Introducción al aprendizaje automático y Datasets

Módulo Aprendizaje Máquina (automático)

TC3006C

Basic Concepts

Introduction

What is learning?

- Ability to
  - use percepts from the outside world
    - not only for reacting,
    - but for **improving actions** in future events.

- Implies that we know **when** and **how** to use this new knowledge.
  - When: pattern detected
  - How: algorithm created.

What is machine learning?

- Example:
  - Imagine a supermarket chain with a hundred of stores selling groceries to millions of customers.
  - Each sale has a lot of data that can be analyzed and converted into information.
  - These information can be used to give people suggestions when buying.
- If we knew who would buy an item, we would just write code for the computer to remind them.
- Because we do not know, we collect data and hope to extract enough information to recommend articles to people.

What is machine learning?

- Example:
  - In RoboCup agents play soccer.
  - There are 11 players against 11 players.
  - Each team has its own strategy for playing soccer.
- If we knew which strategy a team is using, we would play a counter-attack strategy to stop them.
- Because we do not know their strategy, we collect data and try to extract enough information to detect their strategies.
- Once strategies are detected and classified, we could select the best strategy to exploit this knowledge.

What is ML? …

- The computer algorithm should be able to:
  - Identify **patterns** in the data (When)
  - Construct a good and useful **approximation** of the solution to the problem (How)

What is ML? …

- "Machine learning uses data and answers to discover rules behind a problem" Chollet (2017)

- "Machine learning is programming computers to optimize a performance criterion using example data or past experience."
  Alpaydin, E. (2004)

- Has a model defined for some parameters.
  - Learning is the execution of a computer program to optimize the parameters of the model using training data or past experience.

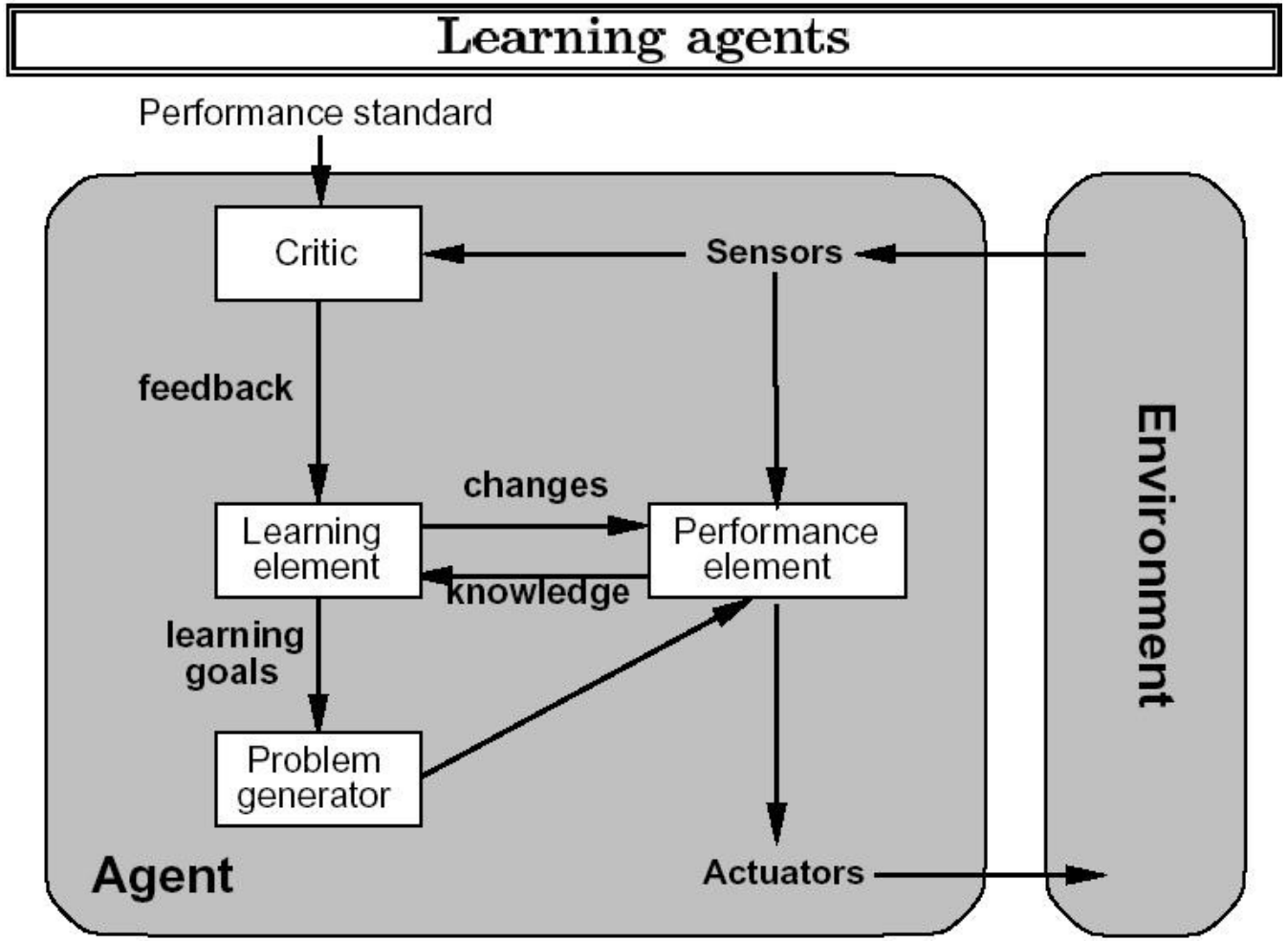
- Two types of models:
  - Predictive model: predictions in the future.
  - Descriptive model: gain knowledge from data.

What is ML? …

- "A computer program is said to **learn** from experience *E* with respect to some class of tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*." Mitchell, T. (1997)

- Example: handwriting recognition
  - Task *T*: recognizing and classifying handwritten words within images.
  - Performance measure *P*: percent of words correctly classified
  - Training experience *E*: a database of handwritten words with given classifications

Learning Agent



Feedback

- Components can be learned from appropriate feedback.
  - Example: training Tae Kwon Do, Driving a Taxi.

- Type of feedback:
  - The most important factor in determining the nature of the learning problem.

  - Three cases:
    1. Supervised learning
    2. Unsupervised learning
    3. Reinforcement learning

Supervised Learning

- Learning a function from examples of its inputs and **outputs**.
  - There is an input X, an output Y, and the task is to learn mapping from input to output.
- Outputs values can be provided
  - By a supervisor – someone feed the output.
  - By the environment – detected by sensors.
- Examples:
  - Learn a condition-action rule for punching.
  - Learn to differentiate between a dog and a cat.

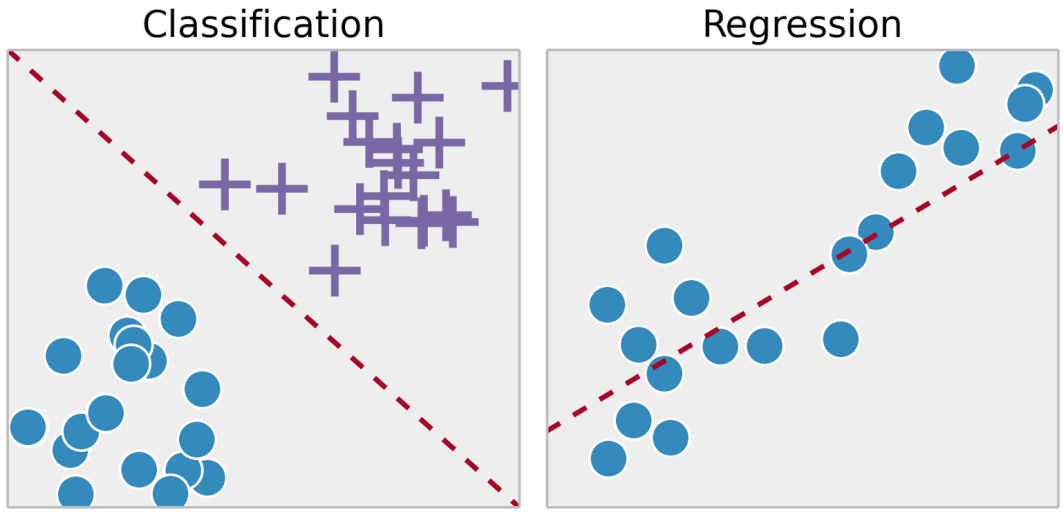
Supervised Learning

Classification

- Output y is one of a finite set of values
  - Example of sets:
    - Red, blue, yellow, green
    - Hot, mild, cold
- If set has only two elements is called Boolean (Binary) Classification

Regression

- Output y is a number
  - Examples:
    - Temperature
    - Velocity
- **Consistent hypothesis**: function h agrees with all the data



https://miro.medium.com/max/3200/1*ASYpFfDh7XnreU-ygqXonw.png

Unsupervised learning

- Learning patterns in the input when **no specific output** values are supplied.
- Aim: to find regularities in the input.
- There are two main categories of algorithms:
  - Clustering
    - Discover inherent grouping in data
  - Association
    - Discover rules that describe large portions of data
- Example:
  - Learn to separate colors.
  - Learn when it might rain.
  - Learn how to detect people that will not pay their credit cards.

Semi-supervised learning

- Mix between supervised and unsupervised learning
- Some data is labelled – usually a very small part
- Labelled data is used to create more data
- Learner learns to:
  - Generate labelled data and to
  - Detect regularities in the input

Reinforcement learning

- The output of the system is a **sequence of actions**.
- Uses rewards to guide the sequence of actions
- These actions are part of a **policy**.
  - A single action is not important.
  - The policy is what must be learned.
- Agent must learn from reinforcement which actions are best, i.e., the policy.
- Examples:
  - Playing chess.
  - Driving politely.
  - Robot navigation.

Representation of the learned information

- Polynomials
- Propositional logic
- Predicate calculus
- Bayesian networks
- Neural networks
- Etc.

Applications of machine learning

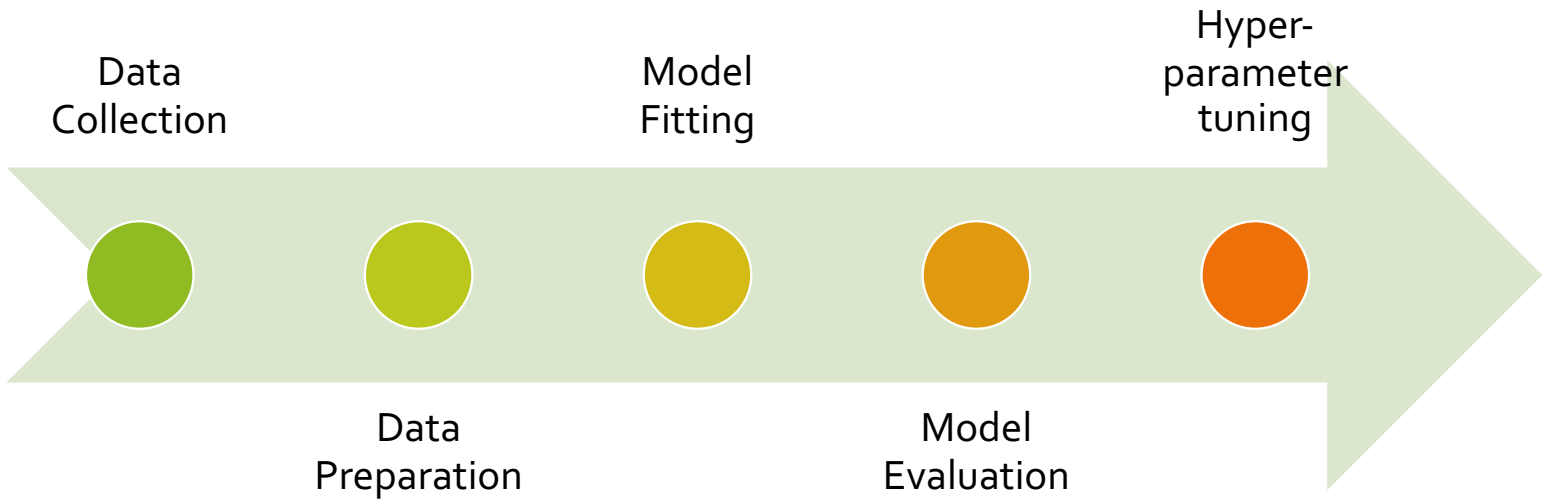
- Learning associations
  - Learn how people associate elements (ex. buying groceries)
- Classification
  - Learn to classify elements in different categories
- Prediction
  - Learn to predict if some action will happen
- Pattern recognition
  - Learn to find familiar patterns (characters, faces, objects, etc.)
- Knowledge extraction
  - Learning a rule from data – it explains the data
  - Rules are a form of data compression
- Outlier detection
  - Data that does not belong to a class
- Regression problems
  - Learn the curve that best fits a function to a set of points

ML is multidisciplinary

- Artificial Intelligence
- Bayesian methods
- Computational complexity theory
- Control theory
- Information theory
- Philosophy
- Psychology and neurobiology
- Statistics

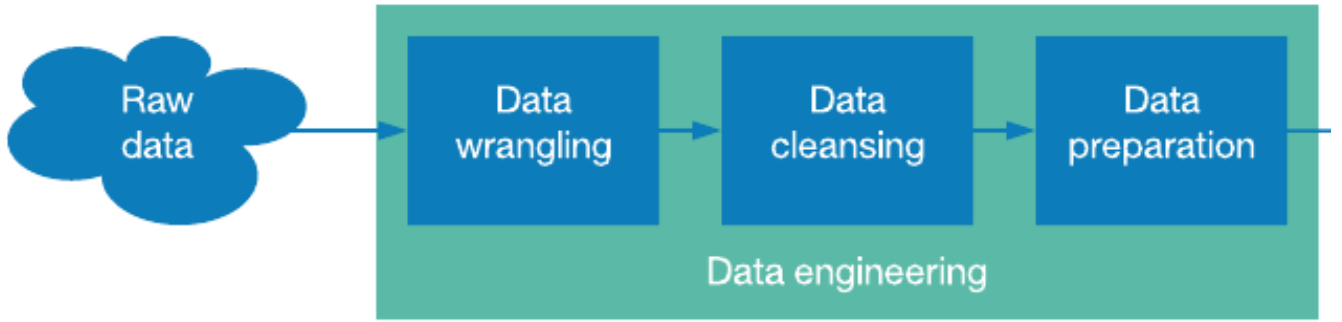
Designing a learning system

ML Process


Data Collection
Data Preparation
Model Fitting
Model Evaluation
Hyper-parameter tuning

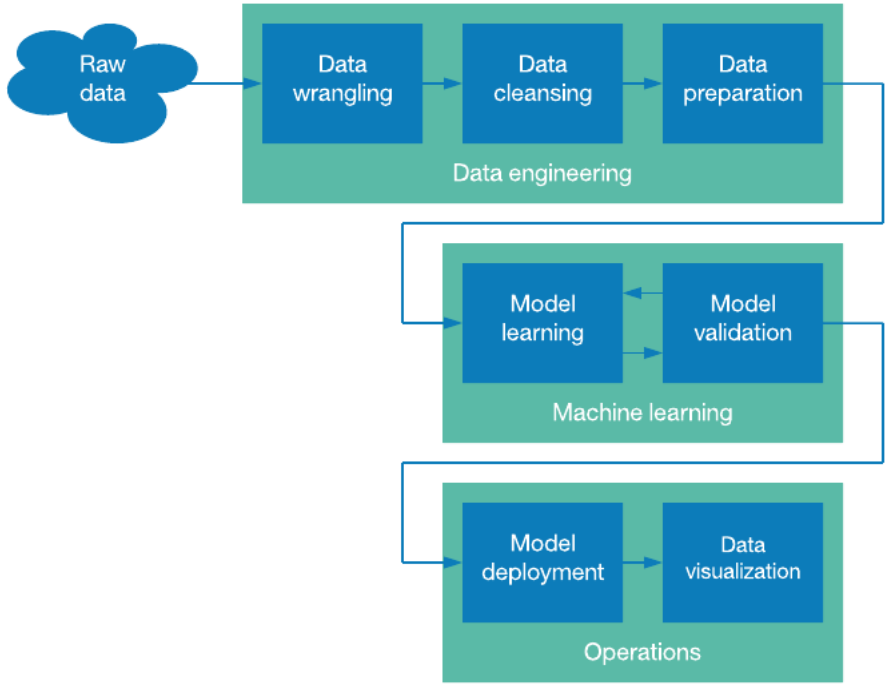
Data Preparation

- Data Wrangling
  - Data might be in different files
  - Cleaning, structuring, enriching raw data
  - Assure quality and useful data
- Data Cleansing
  - Missing values (delete?)
  - Unwanted characters
  - Unwanted elements
- Data Preparation
  - Analysis and optimization of features
  - Select/remove features
  - Consider prediction needs and computation time



https://miro.medium.com/max/666/0*ScsuON73dMJDC9XO.png

Complete data science pipeline



https://developer.ibm.com/articles/ba-intro-data-science-1/

Representing data

Text

Feature representation

List of words showing their frequency count

Feature	Count
Subject	1
Material	2
Del	2
Curso	2
From	1
…	…

```
Subject: Material del curso
From: "Jorge Adolfo Ramírez Uresti" <juresti@tec.mx>
To: "Miguel González Mendoza" <mgonza@tec.mx>
Content-Type: multipart/alternative; boundary="0000000000000272e505e4cc0f55"

--0000000000000272e505e4cc0f55
Content-Type: text/plain; charset="UTF-8"
Content-Transfer-Encoding: quoted-printable

Hola Miguel:
Tan solo un breve mensaje para comentarte que estoy haciendo el material
del curso.

Seguimos en contacto,

Jorge
```

Image

Feature representation

Matrix of color values

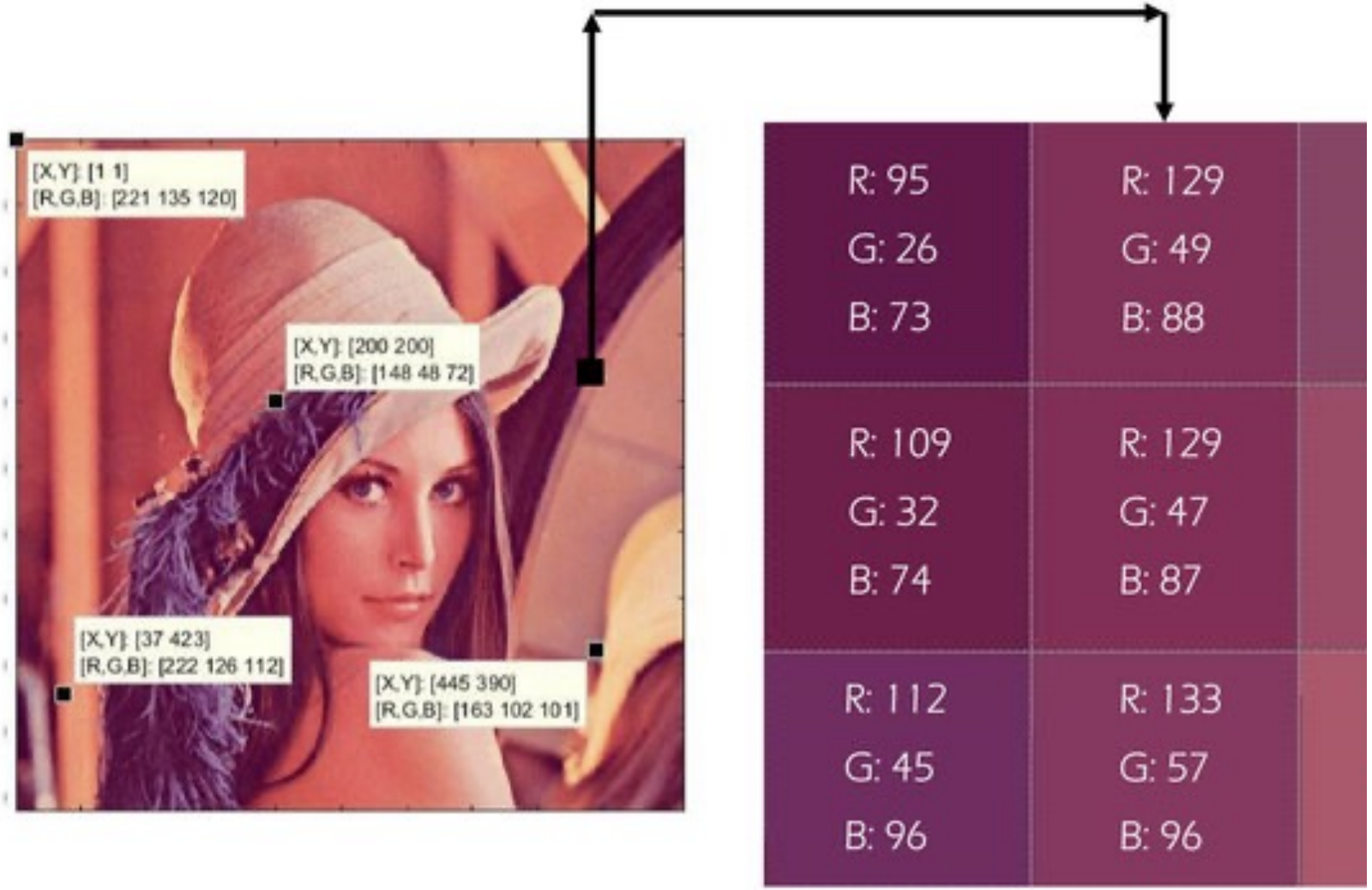


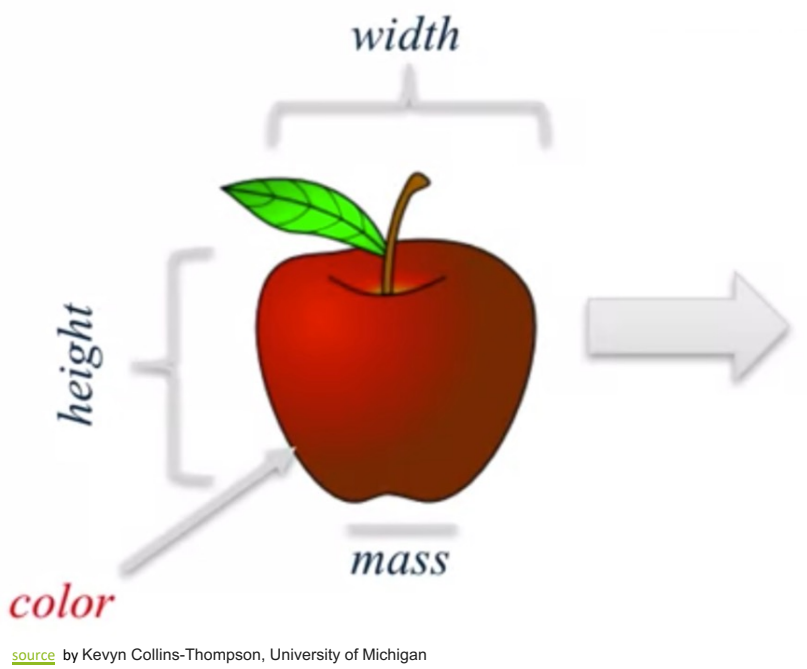
Image form Heidari, Shahrokh & Pourarian, Mohammad Rasoul & Gheibi, Reza & Naseri, Mosayeb & Houshmand, Monireh. (2017). Quantum red–green–blue image steganography. International Journal of Quantum Information. 15. 1750039. 10.1142/S0219749917500393.

Objects

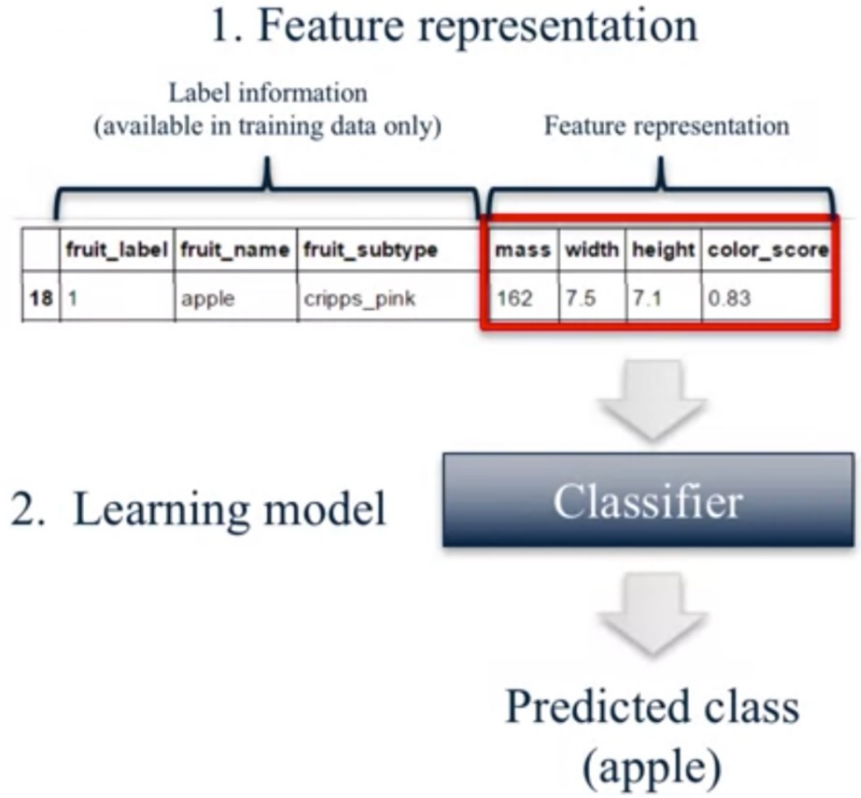
Feature representation

Set of attribute values

Feature	Value
Color	Black
Legs	4
Tail	Yes
Length	2.4 mts
…	…



source by Kevyn Collins-Thompson, University of Michigan



Representing Data

How to build a dataset

Structured Data

ML models learn from examples

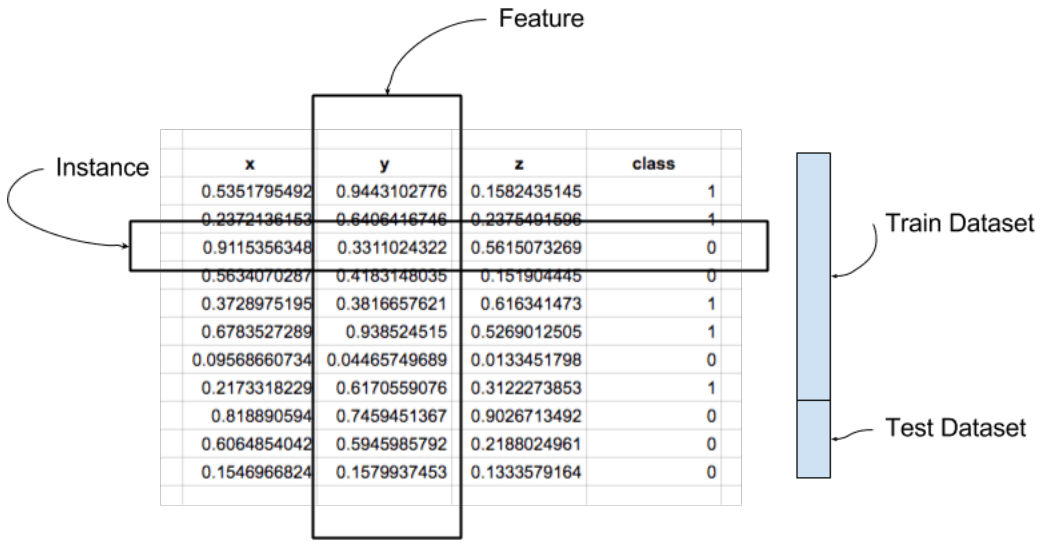
Each example is called an **instance** or pattern

Dataset is formed with multiple examples

Structured Data is organized in rows and columns

A column is called a **feature**

Images, videos and text are called Unstructured Data



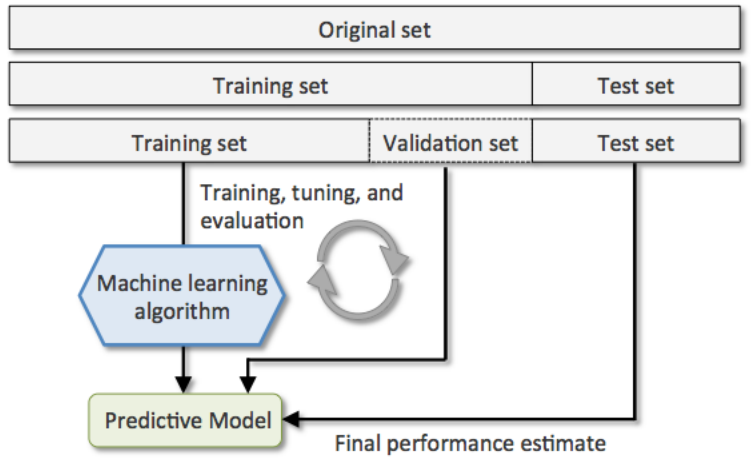
https://machinelearningmastery.com/wp-content/uploads/2013/12/Table-of-Data-Showing-an-Instance-Feature-and-Train-Test-Datasets.png

Dataset organization and division

Training set is usually 80% of original set

Test set is usually 20%

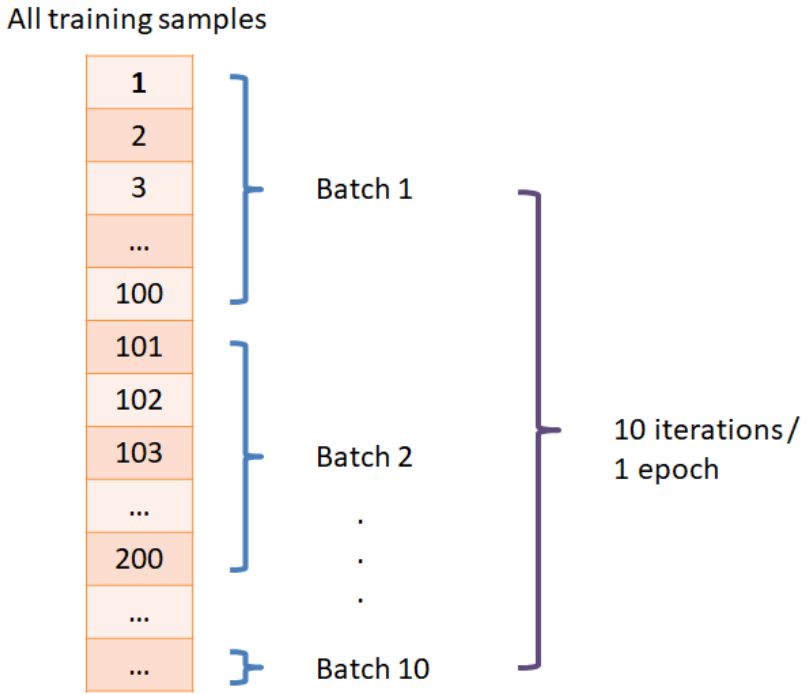
Validation set is usually 20% of training set



https://miro.medium.com/max/585/0*lbveKaL-MGRgppD8.png

Epoch, batch & iteration

- When data is too big and we can't pass all data to computer at once.
- One **Epoch** is when an *entire* dataset is passed forward and backward through the learning model only *once*.
- **Batch size**: divide dataset into *number of batches* or sets or parts.
- **Iterations** is the number of batches needed to complete one epoch.
- *The number of batches is equal to number of iterations for one epoch.*



https://towardsdatascience.com/epoch-vs-iterations-vs-batch-size-4dfb9c7ce9c9

Example datasets

Datasets

- Dataset for regression
  - Iris Data Set
  - https://archive.ics.uci.edu/ml/datasets/iris
- Dataset for classification
  - Wine Data Set
  - https://archive.ics.uci.edu/ml/datasets/wine

Repositories

- UC Irvin https://archive.ics.uci.edu/ml/index.php
- Data set from https://index.okfn.org/place/
- UN http://data.un.org/
- World bank https://data.worldbank.org/

Exploring a dataset

Using pandas

- Example code

Class Exercise

- Go into one of the repositories links
- Download a dataset
- Open it in a plain text editor (notepad, nano, pico, etc…) and locate the instances, the features/attributes, the values of the attributes, the labels/classes.
- Now load it into pandas in your own notebook

References

- Alpaydin, Ethem (2004). *Introduction to Machine Learning*. The MIT Press.
- Mitchell, Tom (1997). *Machine Learning*. WCB McGraw-Hill.
- Edwards, Gavin (2018). *Machine Learning, an introduction*. Towards Data Science (https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0)