

# Introduction to machine learning course

## Lecture 1: Introduction and presentation of the course

Alexis Zubiolo

`alexis.zubiolo@gmail.com`

Data Science Team Lead @ Adcash

January 19, 2017

## Before we start

I'd like to know a little bit more about you

- ▶ Short presentation: Name, occupation, ...
- ▶ Background in machine learning?
- ▶ Background in programming?
- ▶ Background in mathematics?
- ▶ Expectations from the course (if any)?

Please send me an email so that I have your contact:

`alexis.zubiolo@gmail.com`

All the material will be available on my personal GitHub:

`https://github.com/azubiolo/itstep`

# Outline

- ▶ What machine learning is, what it is not
- ▶ A few practical examples
  - ▶ classification
  - ▶ regression
- ▶ Goals and presentation of the course
- ▶ Questions and answers

# What is machine learning?

A simple example. . .



How to filter spam emails **automatically**?

# Machine learning paradigm

Goal: Build algorithms that can

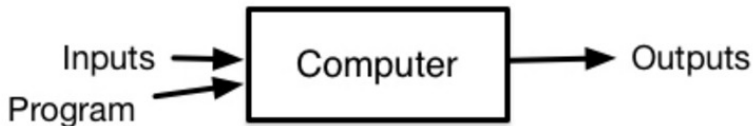
- ▶ **learn** from data
- ▶ **make predictions** on (new) data

# Machine learning paradigm

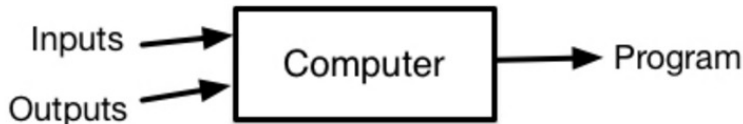
Goal: Build algorithms that can

- ▶ **learn** from data
- ▶ **make predictions** on (new) data

## Traditional Programming



## Machine Learning



# Main components of machine learning

## Mathematics

- ▶ Linear algebra
- ▶ Calculus
- ▶ Numerical optimization

## Statistics, probability theory

## Computer science

I will assume some knowledge in computer science (= a language you are comfortable with).

## Example 1: Regression

Regression = output is a **continuous** numerical value

Example: **Estimate the price** of an apartment

- ▶ input: **information** about the apartment
- ▶ output: **price**



## Example 1: Regression

Regression = output is a **continuous** numerical value

Example: **Estimate the price** of an apartment

- ▶ input: **information** about the apartment
- ▶ output: **price**

living area (m <sup>2</sup> )	price (1000's euros)
50	30
76	48
26	12
102	90

## Example 1: Regression

Regression = output is a **continuous** numerical value

Example: **Estimate the price** of an apartment

- ▶ input: **information** about the apartment
- ▶ output: **price**

living area (m <sup>2</sup> )	price (1000's euros)
50	30
76	48
26	12
102	90
61	?

Linear model:  $\text{price} = \mathbf{a} \times \text{area} + \mathbf{b}$

Problem: optimal values for **a** and **b**?

# Regression

More data for a richer model:

living area (m <sup>2</sup> )	# bedrooms	price (1000's euros)
50	1	30
76	2	48
26	1	12
102	3	90
61	2	?

**Linear model:**  $\text{price} = \mathbf{a} \times \text{area} + \mathbf{b} \times \# \text{ bedrooms} + \mathbf{c}$

**Problem:** Optimal values for **a**, **b** and **c**?

**Remark:** More data does not always imply a better model

## Example 2: Classification

Classification = output is a **label**

Examples:

## Example 2: Classification

Classification = output is a **label**

Examples:

- ▶ Spam filtering
  - ▶ input: email (text, subject, address, ...)
  - ▶ output: **spam** or **not spam**

## Example 2: Classification

Classification = output is a **label**

Examples:

- ▶ Spam filtering
  - ▶ input: email (text, subject, address, ...)
  - ▶ output: **spam** or **not spam**
- ▶ Object recognition in images or videos
  - ▶ input: image or video
  - ▶ (example) output: **face** or **not a face**

## Example 2: Classification

Classification = output is a **label**

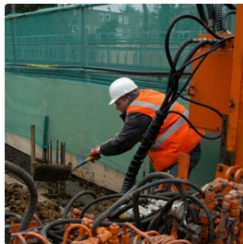
Examples:

- ▶ Spam filtering
  - ▶ input: email (text, subject, address, ...)
  - ▶ output: **spam** or **not spam**
- ▶ Object recognition in images or videos
  - ▶ input: image or video
  - ▶ (example) output: **face** or **not a face**
- ▶ Image classification/description
  - ▶ input: image
  - ▶ output: image **description** or **label** (apple, car, ...)

# Automated image description generation



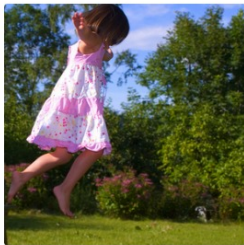
"man in black shirt is playing guitar."



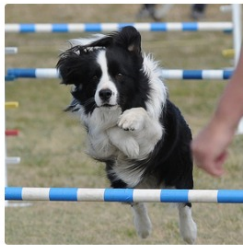
"construction worker in orange safety vest is working on road."



"two young girls are playing with lego toy."



"girl in pink dress is jumping in air."



"black and white dog jumps over bar."



"young girl in pink shirt is swinging on swing."



## Other topics

Machine learning is a wide and growing field. It also includes:

## Other topics

Machine learning is a wide and growing field. It also includes:

- ▶ Clustering (no predefined label/output)

## Other topics

Machine learning is a wide and growing field. It also includes:

- ▶ Clustering (no predefined label/output)
- ▶ Dimensionality reduction

## Other topics

Machine learning is a wide and growing field. It also includes:

- ▶ Clustering (no predefined label/output)
- ▶ Dimensionality reduction
- ▶ ...

# The course

Goals:

- ▶ Introduce **main concepts** of machine learning
- ▶ **Implement** these concepts in Python (Scikit-learn)

# The course

## Goals:

- ▶ Introduce **main concepts** of machine learning
- ▶ **Implement** these concepts in Python (Scikit-learn)

## Practical information:

- ▶ ~ **10 90-min sessions**
- ▶ Alternating between **lectures and lab sessions**
- ▶ I will send you the slides + code after each course

# Course outline

- ▶ **General introduction** to ML and Python (course + lab)
- ▶ **Regression** model (course + lab)
- ▶ **Classification** models (course + lab)
- ▶ **Clustering** models (course + lab)
- ▶ **Evaluation** methods
- ▶ **General conclusion** + questions + remarks and feedback

**Note:** This is a first estimation. I will adapt to your needs.

# Python and Scikit-learn

## Why Python?

- ▶ Intuitive, interpreted language
- ▶ Cross-platform (Windows, MacOS, Linux)
- ▶ An active open-source community
- ▶ Easy-to-use ML libraries (e.g. Scikit-learn)

Let's see an example. . .

First lab session (about Python) next Thursday

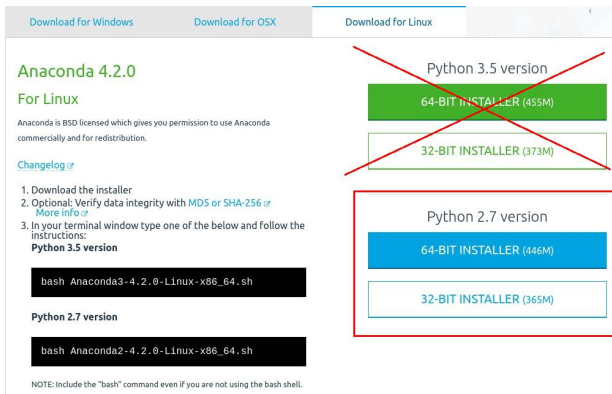


# Want to work on your own computer?

Recommended setup:

## ► **Anaconda** for Python 2.7

<https://docs.continuum.io/anaconda/install>



The screenshot shows the Anaconda Linux download page. The 'Download for Linux' tab is selected. The page title is 'Anaconda 4.2.0 For Linux'. Below the title, it states 'Anaconda is BSD licensed which gives you permission to use Anaconda commercially and for redistribution.' and provides a 'Changelog' link. A list of three steps is provided: 1. Download the installer, 2. Optional: Verify data integrity with MD5 or SHA-256 (with a 'More info' link), and 3. In your terminal window type one of the below and follow the instructions. Two terminal commands are shown: one for Python 3.5 version and one for Python 2.7 version. On the right side, there are two sections for installers. The 'Python 3.5 version' section is crossed out with a large red X and contains a green '64-BIT INSTALLER (455M)' button and a light green '32-BIT INSTALLER (373M)' button. The 'Python 2.7 version' section is highlighted with a red rectangular box and contains a blue '64-BIT INSTALLER (446M)' button and a light blue '32-BIT INSTALLER (365M)' button. A note at the bottom states: 'NOTE: Include the "bash" command even if you are not using the bash shell.'

Download for Windows    Download for OSX    **Download for Linux**

### Anaconda 4.2.0

#### For Linux

Anaconda is BSD licensed which gives you permission to use Anaconda commercially and for redistribution.

[Changelog](#)

1. Download the installer
2. Optional: Verify data integrity with [MD5](#) or [SHA-256](#) [More info](#)
3. In your terminal window type one of the below and follow the instructions:

**Python 3.5 version**

```
bash Anaconda3-4.2.0-Linux-x86_64.sh
```

**Python 2.7 version**

```
bash Anaconda2-4.2.0-Linux-x86_64.sh
```

NOTE: Include the "bash" command even if you are not using the bash shell.

**Python 3.5 version**

**64-BIT INSTALLER (455M)**

**32-BIT INSTALLER (373M)**

**Python 2.7 version**

**64-BIT INSTALLER (446M)**

**32-BIT INSTALLER (365M)**

## ► **Spyder** text editor

Thank you! Questions?

Don't forget to send me an email:  
`alexis.zubiollo@gmail.com`