Advanced School in Artificial Intelligence

Introduction to Machine Learning

Elena Bellodi

elena.bellodi@unife.it

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Regione Emilia-Romagna

FERRARIAE UNIVERSITAS · 13 91 · EX LABORE FRUCTUS

Università degli Studi di Ferrara

Outline

- Machine learning (ML) definitions
- Learning paradigms
  - supervised
  - unsupervised
  - reinforcement
  - semi-supervised
- Use of Data in ML
  - training, validation and test set
  - generalization, underfitting and overfitting
  - capacity
  - bias and variance
- Learning protocols

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Outline

- Machine learning (ML) definitions
- Learning paradigms
  - supervised
  - unsupervised
  - reinforcement
  - semi-supervised
- Use of Data in ML
  - training, validation and test set
  - generalization, underfitting and overfitting
  - capacity
  - bias and variance
- Learning protocols

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Machine Learning and AI

- Machine learning (ML) is a field of Artificial Intelligence (AI)

- Definition 1:
  - Learning is constructing or modifying representations of what is being *experienced* [Michalski 1986, pag. 10]
- Definition 2:
  - Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the same task or tasks drawn from the same population *more efficiently and more effectively the next time* [Simon 1984, pag. 28]

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Machine Learning and AI

- **To be intelligent**, a system that is in a changing environment should have the ability to learn. If the system can learn and adapt to such changes, the system designer need not foresee and provide solutions for all possible situations

- Machine learning allows computers to learn and infer from data

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Machine Learning and other disciplines

Multidisciplinary field, it draws results from:

- Artificial intelligence
- Probability and statistics
- Computational complexity theory
- Information theory
- Philosophy
- Psychology
- Neurobiology
- ….

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Machine Learning Goals

- We are overwhelmed with **data**:
  - ubiquitous electronics record our decisions, our choices in the supermarket, our financial habits, our choices in purchases,…
  - World Wide Web

- in all this raw data lies implicit and potentially useful information

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Machine Learning Goals

A. Knowledge extraction

1. To be employed in knowledge-based systems (e.g. in classification systems)
2. To be presented to humans (e.g. for scientific purposes, i.e., discovery of new scientific theories)

B. Improvement of the performances of a machine

- E.g. improvement of the motion and sensing capabilities of a robot

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Machine Learning Goals

- **Prediction** - what can we predict about this phenomenon?
  - case A.1 and B
- **Description** - how can we describe/understand this phenomenon in a new way?
  - case A.2

- Machine Learning algorithms extract explicit **knowledge structures from data** (models) in both cases

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Fielded applications in ML

- Spam emails: How can we predict if something is spam/genuine?
- Medicine/Novel Drugs: What characteristics of a patient indicate they may react well/badly to a new drug? How can we predict whether it will potentially hurt rather then help them?
- Bioinformatics
- Text/audio recognition
- Fault diagnosis
- Marketing and sales
- Search suggestions (Google)
- Information retrieval
- Recommender systems (Amazon)
- Computer vision (screening images)
- Natural language processing (NLP)

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Machine Learning Systems

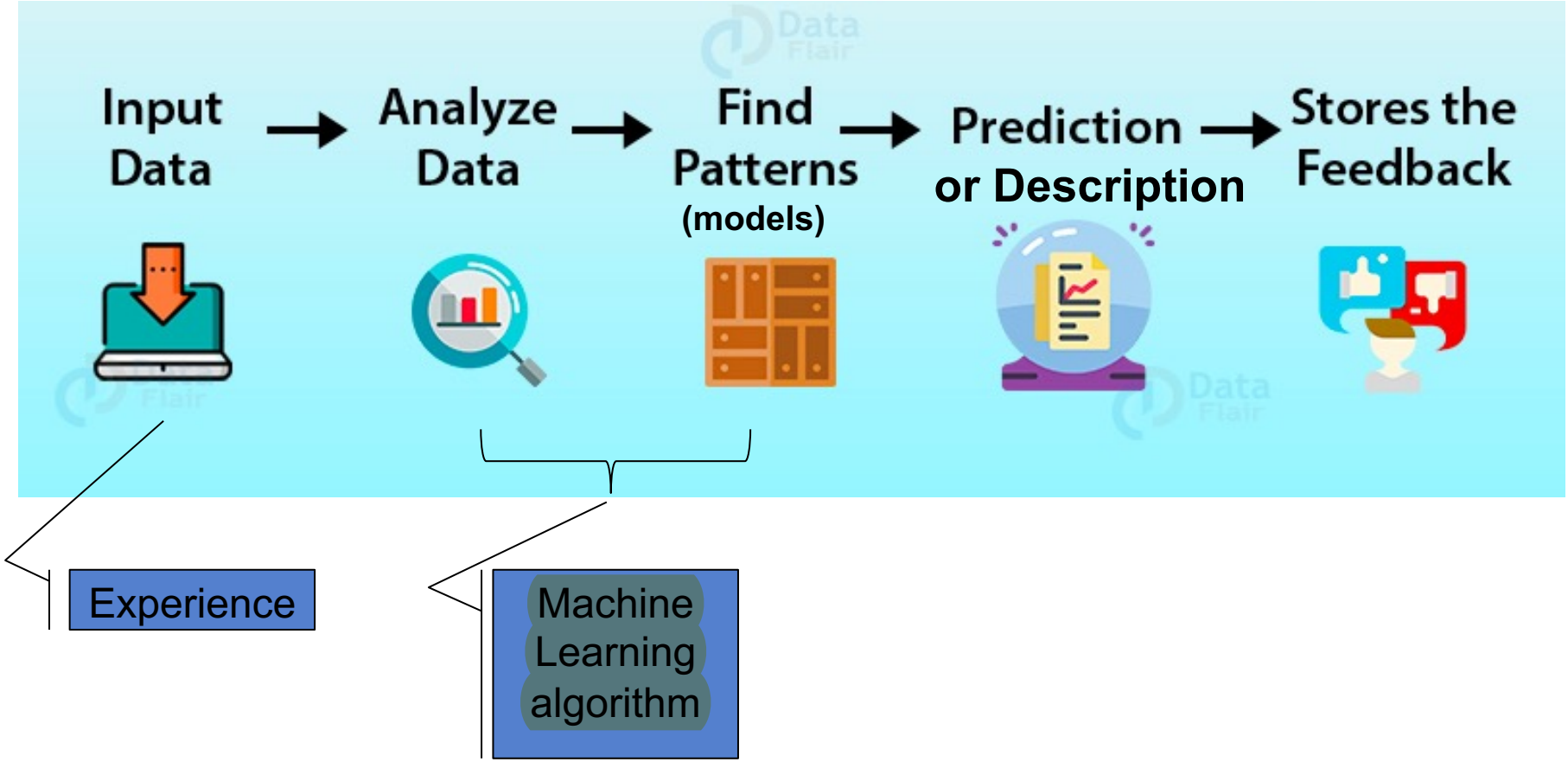
- **Experience**: past information available to the learner, which typically takes the form of electronic data collected and made available for analysis

- In all cases, its quality and size are crucial to the success of the predictions/descriptions made by the learner

- Critical measures of the quality of the ML algorithms
  - time complexity
  - space complexity
  - sample complexity

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

How does Machine Learning work?



Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Machine Learning Vocabulary

- **Dataset:** the whole collection of data points

- **Example:** a single data point *x* within the dataset
  - synonyms: observation, instance, item, instance, datapoint, row, record

- **Features:** set of attributes, often represented as a vector **x**, associated to an example
  - synonyms: predictors, input, independent variables, attributes

- **Label:** a category assigned to examples (not always available!)
  - synonyms: **class**, answer, ***y***, response, output, dependent variable, **target**

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Learning Paradigms

- **Learning with Different Outputs**
  - supervised
  - unsupervised
  - semi-supervised
  - reinforcement

- **Learning with Different Protocol**
  - batch learning
  - online learning
  - active learning

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Learning with Different Outputs

Supervised	Data points have known outputs
Unsupervised	Data points have unknown outputs
Semi-supervised	Data points have some known outputs
Reinforcement	Data points have implicit output by goodness

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Supervised/Unsupervised Learning Goals

- Machine Learning Goals:
  - Prediction - what can we predict about this phenomenon?
  - Description - how can we describe/understand this phenomenon in a new way?

	Predictive model	*Descriptive model*
Supervised learning	classification, regression	subgroup discovery
Unsupervised learning	predictive clustering	descriptive clustering, association rule discovery

Table 1.1. An overview of different machine learning settings. The rows refer to whether the training data is labelled with a target variable, while the columns indicate whether the models learned are used to predict a target variable or rather describe the given data.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Supervised/Unsupervised Learning

1) The most common setting is supervised learning of predictive models – in fact, this is what people commonly mean when they refer to supervised learning.
    - Typical tasks are classification and regression
2) *subgroup discovery:* use labelled training data to build a descriptive model that is not primarily intended to predict the target variable, but instead **identifies subsets of the data that behave differently with respect to the target variable**
3) Descriptive models can naturally be learned in an unsupervised setting (clustering, association rule discovery and matrix decomposition). This is often the implied setting when people talk about unsupervised learning.
4) *predictive clustering:* unsupervised learning of a predictive model where we cluster data with the intention of using the clusters to assign class labels to new data

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

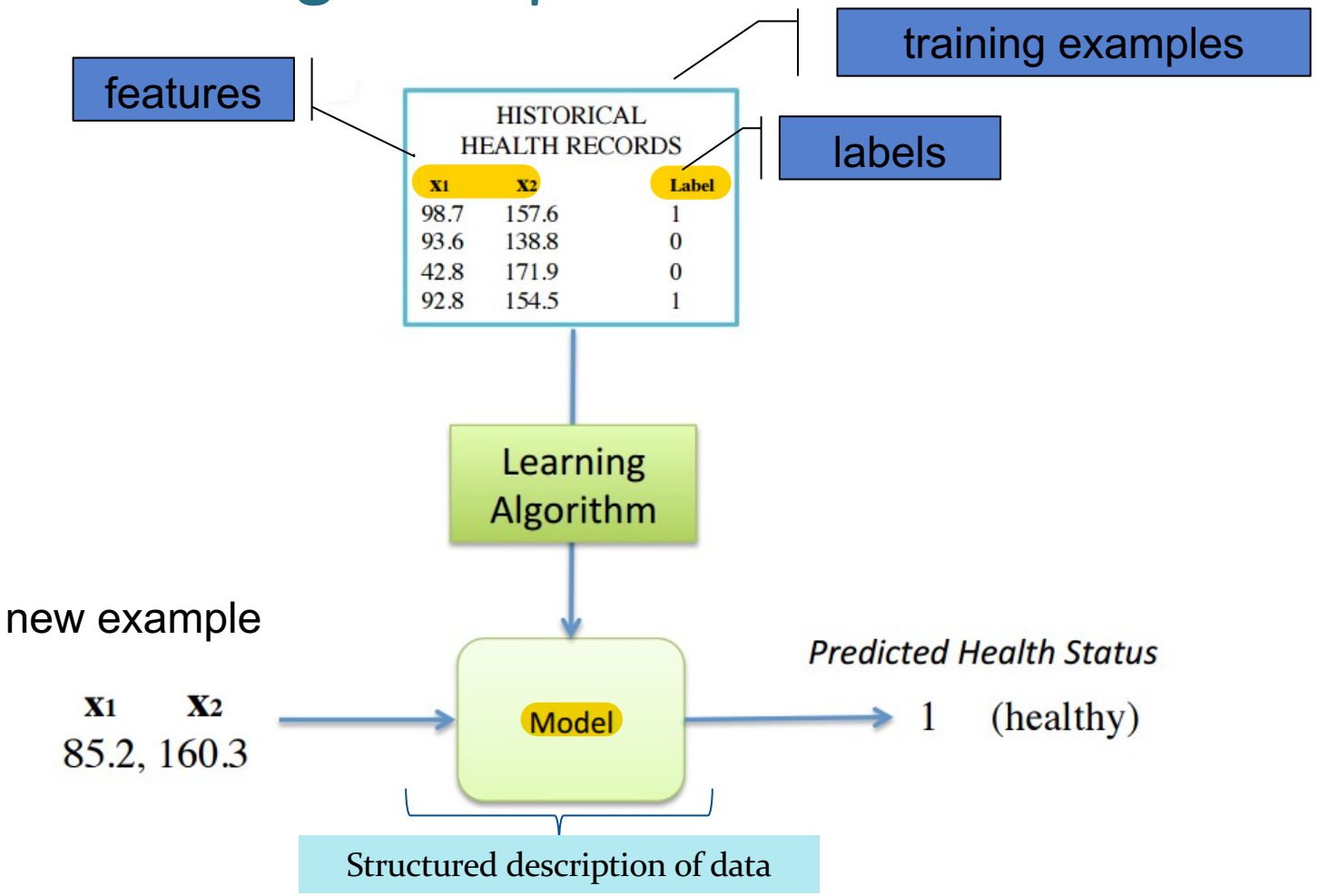
Supervised Learning

- We know (input, correct output) pairs in the «training» examples (input data)
  - for every example *i* its feature vector **x** comes with corresponding y: $(\mathbf{x}_i, y_i)$

- We want to *predict*, given the values of the features of a *new* item **x**, the unknown value of its target y
  - Target in $\mathbb{R}$: regression.
  - Target in {1, . . . , K} or {−1, +1} : classification (multiclass or binary).

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Supervised Learning Example



Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Unsupervised Learning

- Instead of (input, correct output), we get (input, ?) in the training examples (input data)
  - We are just given input vectors $\mathbf{x}_1, \cdots, \mathbf{x}_n$ for *n* examples

- We wish to extract some synthetic information:
  - subsets of similar items (**clustering**)
    - articles ⇒ topics
  - the distribution of items in their domain (**density estimation**)
    - traffic reports with location ⇒ dangerous areas

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

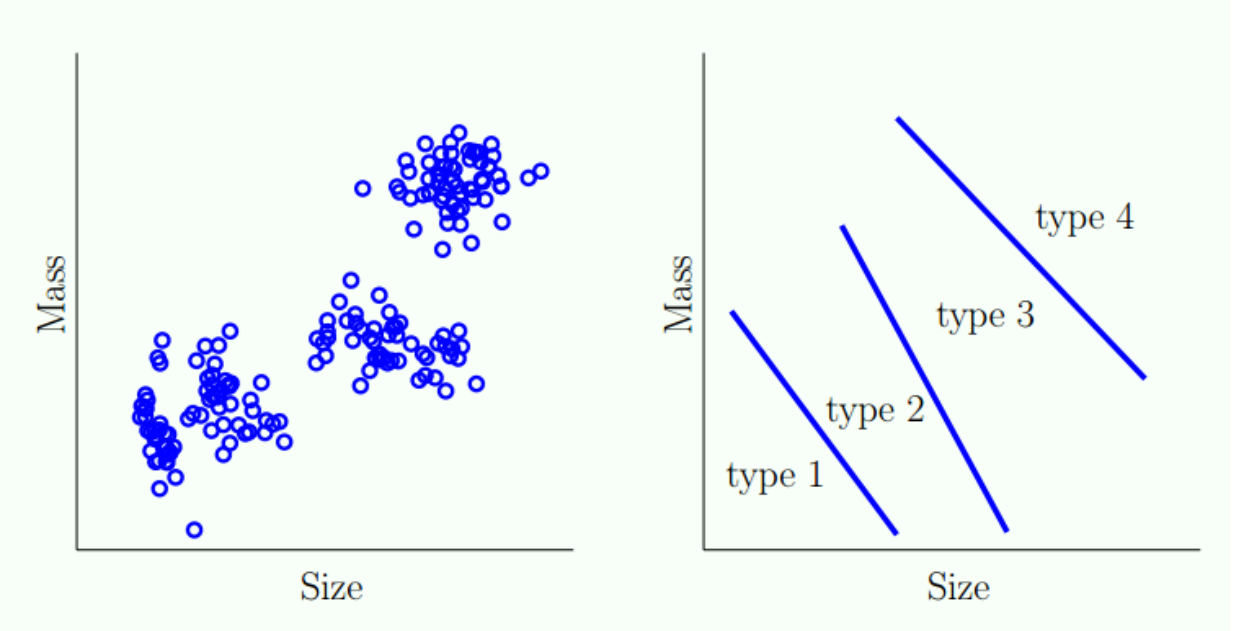
Unsupervised Learning

- The projection of items on lower dimensional subspaces, that is, their characterization by means of a smaller set of features (**feature selection, feature extraction**)
- **Outlier detection**:
  - Internet logs ⇒ intrusion alert

- Typical in Data Mining

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Unsupervised Learning



An unsupervised classification rule treats the four clusters as different types

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Reinforcement Learning

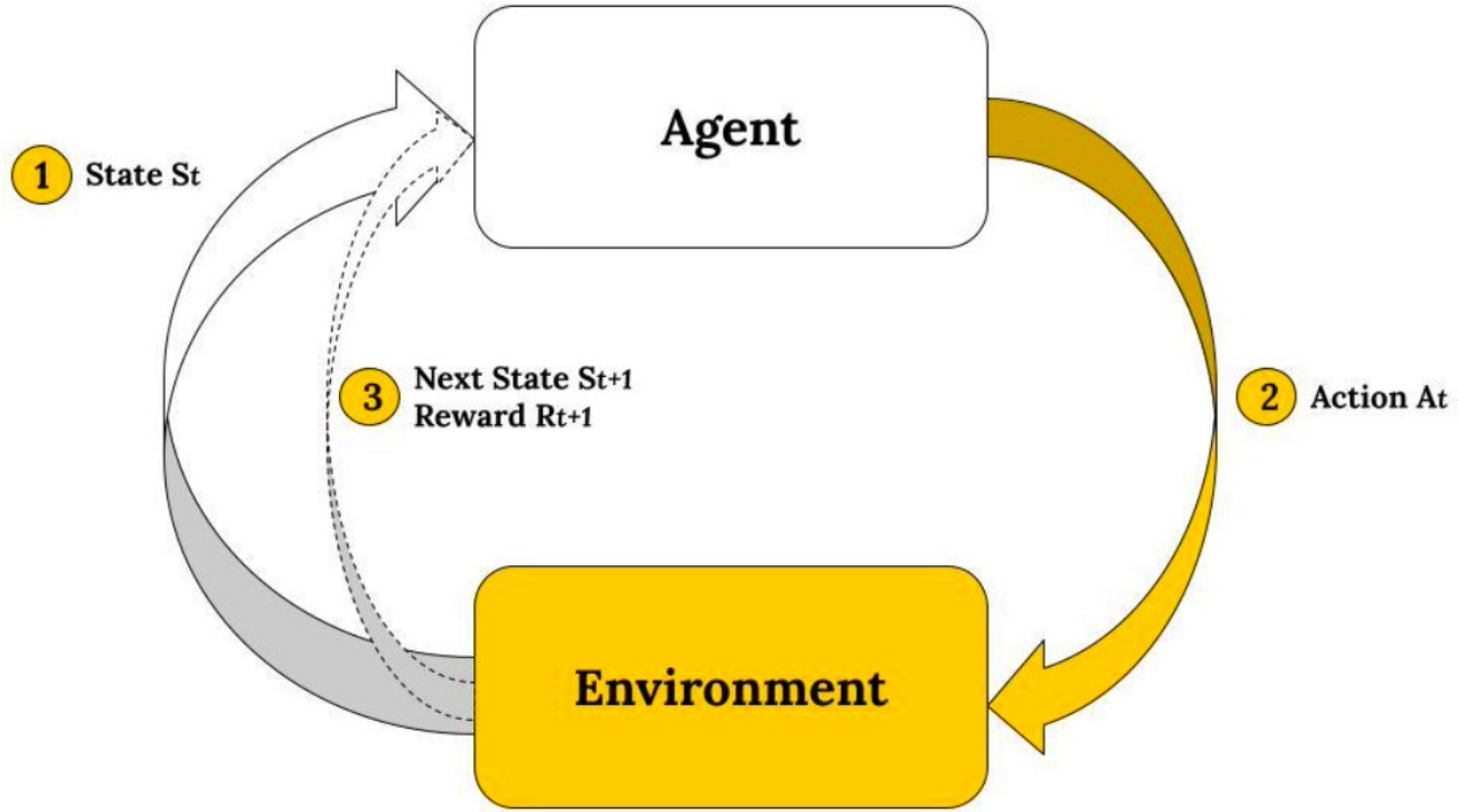
- Instead of (input, correct output), we get

  (input, some output, grade for this output)

- How an autonomous agent that acts in its environment can **learn to choose optimal actions** to achieve its goals?
- Each time the agent performs an action in its environment, a trainer may provide a **reward or penalty** to indicate the desirability of the resulting state
- The task of the agent is to learn from this indirect, delayed reward, to choose sequences of actions that produce **the greatest cumulative reward**

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Reinforcement Learning



Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Reinforcement Learning: example

- A mobile robot with sensors (camera) and actions such as “move forward” and “turn”. Its task is to learn a control strategy, a *policy*, for choosing actions that achieve the goal of docking onto its battery charger whenever its battery level is low
- a *reward* function assigns a numerical value to each distinct action the agent may take from each distinct state
  - the goal of docking to the battery charger can be captured by assigning a **positive reward (e.g. +100) to state-action transitions that immediately result in a connection to the charger and a reward of 0** to every other state-action transition
- the reward function may be built into the robot, or known only to an external teacher who provides the reward value
- the learnt control policy is the one that, from any initial state, chooses actions that **maximize the reward accumulated over time by the agent**

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

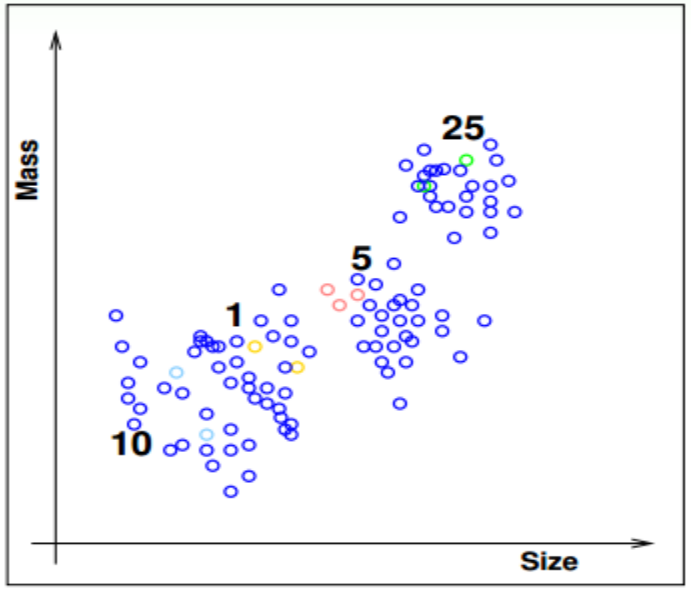
Reinforcement Learning applications

- **Manufacturing** optimization problems
- **Sequential** scheduling problems
- **Playing games:** Google has reinforcement learning agents that learn to solve problems by playing simple games like Go, which is a game of strategy
- **Personalised product recommendation system**: Personalise / customize what products need to be shown to individual users to realise maximum sale
- **Ad recommendation system:** Customise / personalise what Ads need to be shown to the end user to have higher click-through rate. The reinforcement signal is defined as the total click-through rate (CTR) of the ad.
- …

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Semi-supervised Learning (SSL)

- We get (some input, correct output) + (some input, ?)
- The goal is classification, having a small amount of labeled data with a large pool of unlabeled data
- Leverage unlabeled data to avoid 'expensive' manual labeling
  - face images with a few labeled ⇒ face identifier
  - medicine data with a few labeled ⇒ medicine effect predictor



Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Semi-supervised Learning applications

- Speech recognition
- Web content classification
- Text document classification
- Protein Sequence Classification
- Manufacturing
- ….

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021