Machine learning from scratch

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
- Big picture of a machine learning algorithm
- 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

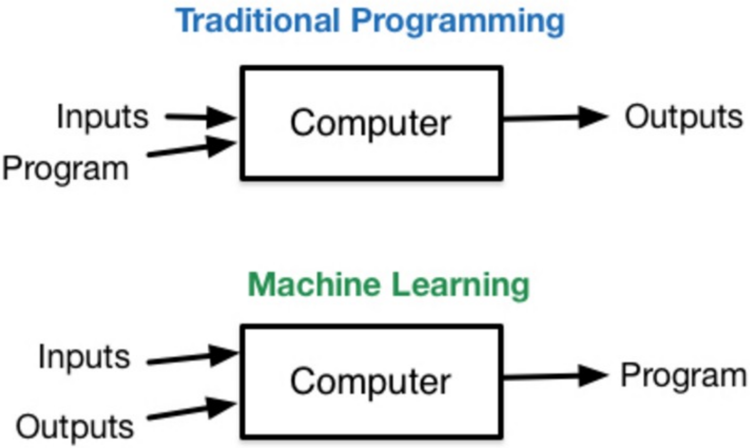
Goal: Build algorithms that can

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

Machine learning paradigm

Goal: Build algorithms that can

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



Main components of machine learning

- Mathematics
  - Linear algebra
  - Calculus
  - Numerical optimization
- Statistics, probability theory
- Computer science

Main components of machine learning

- Mathematics
  - Linear algebra
  - Calculus
  - Numerical optimization
- Statistics, probability theory
- Computer science

In the course, we will review these aspects.

Prerequisites: I will assume

- Some knowledge in computer science (understand: at least **a language you are comfortable with**)
- You do not pass out when you see a mathematical formula

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^2)	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^2)	price (1000's euros)
50	30
76	48
26	12
102	90
61	?

Linear model: price = **a** × area + **b**

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

Regression

More data for a richer model:

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

Linear model: price = **a** × area + **b** × # bedrooms + **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:

- Unsupervised learning/clustering (no predefined label/output)

Other topics

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

- Unsupervised learning/clustering (no predefined label/output)
- Dimensionality reduction

Other topics

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

- Unsupervised learning/clustering (no predefined label/output)
- Dimensionality reduction
- Feature engineering

Other topics

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

- Unsupervised learning/clustering (no predefined label/output)
- Dimensionality reduction
- Feature engineering
- . . .

This course will focus on **supervised learning**.

What machine learning is not

Even though ML can provide great results, it is not a magic black box that solves all issues.

ML users/engineers need proper understanding and some experience.

ML Algorithm: Big Picture

There are several key steps when using supervised learning. Several pieces have to be wisely chosen:

- A **data** set
- A **model**
- A **loss** function
- A **regularization**
- An **optimizer**

ML Algorithm: Big Picture

There are several key steps when using supervised learning. Several pieces have to be wisely chosen:

- A **data** set
- A **model**
- A **loss** function
- A **regularization**
- An **optimizer**

These choices have to take into account a few constraints, depending on the application, *e.g.*:

- A minimum **accuracy** (or other performance index)
- **Time** constraints
- **Resources** constraints (storage, computation power, architecture, . . .)

ML Algorithm

In this course, we will focus on

- **Models**: linear, kernel, ...
- **Loss** functions: Least squares, logistic loss, ...
- **Regularization**: ℓ_2 or ℓ_1
- **Optimization** techniques: Stochastic/batch gradient descent
- **Evaluation** of models

The course

Goals:

- Understand **how a supervised ML algorithm works**
- Being **able to implement a ML algorithm**
- Anything else you might have in mind

The course

Goals:

- Understand **how a supervised ML algorithm works**
- Being **able to implement a ML algorithm**
- Anything else you might have in mind

Practical information:

- $\sim$ **10 60-90 min sessions** on Thursdays at 6:30 pm
- Starting with a few lectures about the main concepts followed by lab sessions where you implement these concepts
- All material will be **available on GitHub**, with links to extra material for those who want to go deeper

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

Course outline (attempt)

Course outline (attempt)

- **Mathematical background**
  - Linear algebra (vector, matrices, operations)
  - Derivatives (gradient, Hessian matrix)
  - Convexity

Course outline (attempt)

- **Mathematical background**
  - Linear algebra (vector, matrices, operations)
  - Derivatives (gradient, Hessian matrix)
  - Convexity
- **Mathematical formalization of ML problems**
  - Linear models
  - Kernels
  - Loss functions (least squares, logistic regression, SVM)
  - Regularization

Course outline (attempt)

- **Mathematical background**
  - Linear algebra (vector, matrices, operations)
  - Derivatives (gradient, Hessian matrix)
  - Convexity
- **Mathematical formalization of ML problems**
  - Linear models
  - Kernels
  - Loss functions (least squares, logistic regression, SVM)
  - Regularization
- **Optimization in machine learning**
  - Gradient descent
  - Stochastic vs. batch methods
  - Second-order methods
  - Learning rate

Course outline (attempt)

- **Mathematical background**
  - Linear algebra (vector, matrices, operations)
  - Derivatives (gradient, Hessian matrix)
  - Convexity
- **Mathematical formalization of ML problems**
  - Linear models
  - Kernels
  - Loss functions (least squares, logistic regression, SVM)
  - Regularization
- **Optimization in machine learning**
  - Gradient descent
  - Stochastic vs. batch methods
  - Second-order methods
  - Learning rate
- **Model combination** (boosting)

Course outline (attempt)

- **Mathematical background**
  - Linear algebra (vector, matrices, operations)
  - Derivatives (gradient, Hessian matrix)
  - Convexity
- **Mathematical formalization of ML problems**
  - Linear models
  - Kernels
  - Loss functions (least squares, logistic regression, SVM)
  - Regularization
- **Optimization in machine learning**
  - Gradient descent
  - Stochastic vs. batch methods
  - Second-order methods
  - Learning rate
- **Model combination** (boosting)
- **Model validation**

Course outline (attempt)

- **Mathematical background**
  - Linear algebra (vector, matrices, operations)
  - Derivatives (gradient, Hessian matrix)
  - Convexity
- **Mathematical formalization of ML problems**
  - Linear models
  - Kernels
  - Loss functions (least squares, logistic regression, SVM)
  - Regularization
- **Optimization in machine learning**
  - Gradient descent
  - Stochastic vs. batch methods
  - Second-order methods
  - Learning rate
- **Model combination** (boosting)
- **Model validation**

Note: This is a first rough estimation. I will adapt to your needs and how fast things go.

About programming languages

For the practical sessions, I will be using **Python** with **Jupyter**.

`http://jupyter.org/`

If you prefer another language, feel free to use it. Remember that I assume some programming knowledge.

Thank you! Questions?

alexis.zubiolo@gmail.com

https://github.com/azubiolo/itstep