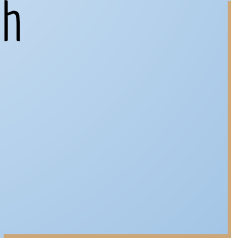


The Machine learning

Salar Mokhtari Laleh

Spring 2021





01_The machine learning landscape



01_The machine learning landscape

- What's machine learning ?
- Why use machine learning ?
- Types of machine learning systems
- Main challenges of machine learning
- Testing and validation
- Exercises

What Is Machine Learning ?

Machine Learning is the science (and art) of programming computers so they can learn from data.

Here is a slightly more general definition:

[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

—Arthur Samuel, 1959

And a more engineering-oriented one:

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

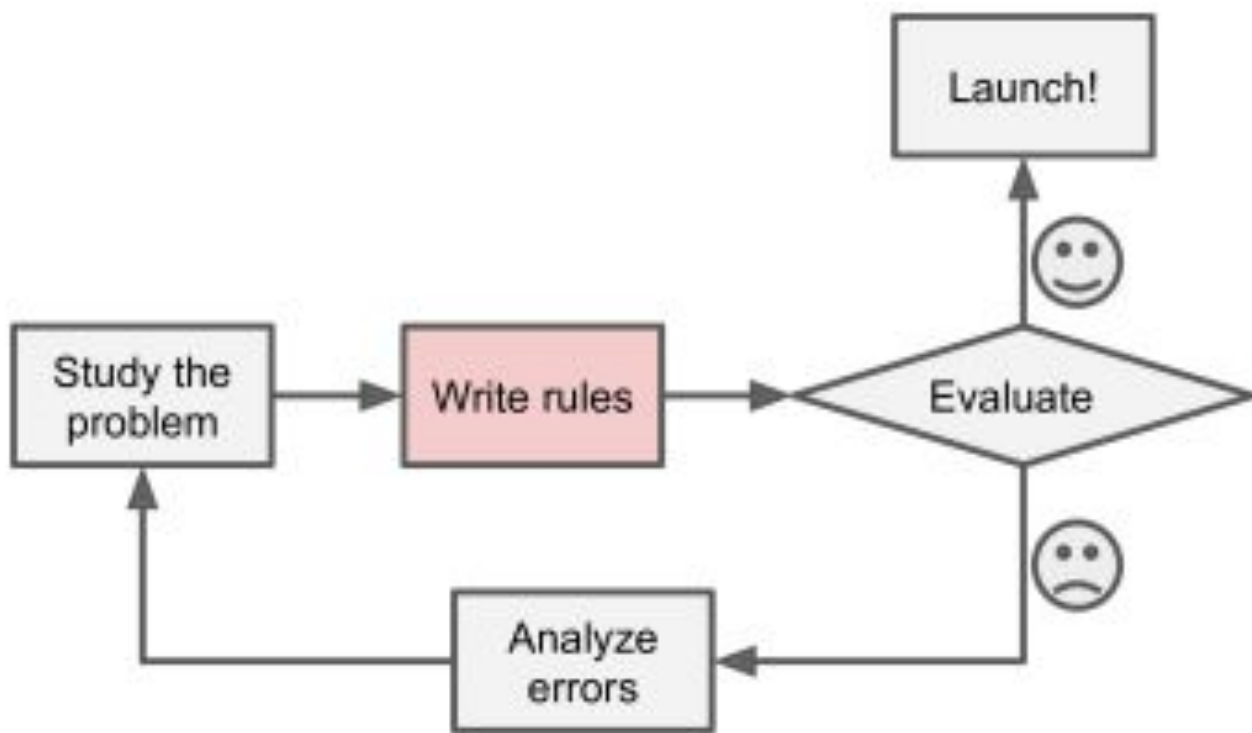
For example, your spam filter is a Machine Learning program that can learn to flag spam given examples of spam emails (e.g., flagged by users) and examples of regular (nospam, also called “ham”) emails. The examples that the system uses to learn are called the training set. Each training example is called a training instance (or sample). In this case, the task T is to flag spam for new emails, the experience E is the training data, and the performance measure P needs to be defined; for example, you can use the ratio of correctly classified emails. This particular performance measure is called accuracy and it is often used in classification tasks.

If you just download a copy of Wikipedia, your computer has a lot more data, but it is not suddenly better at any task. Thus, downloading a copy of Wikipedia is not ML

Why Use Machine Learning?

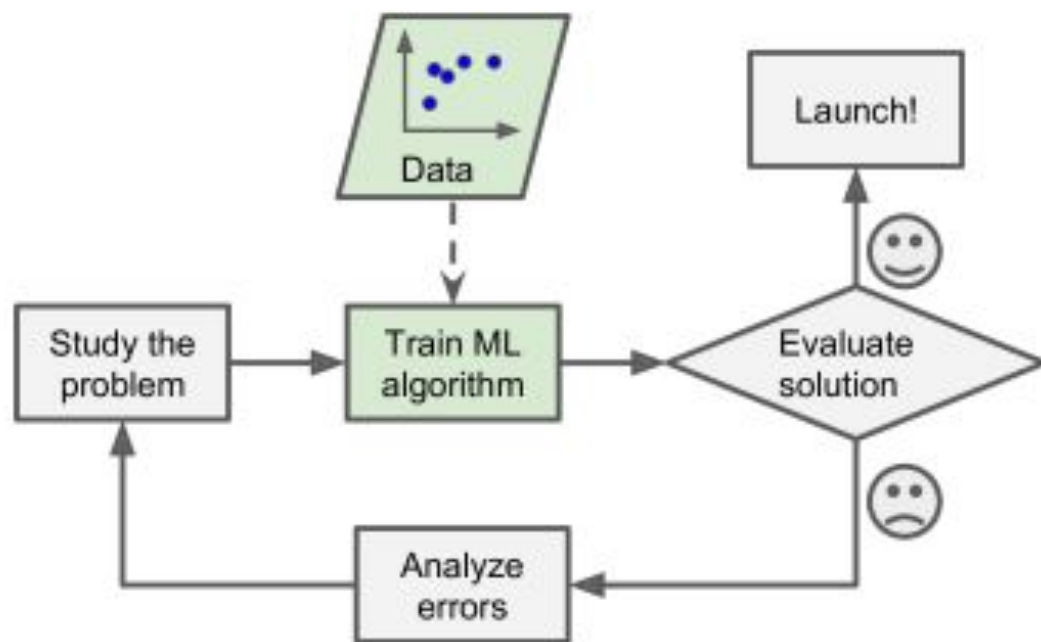
Consider how you would write a spam filter using traditional programming techniques

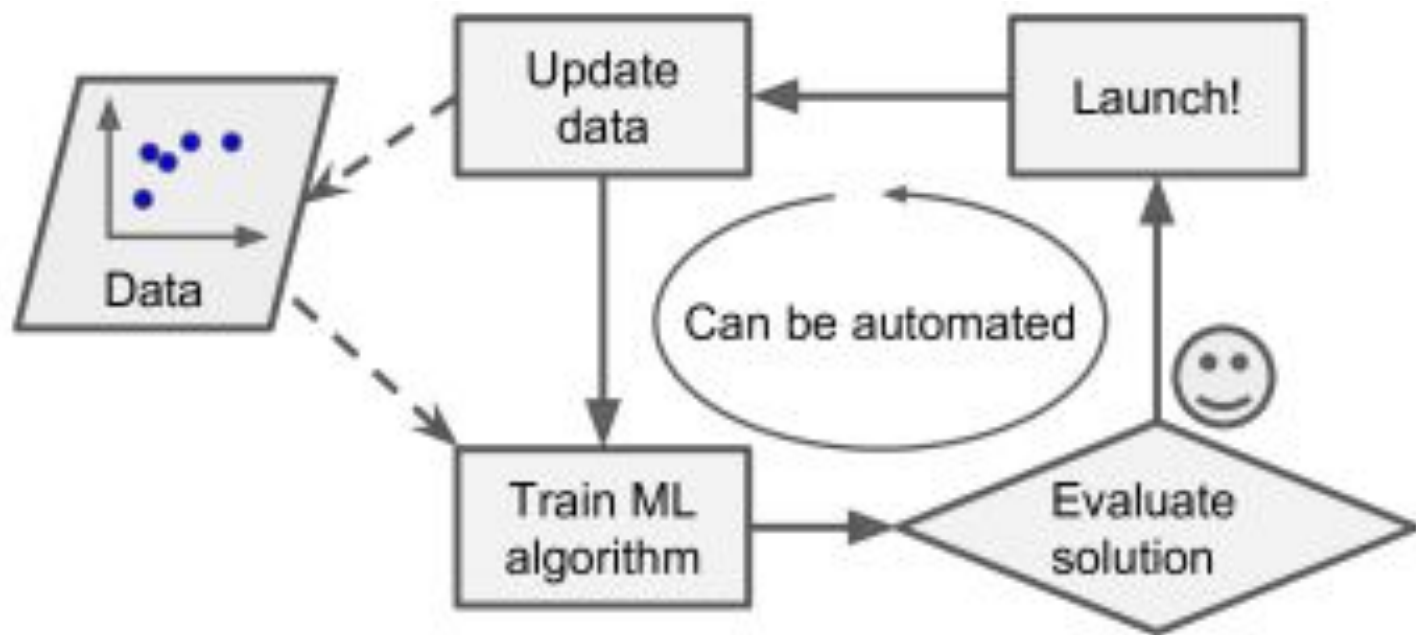
1. First you would look at what spam typically looks like. You might notice that some words or phrases (such as “4U,” “credit card,” “free,” and “amazing”) tend to come up a lot in the subject. Perhaps you would also notice a few other patterns in the sender’s name, the email’s body, and so on
2. You would write a detection algorithm for each of the patterns that you noticed, and your program would flag emails as spam if a number of these patterns are detected
3. You would test your program, and repeat steps 1 and 2 until it is good enough

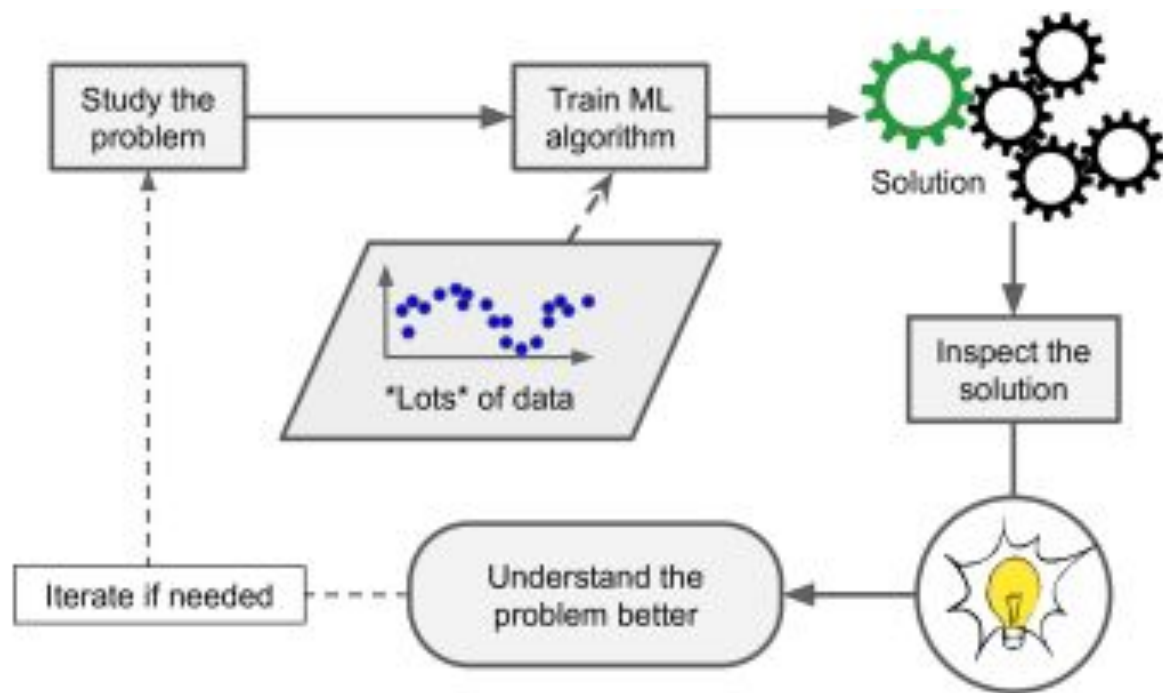


Since the problem is not trivial, your program will likely become a long list of complex rules—pretty hard to maintain.

In contrast, a spam filter based on Machine Learning techniques automatically learns which words and phrases are good predictors of spam by detecting unusually frequent patterns of words in the spam examples compared to the ham examples . The program is much shorter, easier to maintain, and most likely more accurate.







To summarize, Machine Learning is great for:

- Problems for which existing solutions require a lot of hand-tuning or long lists of rules: one Machine Learning algorithm can often simplify code and perform better.
- Complex problems for which there is no good solution at all using a traditional approach: the best Machine Learning techniques can find a solution.
- Fluctuating environments: a Machine Learning system can adapt to new data.
- Getting insights about complex problems and large amounts of data.

Types of Machine Learning Systems

There are so many different types of Machine Learning systems that it is useful to classify them in broad categories, based on the following criteria:

- Whether or not they are trained with human supervision (supervised, unsupervised, semisupervised, and Reinforcement Learning)
- Whether or not they can learn incrementally on the fly (online versus batch learning)
- Whether they work by simply comparing new data points to known data points, or instead detect patterns in the training data and build a predictive model, much like scientists do (instance-based versus model-based learning)

These criteria are not exclusive; you can combine them in any way you like. For example, a state-of-the-art spam filter may learn on the fly using a deep neural network model trained using examples of spam and ham; this makes it an online, model based, supervised learning system.

Let's look at each of these criteria a bit more closely



Supervised/Unsupervised Learning

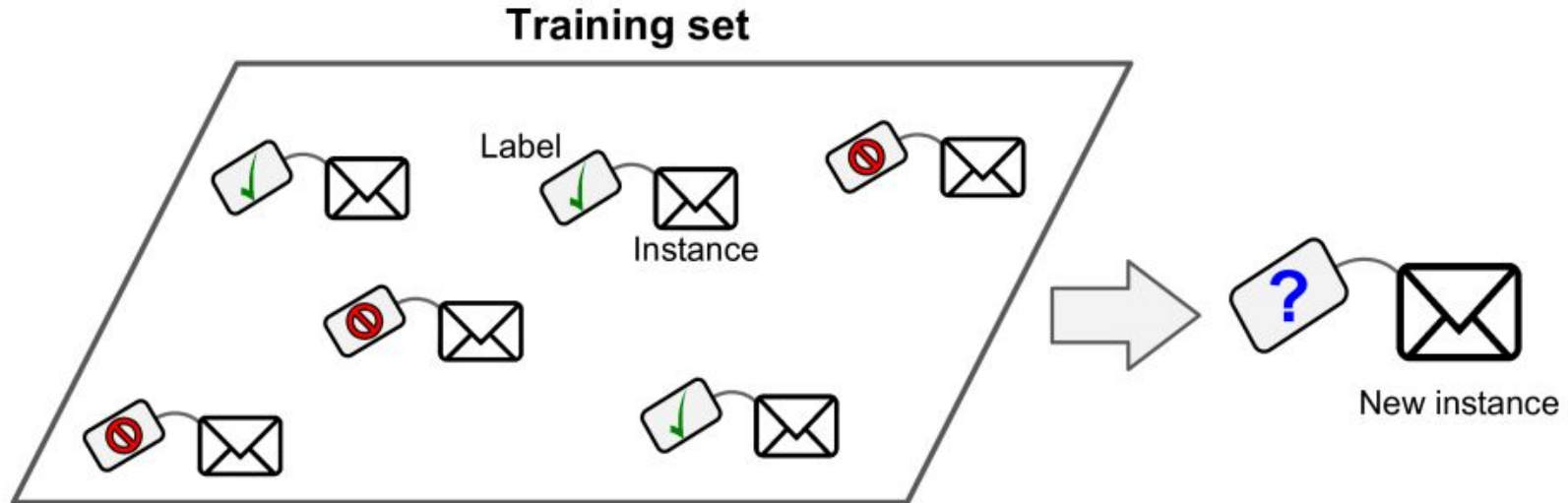


Machine Learning systems can be classified according to the amount and type of supervision they get during training. There are four major categories:

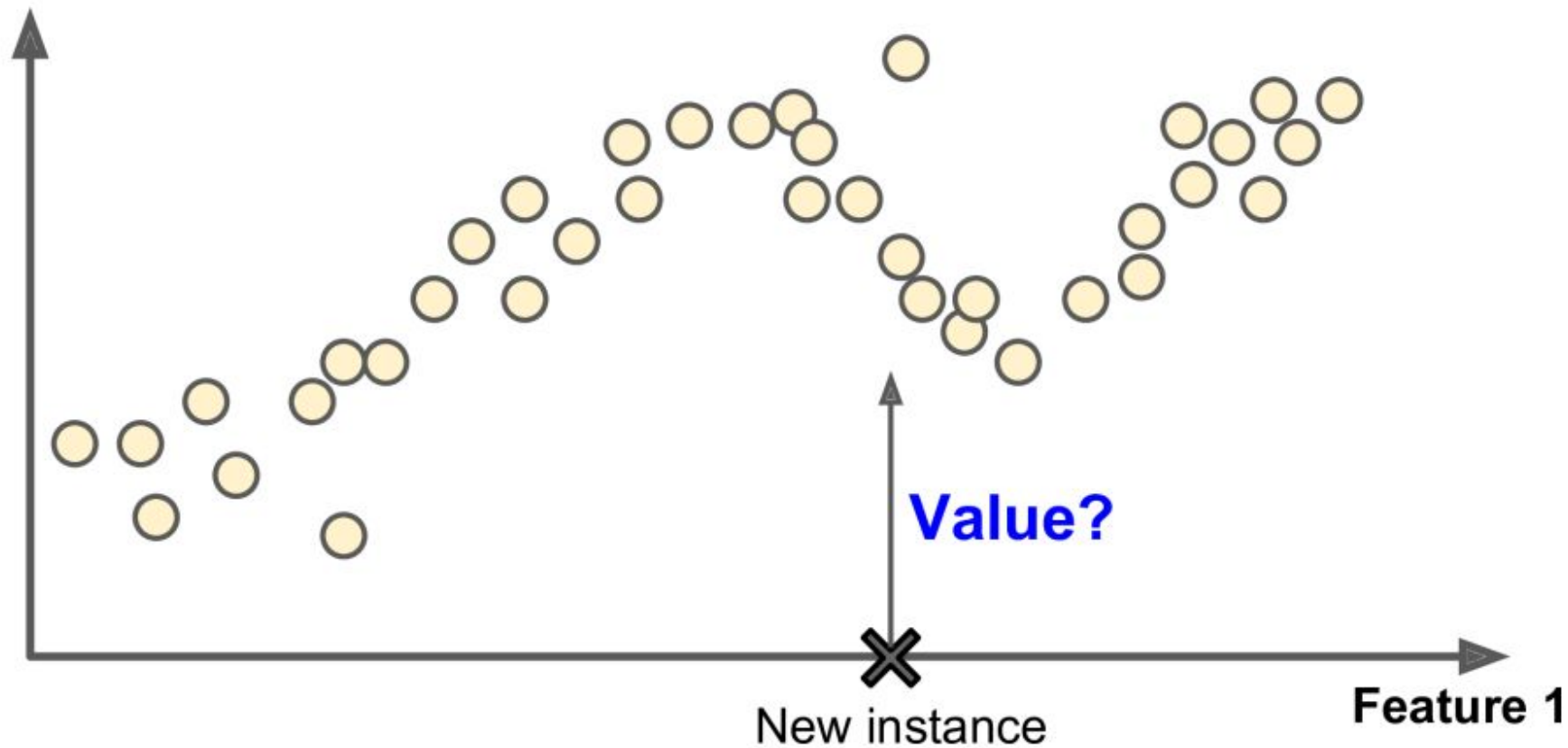
supervised learning, unsupervised learning, semisupervised learning, and Reinforcement Learning.

Supervised learning

In supervised learning, the training data you feed to the algorithm includes the desired solutions, called *labels*



Value

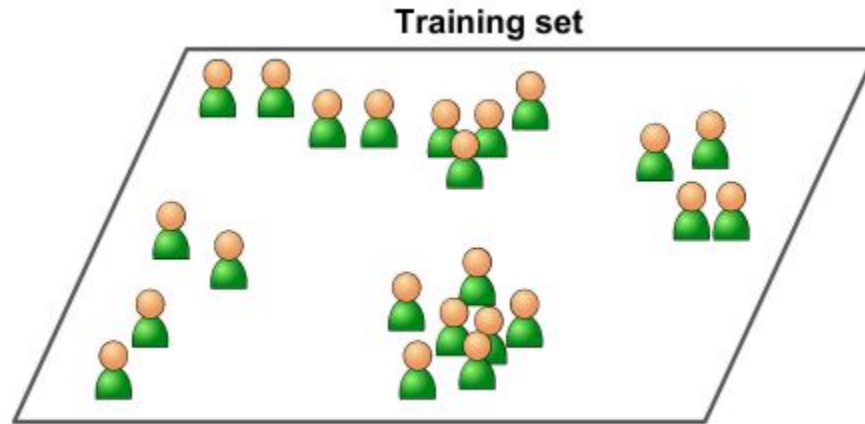


Here are some of the most important supervised learning algorithms:

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

Unsupervised learning

In unsupervised learning, as you might guess, the training data is unlabeled. The system tries to learn without a teacher.



Here are some of the most important unsupervised learning algorithms:

- Clustering
 - K-Means
 - DBSCAN
 - Hierarchical Cluster Analysis (HCA)
- Anomaly detection and novelty detection
 - One-class SVM
 - Isolation Forest
- Visualization and dimensionality reduction
 - Principal Component Analysis (PCA)
 - Kernel PCA
 - Locally-Linear Embedding (LLE)
 - t-distributed Stochastic Neighbor Embedding (t-SNE)
- Association rule learning
 - Apriori
 - Eclat

