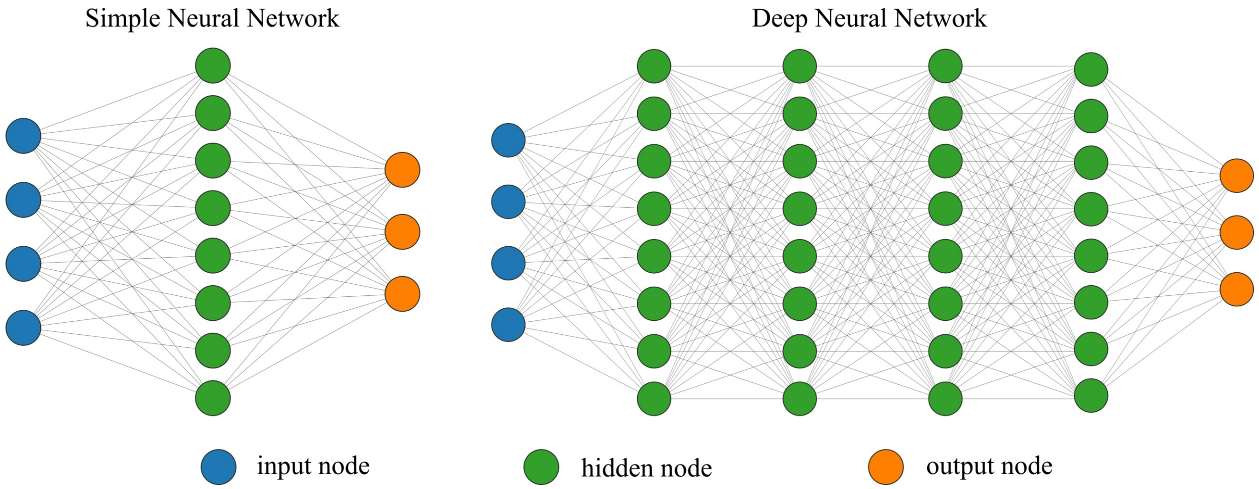
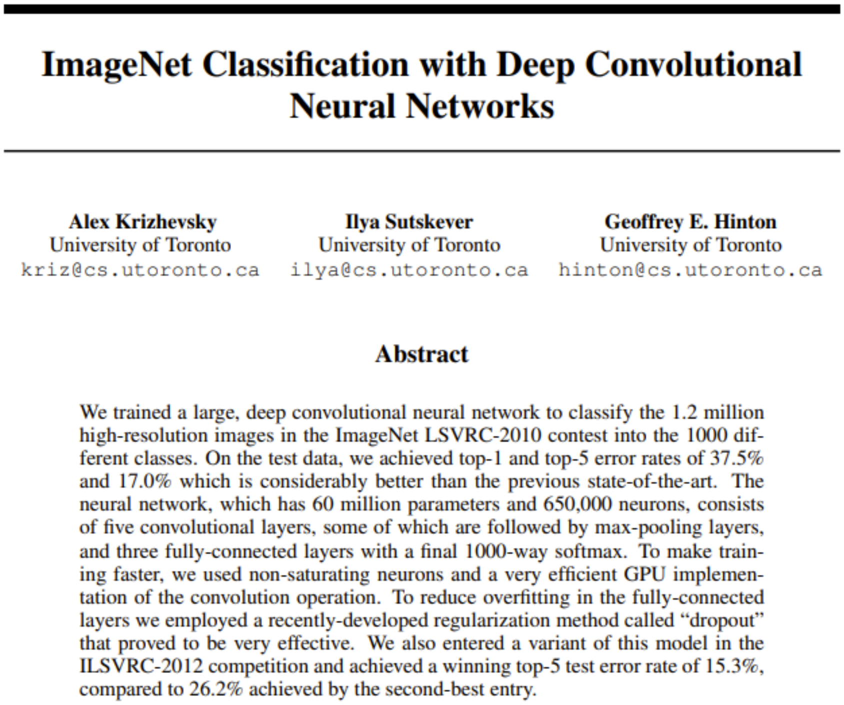


Figure 1.4: simple neural network vs deep learning

1.1.3 Examples of Deep Learning

Deep learning, an enhancement of machine learning, utilizes “deep” neural networks to construct knowledge graphs, as shown in figure 1.4. Though conceptualized in the 1970s, it was initially unfeasible due to the problem of vanishing gradients within the networks. In 2012, Geoffrey E. Hinton’s team demonstrated that a network with 60 million parameters and 650,000 neurons could effectively perform image classification across a dataset containing 1,000 categories (paper shown in figure 1.5). This breakthrough was facilitated by the use of GPUs and a novel regularization technique known as “dropout.” The team’s modified model participated in the ILSVRC-2012 competition, securing a first-place top-5 test error rate of 15.3%, a significant improvement over the 26.2% recorded by the runner-up and a major leap when compared to the previous year as diagrammed in figure 1.6.

Review 1.1 Imagenet in 2012 represented a significant step forward in machine learning by introducing the first practical example of deep learning, famously known as AlexNet (Figure 1.5) This model, developed by Geoffrey Hinton and his team, utilized deep convolutional neural networks to dramatically improve the accuracy of image classification, which was a longstanding challenge in the field.



ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

Figure 1.5: Geoffrey's 2012 paper "ImageNet Classification with Deep Convolutional Neural Networks".[a]

The results of AlexNet were rather monumental, reducing the top-5 test error rate to 15.3% compared to 26.2% by the next best entry, as shown in Figure 1.6. This was clear evidence of deep learning's superior capability over traditional machine learning methods, effectively revolutionizing the approach towards machine learning in the broader scientific community.

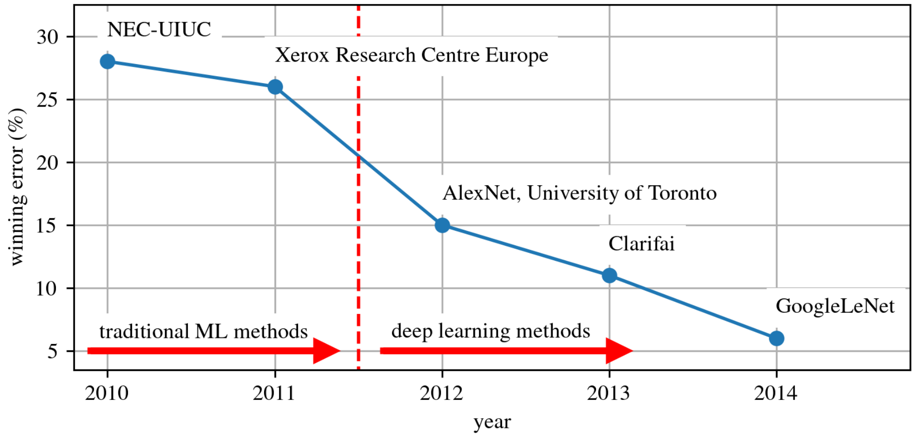


Figure 1.6: ImageNet Competition Results showing the impact of deep learning methods on image classification.

The implications of this leap forward led to broad applications of deep learning that permeate numerous aspects of technology and science today. The impact of AlexNet and subsequent deep learning developments culminated in the awarding of the 2024 Nobel Prize in Physics to Geoffrey Hinton, recognizing his contributions to the field of artificial neural networks. Figure 1.7 shows the 2024 Nobel Award Ceremony in Stockholm Sweden with John Hopfield and Geoffrey Hinton receiving their awards from the King of Sweden.

Figure 1.7: The 2024 Nobel Prize in Physics for "for foundational discoveries and inventions that enable machine learning with artificial neural networks" was awarded to: (a) John Hopfield and (b) Geoffrey Hinton. [a]

[a] Copyright held by Authors and Neural Information Processing Systems Foundation, Inc., used under fair use. <https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>

[a] The author of this text was lucky enough to attend the 2024 Nobel Prize Ceremony in Stockholm and took these pictures.

1.2 Definition of Machine Learning

The definitions and conventions that follow provide a common language and point of reference for all subsequent material:

- Machine learning is the discipline of creating computer programs that improve automatically by analyzing data. Here a broader and a more engineering-focused definition of Machine learning is provided5:
  - [Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed. Arthur Samuel, 1959.
  - A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E. Tom Mitchell, 1997.
- The use of X and Y for variables.
  - For the i^{th} sample, $x^{(i)}$ contains all its input features (excluding the target), while $y^{(i)}$ denotes the corresponding target value.
  - General definition: “Machine learning algorithms are described as learning a target function (f) that best maps input variables (X) to an output variable (Y).”

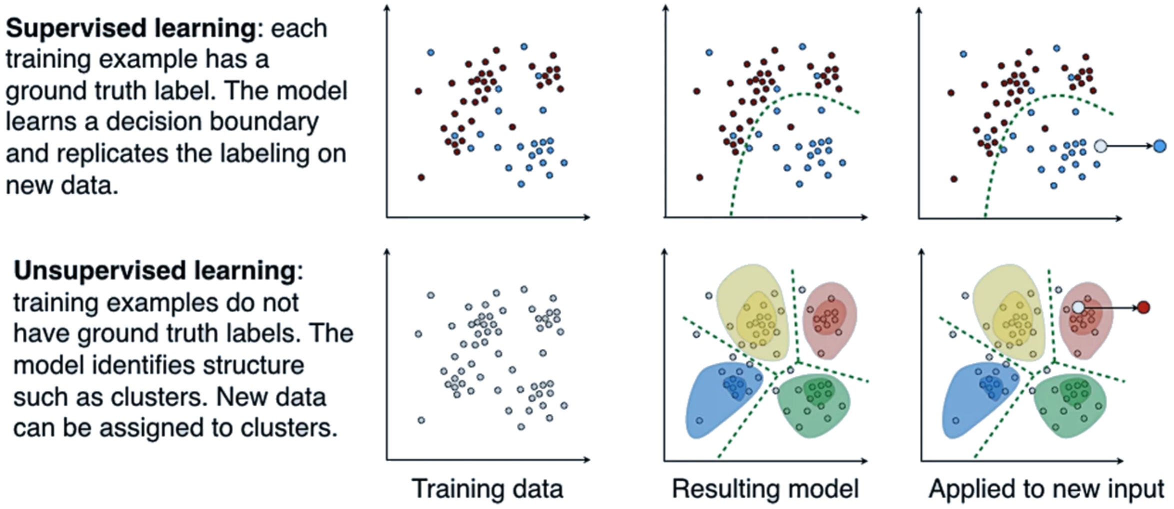
$$Y = f(X) \tag{1.1}$$

This is described as a standard learning challenge where the goal is to predict future values Y using new samples of input variables X. The function f that relates inputs to outputs is not known. If it were, direct application would be possible, eliminating the necessity for learning it via machine learning methods. This process is more complex than it may initially seem. Furthermore, there is an error e associated with this task that is independent of the input data X.

$$Y = f(X) + e \tag{1.2}$$

This yields two primary phases in the machine-learning workflow:

- Training: Creating the model is a compute-intensive process often run in a data center
- Inference: Using the model can be computationally cheap and even performed “at the edge”



Supervised learning: each training example has a ground truth label. The model learns a decision boundary and replicates the labeling on new data.

Unsupervised learning: training examples do not have ground truth labels. The model identifies structure such as clusters. New data can be assigned to clusters.

Figure 1.8: Supervised vs unsupervised unsupervised machine learning methods[a].

1.3 Supervision in Machine Learning

The general domains of ML can be classified into supervised and unsupervised learning, as shown in figure 1.8.

1.3.1 Supervised Learning

In supervised learning, the data is pre-categorized with labels

- **Classification**, The process of categorizing a given set of data into classes
- **Regression**, The process of estimating the relationships between a dependent variables (i.e. output) and one or more independent variables (i.e. input).

In supervised learning, the training data provided to the algorithm includes the desired solutions, known as labels. A common task within this type of learning is classification. An example of classification is a spam filter, which is trained using numerous emails that are each labeled as either spam or ham. The objective for the spam filter is to learn how to accurately classify new emails based on this training. Regression is another typical task is to predict a target numeric value, such as the price of a house, given a set of features (size, proximity to railroad or highways, age of house, style, etc.) called predictors. To train a system you need to build a model and train in on many house sale examples, including both predictors (house features) and their labels (i.e., prices). Some of the most important supervised learning algorithms:

- k-Nearest Neighbors
- Linear Regression
- Logistic Regression
- Support Vector Machines (SVMs)
- Decision Trees and Random Forests

[a] modified from: Langs, G., Röhrich, S., Hofmanninger, J. et al., CC BY 4.0 <https://creativecommons.org/licenses/by/4.0>, via Wikimedia Commons

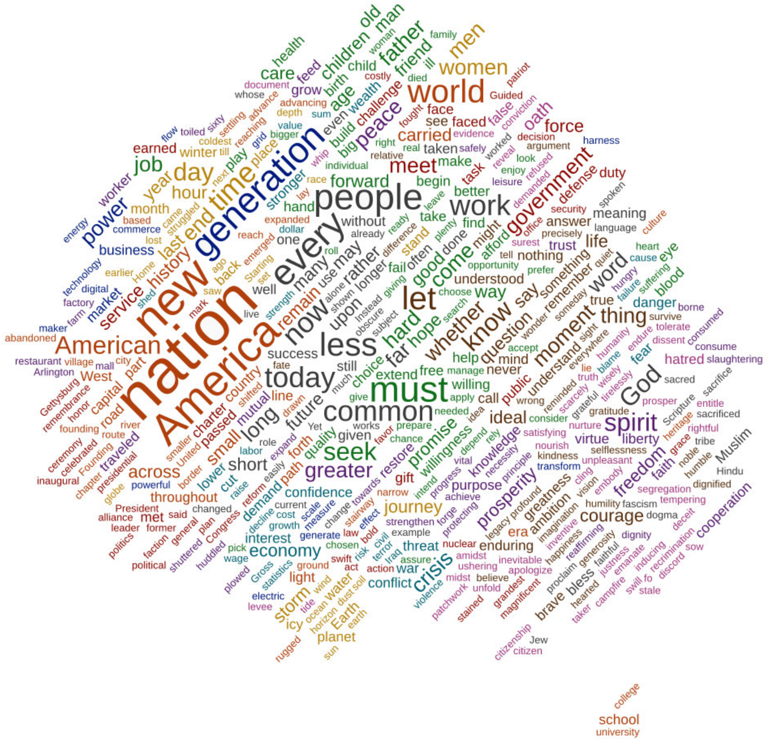


Figure 1.9: A semantic word cloud of Barack Obama's First Inaugural Address [a].

1.3.2 Unsupervised Learning

In unsupervised learning, the training data is unlabeled, and the system endeavors to learn independently without explicit guidance. For instance, consider a situation where you have extensive data on what people watch on YouTube. You might use a clustering algorithm to identify groups of similar users. However, at no point do you instruct the algorithm on which group a user belongs to; it discovers these connections autonomously. For example, it might identify that the types of videos people watch are closely linked to specific age and income metrics. Older viewers may prefer watching videos about vacations, while younger viewers might watch a lot of educational content, such as Khan Academy.

Visualization algorithms such as that in Figure 1.9) serve as an important instance of unsupervised learning algorithms: they take in extensive, complex, and unlabeled data and produce a 2D or 3D representation that can be conveniently visualized. These algorithms strive to maintain as much of the original data structure as possible (for example, ensuring that distinct clusters in the input space do not merge in the visual output). This helps in comprehending the organizational structure of the data and potentially uncovering hidden patterns. Other Important concepts in unsupervised learning include

[a] Modified from: GuoYongzhi, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons

- **Clustering**, The process of identifying and grouping similar data points in larger datasets without concern for the specific outcome
- **Association**, The process learning a rule-based method for discovering relations between variables data data
- **Dimension Reduction**, The process of reducing the number of input variables in training data.

Here are some of the most important unsupervised learning algorithms:

- Clustering
- k-Means
- Hierarchical Cluster Analysis (HCA)
- Expectation Maximization

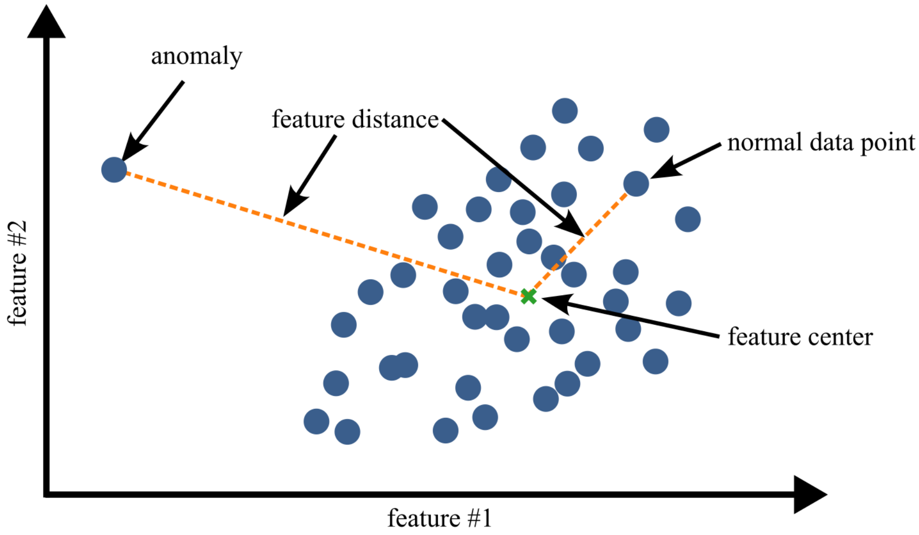


Figure 1.10: Anomaly Detection for a given set of data.

Another key application of unsupervised learning is anomaly detection as shown in figure 1.10. The goal is to flag unusual credit-card transactions that may signal fraud, spot defects on a production line, or filter out outliers before passing a dataset to another learning algorithm. The model is first exposed to many examples of normal behavior so it can build a reference profile. When a new observation arrives it checks how closely the example matches that profile and labels it normal or anomalous accordingly.

1.3.3 Semisupervised Learning

Certain algorithms are designed to manage partially labeled training data, typically characterized by a substantial amount of unlabeled data with a minimal amount of labeled data as shown in figure 1.11. Many semi-supervised learning algorithms are hybrid forms, integrating features of both unsupervised and supervised learning approaches.

An example is seen in image hosting services like Google Photos. Upon uploading your family images, the system automatically recognizes and groups repeated appearances of the same individuals across different photos, such as person 1 in photos A, C, and D and person 2 in photos A, B, and E. This clustering is the unsupervised component of the algorithm. The system then only needs a single label for each person to subsequently identify and tag everyone in all images, enhancing the ease of searching through photos.

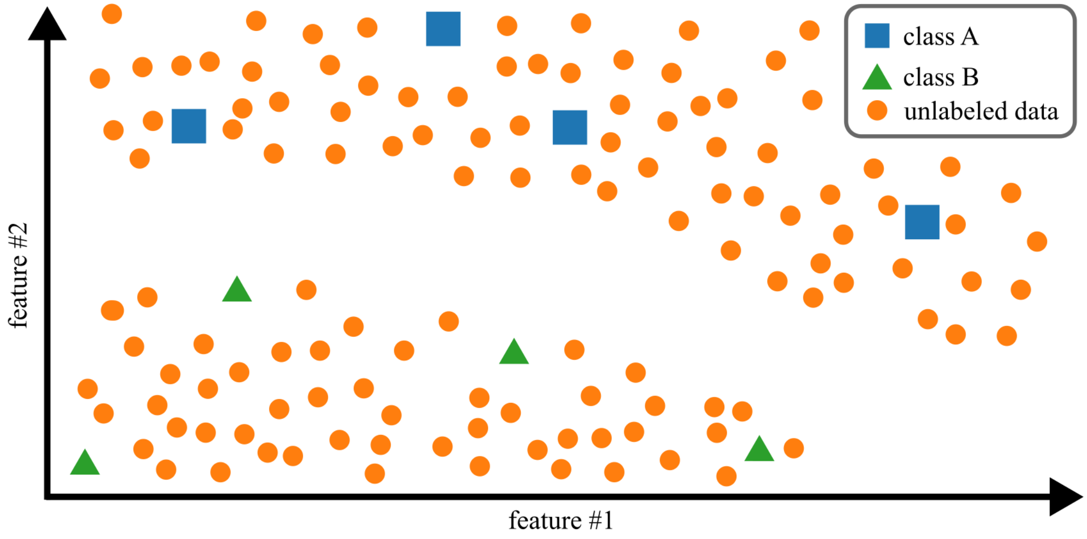


Figure 1.11: Semisupervised learning enabling the use of a limited set of labeled data to infer the labels of larger unlabeled datasets.

1.3.4 Reinforcement Learning

Reinforcement learning follows a distinct paradigm in which an autonomous agent repeatedly interacts with its environment. At each step the agent observes the current state, chooses and executes an action, and then receives feedback in the form of a reward or penalty; as shown in figure 1.12. By sampling many such cycles the agent gradually discovers a strategy, known as a policy, that maps perceived situations to actions so as to maximise the total reward accumulated over time.

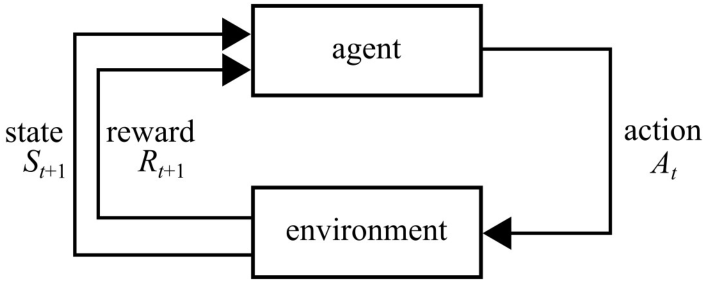


Figure 1.12: reinforcement learning control diagram

1.4 Types of learning (Batch and Online)

Machine learning systems can be broadly categorized based on how they process and learn from data.

1.4.1 Batch Learning

In batch learning, the system is fully trained using the entirety of the available data, which often demands significant time and computational resources, and is thus usually performed offline. Initially, the system undergoes training; after which, it is deployed into production where it no longer learns but merely applies the previously acquired knowledge. This method is referred to as offline learning. The deployment phase, known as inference, is typically executed swiftly.

To update a batch learning system with new information, such as recognizing a novel form of spam, it is necessary to develop a completely new version of the system. This version must be trained from the ground up using the entire dataset, both old and new data, before replacing the older system in production.

Despite these challenges, the training, evaluation, and deployment processes of a machine learning system can be automated, allowing even batch learning systems to adapt to changes. This approach is straightforward and usually effective; however, training on a full dataset can be time-consuming - often taking many hours - hence, systems are usually updated no more frequently than daily or weekly. Moreover, utilizing the full dataset requires extensive computing resources, such as CPU power, memory, disk space, and network bandwidth. For organizations with vast amounts of data, the costs of daily retraining from scratch can be prohibitively expensive.

If the dataset is exceptionally large, employing a batch learning algorithm may become impractical. Additionally, in situations where autonomous learning is essential and computational resources are limited, such as with smartphone applications or extraterrestrial rovers, the need to manage large datasets and conduct lengthy training sessions daily poses significant challenges.

1.4.2 Online Learning

Online learning trains a model incrementally by presenting new data points one at a time or in small mini-batches. Each update is computationally light and fast, allowing the model to refresh its knowledge continuously as fresh data streams in. This approach is particularly appropriate for systems that:

- receive data continuously, such as stock prices,
- need to quickly or autonomously adapt to changes,
- have limited computing resources: once an online learning system processes new data instances, they can be discarded to save space, unless there is a need to revert to a previous state and "replay" the data.

Online learning is also useful for managing large datasets that exceed the memory capacity of a single machine, known as out-of-core learning. The algorithm processes parts of the data, conducts a training step, and repeats this until all the data has been processed.

A critical parameter in online learning systems is the learning rate, which dictates how rapidly the system adapts to changing data. A high learning rate allows for rapid adaptation but may also lead to quick forgetting of old data and training on noise. A low learning rate results in slower learning and reduced sensitivity to variations in new data.

1.4.3 Bad Data in Machine Learning

A significant challenge in online learning is the system's susceptibility to performance degradation when exposed to poor quality data. This is particularly problematic in live systems where clients might quickly notice issues. For instance, bad data could originate from a malfunctioning robot sensor or an attempt to manipulate search engine rankings through spamming. To mitigate these risks, it is crucial to vigilantly monitor the system and quickly disable learning or revert to a previously effective state if a decline in performance is observed. Additionally, monitoring the input data for anomalies and employing anomaly detection algorithms can help identify and respond to aberrant data.

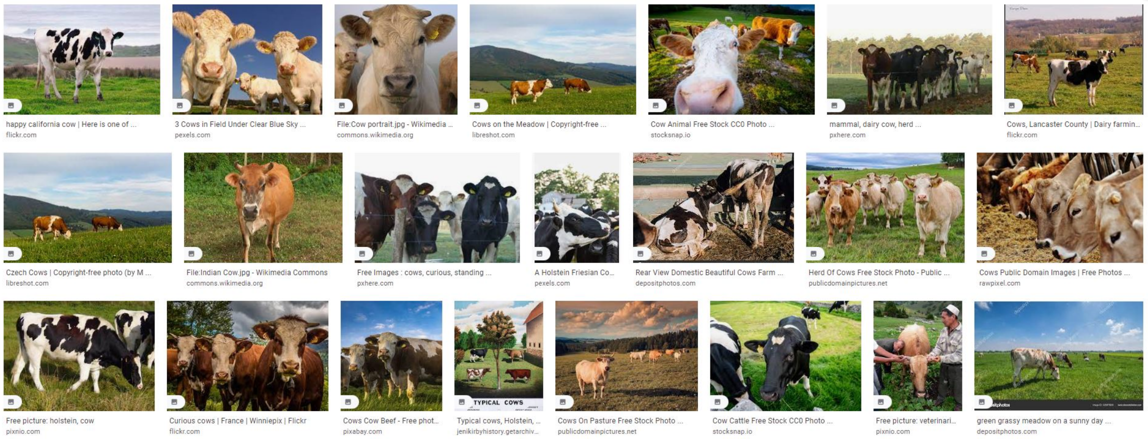

happy california cow | Here is one of ...
flickr.com
3 Cows in Field Under Clear Blue Sky ...
pexels.com
File:Cow portrait.jpg - Wikimedia ...
commons.wikimedia.org
Cows on the Meadow | Copyright-free ...
libreshot.com
Cow Animal Free Stock CC0 Photo ...
stocksnap.io
mammal, dairy cow, herd ...
pxhere.com
Cows, Lancaster County | Dairy farmin...
flickr.com
Czech Cows | Copyright-free photo (by M ...
libreshot.com
File:Indian Cow.jpg - Wikimedia Commons
commons.wikimedia.org
Free Images : cows, curious, standing ...
pxhere.com
A Holstein Friesian Co...
pexels.com
Rear View Domestic Beautiful Cows Farm ...
depositphotos.com
Herd Of Cows Free Stock Photo - Public ...
publicdomainpictures.net
Cows Public Domain Images | Free Photos ...
rawpixel.com
Free picture: holstein, cow
pixnio.com
Curious cows | France | Winniepix | Flickr
flickr.com
Cows Cow Beef - Free phot...
pixabay.com
TYPICAL COWS
Typical cows, Holstein, ...
jenikirbyhistory.getarchiv...
Cows On Pasture Free Stock Photo ...
publicdomainpictures.net
Cow Cattle Free Stock CC0 Photo ...
stocksnap.io
Free picture: veterinari...
pixnio.com
green grassy meadow on a sunny day ...
depositphotos.com

Figure 1.13: Training a machine learning algorithm to recognize cows may end up just learning to recognize grass. This is sometimes called "short-cut learning"[a].

For example, imagine training a convolutional neural network to recognize cows (Figure 1.13). Every image in the training set shows black-and-white cattle standing on lush green pasture, so the

[a] Screen shoot of a Google image search for cows taken by the Austin Downey, all images stated to be under a creative commons license per Google search tools; also assumed to be fair use under given the educational purpose of this text, via https://www.google.com/

easiest statistical cue for the model to latch onto is the dominant green background rather than the animals themselves. At inference time, the network confidently labels any scene filled with grass as “cow,” yet fails when presented with a cow on snow or asphalt. In other words, it has learned “grass recognition,” not “cow recognition”.

1.5 Learning Approaches: Instance-Based and Model-Based

Another method of classifying machine learning systems is based on their generalization capabilities. Typically, the main objective in machine learning is to make predictions, which requires the system to generalize from its training examples to new, unseen instances. Achieving high performance on training data is useful but not the ultimate goal; the system must also excel when confronted with new data. There are primarily two strategies for generalization: instance-based learning and model-based learning.

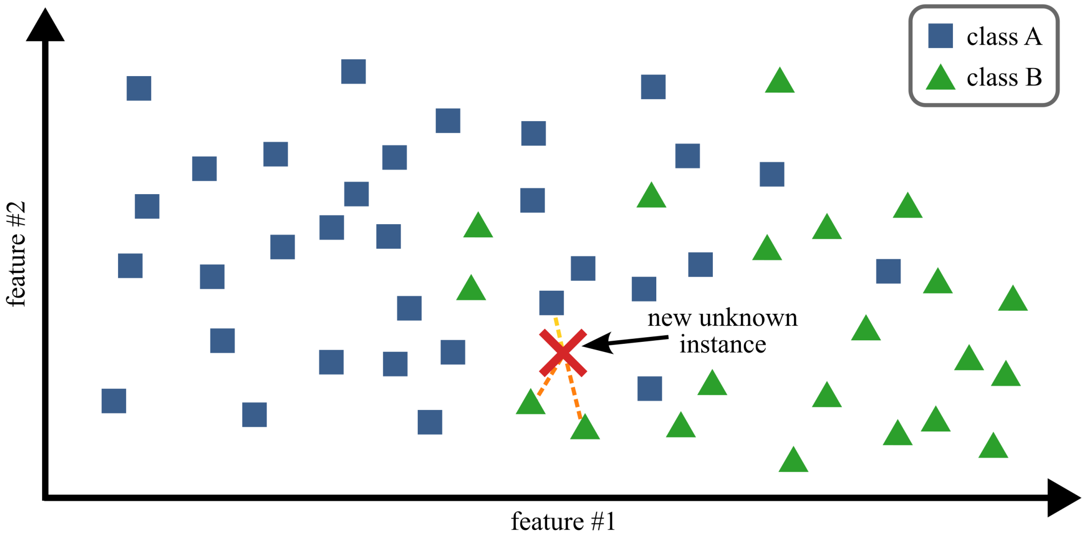


Figure 1.14: Instance-based learning where the class of a unknown instance is inferred from the its distance to data points with known labels.

1.5.1 Instance-based learning

One of the simplest learning strategies is rote memorization. For instance, a spam filter built on this principle would only mark emails as spam if they match previously flagged emails exactly. While this approach is straightforward, it’s hardly the most effective. A more sophisticated spam filter could extend its detection capabilities to include emails that closely resemble known spam messages. This approach necessitates a metric for measuring the similarity between two emails. A basic method might involve counting the shared words between emails. Under this system, an email would be classified as spam if it shares a significant number of words with an email already identified as spam. This method exemplifies instance-based learning (Figure 1.14), where the system memorizes specific examples and uses a similarity metric to generalize to new instances.

1.5.2 Model-based learning

Another approach to generalization involves constructing a model based on a set of examples and then using this model to make predictions as shown in figure 1.15. This method is known as model-based learning.

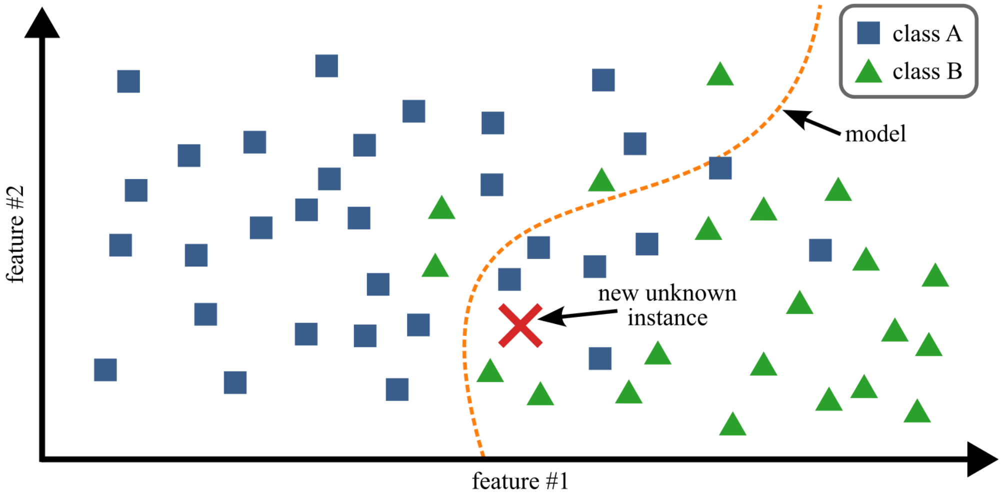


Figure 1.15: Model-based learning where the class of a unknown instance is inferred from its location in reference to a model trained on the data with known labels.

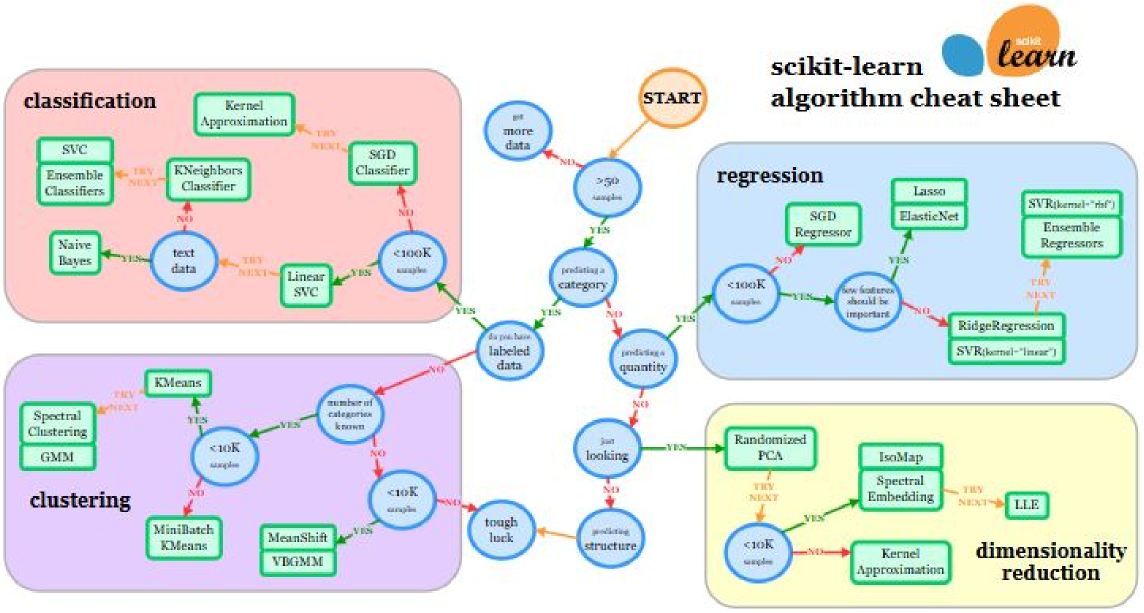


Figure 1.16: Flowchart of estimators used in the scikit learn library that intends guide users for what algorithms to use for a given case[a].

[a] Scikit-learn algorithms cheat sheet, permissive simplified BSD license and assumed to be fair use under given its nature as documentation and the educational purpose of this text, via https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

1.5.3 Selecting methods:

One of the initial challenges faced by newcomers to machine learning is selecting the appropriate algorithms (or estimators) for specific tasks. This is due to the fact that various algorithms excel with different kinds of data and problems. The flowchart depicted in Fig. 1.16, which originates from the scikit-learn library documentation, is intended to guide users in choosing the most suitable algorithms for their particular datasets.

1.6 The Unreasonable Effectiveness of Data

In a landmark study released in 2001, Microsoft researchers Michele Banko and Eric Brill demonstrated that a variety of Machine Learning algorithms, even relatively straightforward ones, achieved similar levels of performance on a complex task of natural language disambiguation when provided with sufficient data. This finding is illustrated in figure 1.17 and further discussed in their paper[a].

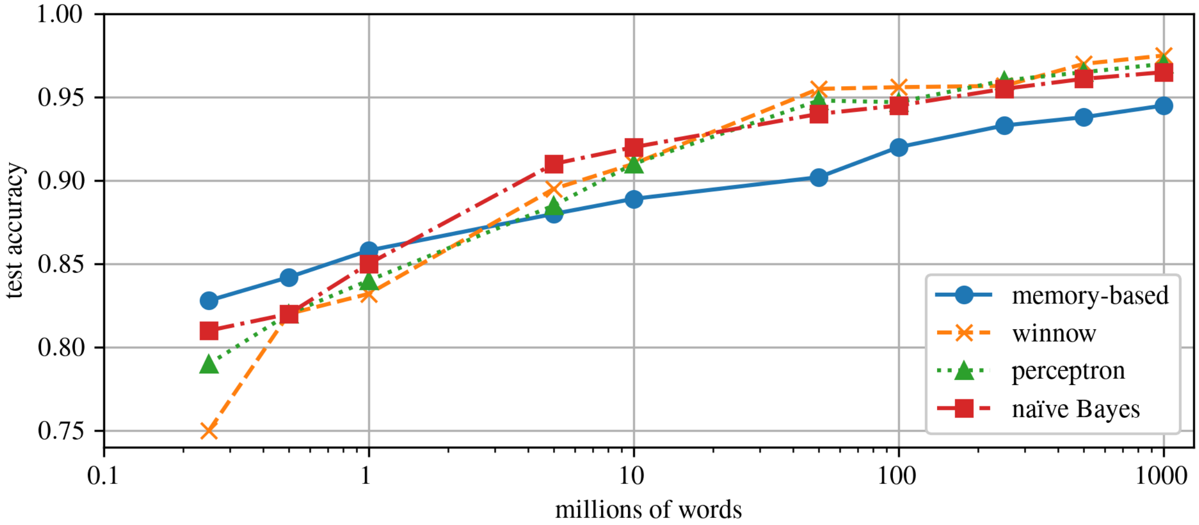


Figure 1.17: Learning curves for four algorithms studied in Banko and Brill showing the importance of the amount of data when compared to algorithm selection.

As the authors put it: “these results suggest that we may want to reconsider the tradeoff between spending time and money on algorithm development versus spending it on corpus development.”. Or put more directly “The Unreasonable Effectiveness of Data”[b]. It’s important to recognize, however, that small and medium-sized datasets remain prevalent, and acquiring additional training data is not always straightforward or economical. Therefore, the significance of algorithm selection should not be overlooked.

[a] Michele Banko and Eric Brill. “Scaling to very very large corpora for natural language disambiguation.” Proceedings of the 39th annual meeting of the Association for Computational Linguistics. 2001.

[b] Halevy, Alon, Peter Norvig, and Fernando Pereira. “The unreasonable effectiveness of data.” IEEE intelligent systems 24.2 (2009): 8-12.