

Introduction to Supervised Learning

Michela Papandrea
michela.papandrea@supsi.ch

Supervised Learning
Bachelor of Data Science and Artificial Intelligence
University of Applied Sciences and Arts of Southern Switzerland

21.09.2021

Overview

- 1 Introduction
- 2 Data Representation
- 3 Classification vs Regression
 - Definition of Classification
 - Definition of Regression
- 4 Generalization, Overfitting and Underfitting

Machine Learning

- extracting knowledge from data.
- intersection of statistics, artificial intelligence, and computer science (aka *predictive analytics* or *statistical learning*)
- ML applications is ubiquitous
many modern websites and devices have machine learning algorithms at their core

Example

- automatic recommendations of which movies to watch, what food to order or which products to buy,
- personalized online radio streaming
- recognizing friends faces on your photos, inferring age and gender

Why Machine Learning?

Past

intelligent applications involved
handcoded rules:

if-then-else

decisions to process data

Example: spam filter with blacklist of words

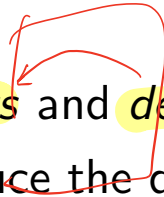
Present

major disadvantages of manually
crafted decision rules

- ① logic required to make a decision is specific to a single domain and task.
⇒ every slight change in the task might require a rewrite of the whole system
- ② human expert deep understanding of how a decision should be made is necessary

Example: faces detection in images

Supervised Learning

- automate decision-making processes by *generalizing* from known examples
 - (**dataset**) data is provided as pairs of *inputs* and *desired outputs*
 - (**model**) the algorithm finds a way to produce the desired output given an input
 - (**generalization**) the algorithm generate autonomously an output for an input it has never seen before
- 

Supervised Learning: the meaning

Supervised Learning

- *Supervised Learning* is a subbranch of *Machine Learning*
- algorithms that learn from examples ($\langle \text{input}, \text{desired output} \rangle$ pairs)
- "*Supervised*" refers to: having a teacher which supervise the whole process
- *supervision* is provided in the form of *desired outputs* for each training example
- require a laborious manual process of inputs and outputs dataset creation
- prediction performances are quantitatively measurable
- **training**: the algorithm will search for patterns in the data that correlate the input with the desired outputs
- **prediction**: the algorithm takes new unseen inputs and determine the output ($\langle \text{new input}, \text{predicted output} \rangle$) based on prior training data
- **objective of a SL model**: predict the correct label for newly presented input data

Examples of supervised machine learning tasks

Example

Identifying the zip code from handwritten digits on an envelope

- **input:** a scan of the handwriting,
- **output:** actual digits in the zip code.
- To create a dataset for building a machine learning model, you need to collect many envelopes. Then you can read the zip codes yourself and store the digits as your desired outcomes.

Example

Determining whether a tumor is benign based on a medical image

- **input:** the image
- **output:** whether the tumor is benign (Y/N)
- To create a dataset for building a model, you need a database of medical images. You also need an expert opinion, so a doctor needs to look at all of the images and decide which tumors are benign and which are not. It might even be necessary to do additional diagnosis beyond the content of the image to determine whether the tumor in the image is cancerous or not.

Examples of supervised machine learning tasks

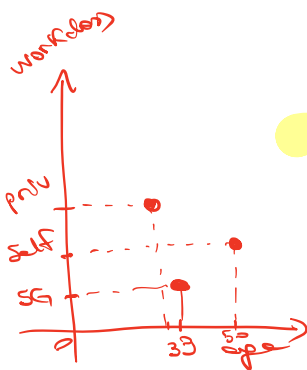
Example

Detecting fraudulent activity in credit card transactions

- **input:** a record of the credit card transaction
- **output:** whether it is likely to be fraudulent or no
- Assuming that you are the entity distributing the credit cards, collecting a dataset means storing all transactions and recording if a user reports any transaction as fraudulent.

Data representation

- Despite the nature of the data, it is important to have a representation of your input data that a computer can understand
- commonly a dataset is representation as a **table**
 - **row** (or **entry**): each **data point** (or **sample**) that we want to reason about
 - **column**: each **property** that describes that data point (**features**)



characteristic, feature, attribute

	age	workclass	education	gender	hours-per-week	occupation	income
0	39	State-gov	Bachelors	Male	40	Adm-clerical	<=50K
1	50	Self-emp-not-inc	Bachelors	Male	13	Exec-managerial	<=50K
2	38	Private	HS-grad	Male	40	Handlers-cleaners	<=50K
3	53	Private	11th	Male	40	Handlers-cleaners	<=50K
4	28	Private	Bachelors	Female	40	Prof-specialty	<=50K
5	37	Private	Masters	Female	40	Exec-managerial	<=50K
6	49	Private	9th	Female	16	Other-service	<=50K
7	52	Self-emp-not-inc	HS-grad	Male	45	Exec-managerial	>50K
8	31	Private	Masters	Female	50	Prof-specialty	>50K
9	42	Private	Bachelors	Male	40	Exec-managerial	>50K
10	37	Private	Some-college	Male	80	Exec-managerial	>50K

object, observation, data sample

Knowing Your Data, and Your Task

- In ML it is not effective to randomly choose an algorithm and throw your data at it.
- Each algorithm is different in terms of what kind of data and what problem setting it works best for

Main steps in a ML analysis:

- 1 Understand the problem we are trying to solve and if the data can solve the problem
- 2 Formalize the problem
- 3 Collect **enough** data to solve the problem
- 4 Identify features and algorithms which allow right predictions
- 5 Define metrics for the performances measurement
- 6 Generate the predictive model and integrate the ML solution within a business product



Supervised Learning: formalism

At its most basic form, a supervised learning algorithm can be written simply as:

$$y = f(x) \quad (1)$$

Handwritten annotations: "predicted output" with an arrow pointing to y , "input" with an arrow pointing to x , and "array" with a double underline and a red 'X' next to it.

Where:

- y is the predicted output that is determined by a mapping function that assigns a class to an input value x . X
- The function used to connect input features to a predicted output is created by the machine learning model during training.



Andreas C. Müller & Sarah Guido (2017)

Introduction to Machine Learning with Python

Chapter 1: Introduction (pp. 1 – 5)

Chapter 2: Supervised Learning (pp. 27 – 31)

Published by O'Reilly Media, Inc..