Definition and a taxonomy of Machine Learning

Alberto Paccanaro
EMAp – FGV

www.paccanarolab.org

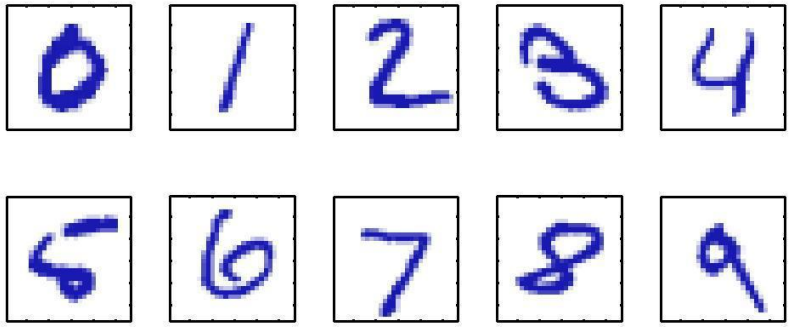
Some material and images are from (or adapted from):
C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006
A. Geron, Hands-on Machine Learning with Scikit-Learn, Keras and TensorFlow, O'Reilly, 2020

Patterns

There are patterns in data !

Pattern recognition: the automatic discovery of regularities (patterns) in data through the use of algorithms

Example: *Handwritten Digit Recognition*



Each digit corresponds to a 28×28 pixel image .
It can be represented by a vector $\boldsymbol{x}$ of 784 real numbers.

GOAL: build a machine that will take a vector $\boldsymbol{x}$ as input and that will produce the identity of the digit $0, \ldots, 9$ as the output.

Handcrafted rules for distinguishing the digits based on the shapes of the strokes won't work…

The ML approach

A dataset is used to tune the parameters of an **adaptive model (a function with parameters).**

GENERALIZATION

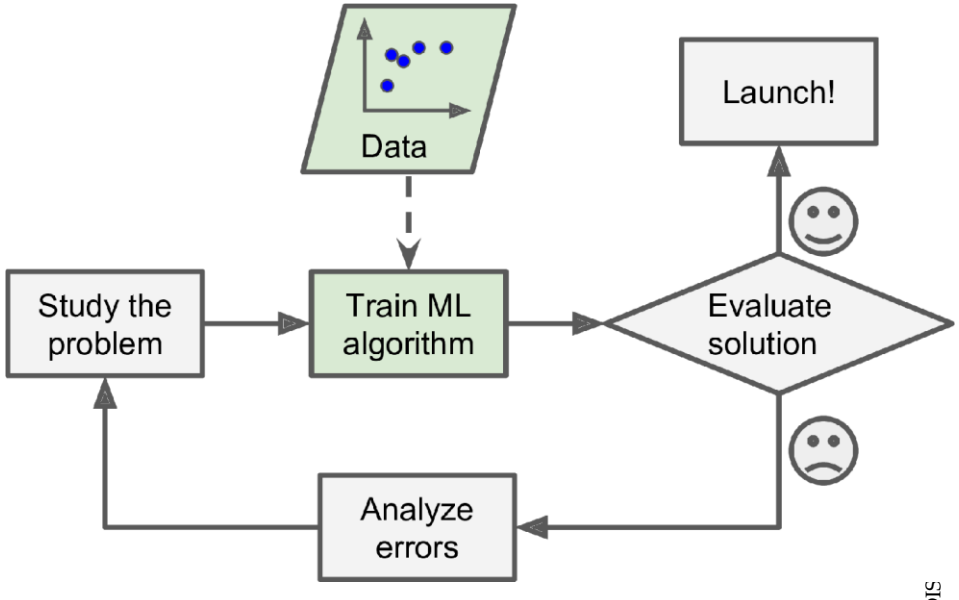
We need:

1. A set of digits $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ called a *training set*.
2. For each digit in the training set, its category is known and represented by a *target vector* $\mathbf{t}$

The **result** of the machine learning algorithm is:

a function $f(\boldsymbol{x})$ which takes a new digit image $\boldsymbol{x}$ as input and that generates an output vector $\boldsymbol{y}$, encoded in the same way as $\boldsymbol{t}$ – thus indicating the predicted class of $\boldsymbol{x}$

Image	label
x_1	"6"
x_2	"8"
. . .	…
x_N	"3"



$$f_{\boldsymbol{p}}(\boldsymbol{x}) \rightarrow \boldsymbol{y}$$

- f depends on a set of parameters $\boldsymbol{p}$
- the precise form of f is determined on the basis of the training data (*training* or *learning phase*)

So given the dataset:

(1) we choose a model

(2) we **learn** the parameters of the model

(3) we use the learned model $y(x,\mathbf{w}^*)$ to make predictions

Learning from data

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

Terminology/concepts

- Training set
- Training or learning phase
- Generalization
- Test set
- Pre-processing
- Feature extraction
- Pattern recognition ←→ machine learning

Types of machine learning algorithms/systems

1. based on the type of data
2. based on how the data is processed
3. based on how the learned functions generalize

Types of machine learning algorithms
1. Based on the type of data

A. Supervised Learning

Input	Output (target)
x_1	t_1
x_2	t_2
. . .	…
x_N	t_N,

Two types of problems:

a) classification

b) regression

$\boldsymbol{x}$ *is a vector of features* $\boldsymbol{x} = (feature_1, feature_2, \ldots feature_n)$

t could also be a vector

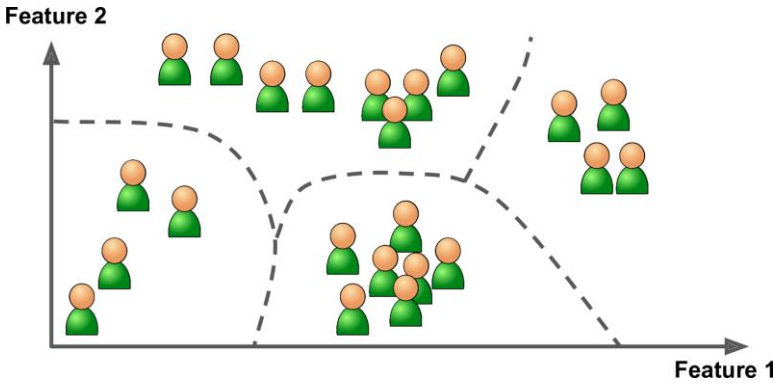
B. Unsupervised Learning

Input
x_1
x_2
. . .
x_N

Examples:

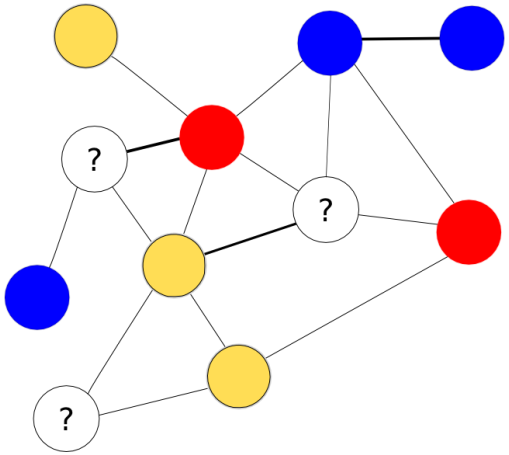
a) clustering
b) density estimation
(e.g. for anomaly detection)

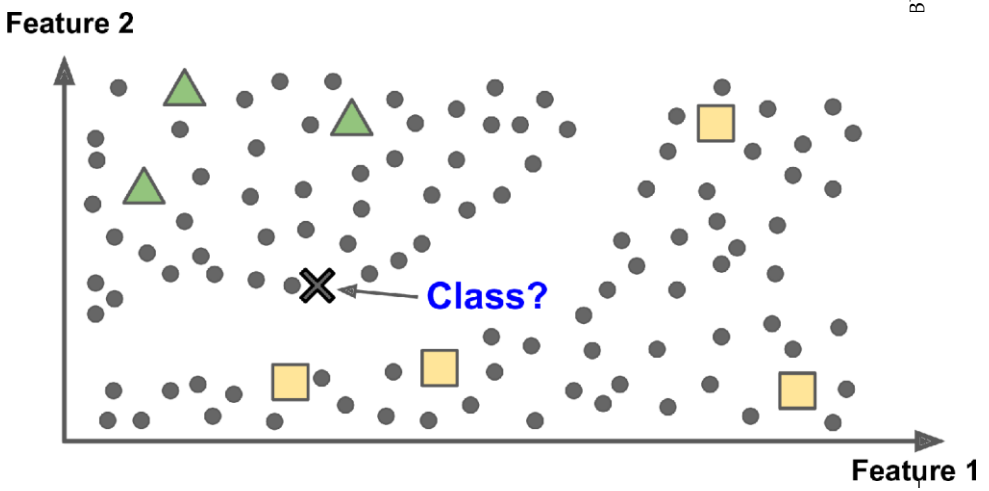
$\boldsymbol{x}$ *is a vector of features* $\boldsymbol{x} = (feature_1, feature_2, \ldots feature_n)$



C. Semi-supervised Learning

Input	Output (target)
x_1	t_1
x_2	t_2
x_3	--
x_4	--
x_5	t_5
. . .	…
x_N	t_N





x is a vector of features $x = (feature_1, feature_2, \ldots feature_n)$

t could also be a vector

D. Reinforcement Learning

- Agent
- sequence of **states** and **actions** in which the learning algorithm interacts with the environment
- only at the end a reward is achieved.
- *credit assignment* problem: the reward must then be attributed appropriately to all of the moves that led to it (good ones and bad ones)
- **Exploration** vs **exploitation**

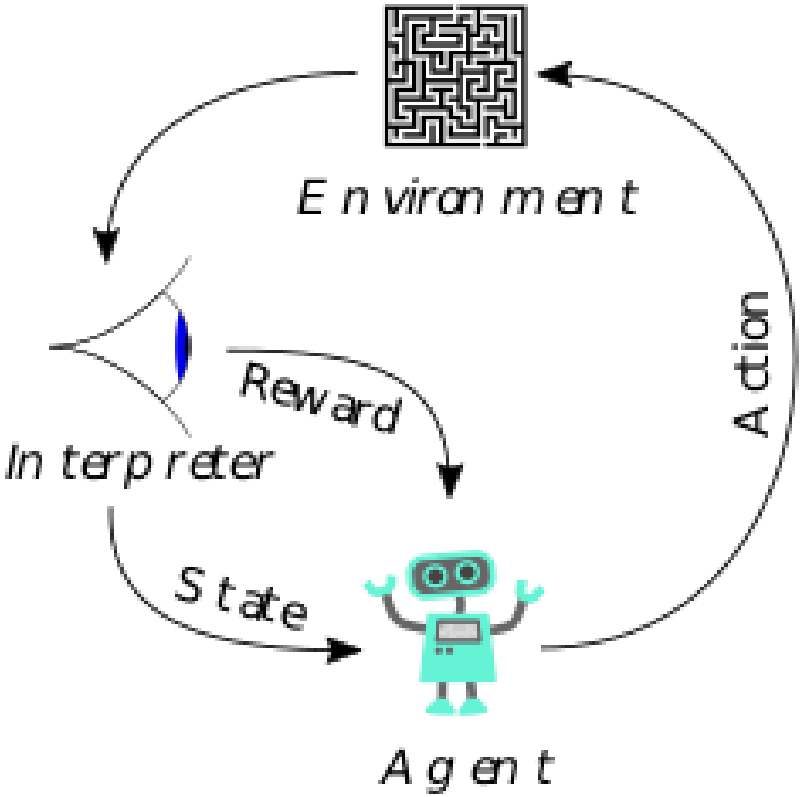
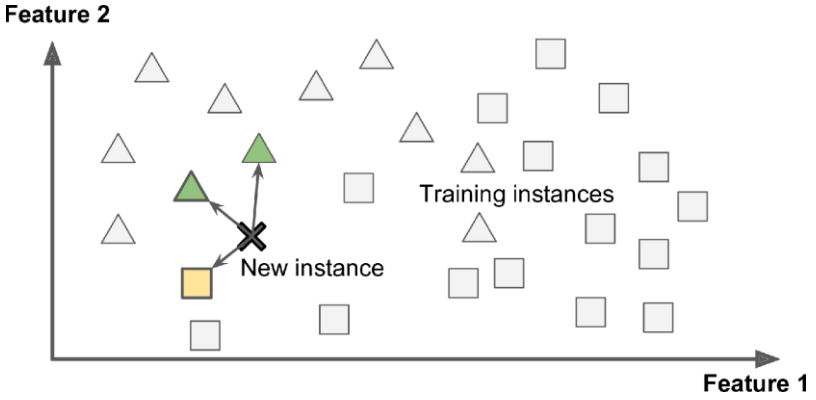


image from Wikipedia

Types of machine learning algorithms
2. Based on based on how the learned functions generalize

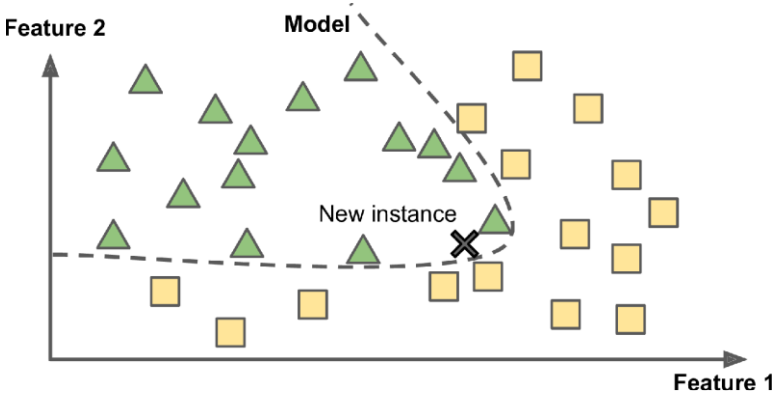
A. Instance based learning:

generalization achieved using a similarity measure to compare new instances to existing (learned) datapoints.



B. Model-based learning:

use the dataset to build a model and then use that model to make predictions.



Types of machine learning systems
3. Based on how the data is processed

A. **Batch learning:** the system is trained using existing data. Typically done offline (offline learning)

B. **Online learning:** you train the system incrementally by feeding it data instances sequentially. For systems that receive data as a continuous flow (e.g., stock prices)