Machine Learning Foundations

Big Data

Kristiania University College

Objectives

1. Define Machine Learning (ML)
2. Delimite ML scope
3. Introduce the main ML tasks
4. Recognize problems as ML tasks

Bibliography

- Bishop, Christopher M. *Pattern Recognition and Machine Learning*. 2nd edition. Springer-Verlag. 2011
- Müller, Andreas C., Guido, Sarah. *Introduction to Machine Learning with Python*. O'Reilly. 2016

Table of Contents

Introduction

Justification

New opportunities

- Huge amount of new data sources: banking, social media, IoT, DNA, ...
- Increased computational power

New needs

- Manual data analysis is unfeasible
- Need of automatic methods

New goal

- Transform data into knowledge

Introduction

Definition (I)

ML definition

ML is the science (and art) of programming computers so they can *learn from data*.

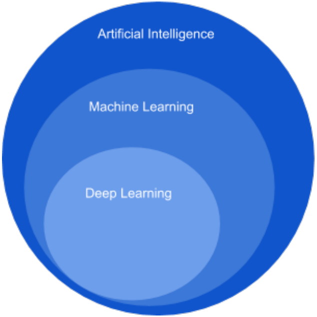
A. Géron, 2017

Alternative definitions

- Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed. Arthur Samuel, 1959.
- A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience. E. Tom Mitchell, 1997.

Introduction

The alphabet soup of data analysis



(Source)

Many related terms

- Artificial Intelligence
- Machine Learning
- Deep Learning
- Big Data

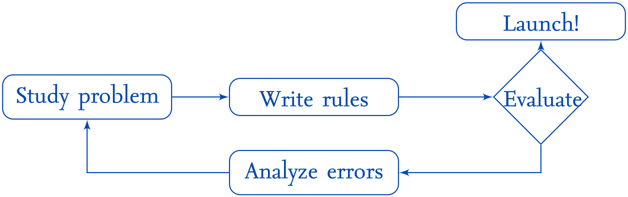
And new careers

- Data Science
- Data scientist
- Data engineer
- ML engineer

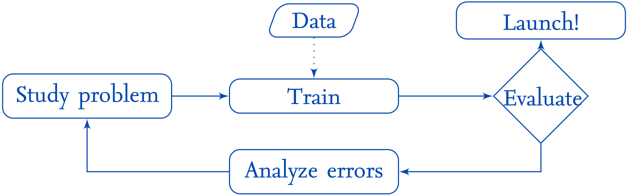
Introduction

Definition (II)

Traditional approach

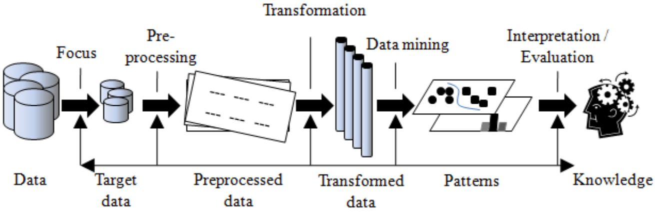


ML approach



The data analysis process

The big picture



(Source)

Steps in any ML application:

1. Data adquisition
2. Selection, cleaning and transformation (preprocessing)
3. Machine Learning
4. Learning evaluation
5. Explotation

The goal in ML is to get a representation of those patterns

The data analysis process

Data adquisition

Goal: Adquire data to perform ML

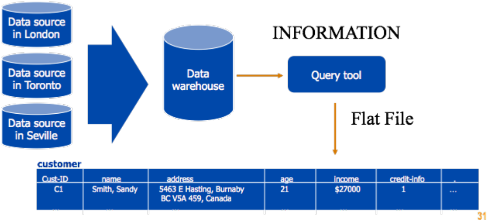
- From extremely easy -CSV file- to extremely complex -full Big Data system-

Public data repositories

- (Kaggle), (NASA Open Data Portal) , (UCI Machine Learning Repository)

Customized adquisition and integration

- Integration from several data sources usually needed



The data analysis process

Selection, cleaning and transformation (I)

Goal: Prepare data for ML

- This phase is usually named preprocess

ML requires a clean data table

- Rows are named instances
- Columns are named features or attributes
- We refer the number of features as dimensionality

In some ML problems we use graphs instead of tables

f_1	f_2	$\cdots$	f_n
$a_{1,1}$	$a_{2,1}$	$\cdots$	$a_{n,1}$
$a_{1,2}$	$a_{2,2}$	$\cdots$	$a_{n,2}$
$a_{1,3}$	$a_{2,3}$	$\cdots$	$a_{n,3}$
$a_{1,4}$	$a_{2,4}$	$\cdots$	$a_{n,4}$
$a_{1,5}$	$a_{2,5}$	$\cdots$	$a_{n,5}$

The data analysis process

Selection, cleaning and transformation (II)

Example: Bank data base

IDC	Years	Euros	Salary	Own house	Defaults
101	15	60000	2200	Yes	2
102	2	30000	3500	Yes	0
103	9	9000	1700	Yes	1
104	15	18000	1900	No	0
...	...	...	...	...	...

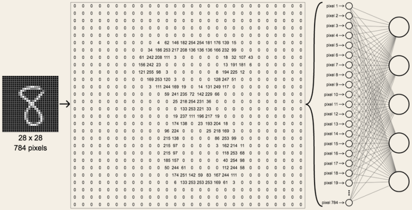
Example: Robot sensors

Timestamp	Sonar1	Sonar2	Sonar3	Sonar4
1	1.687	0.445	2.332	0.429
2	0.812	0.481	1.702	0.473
3	1.572	0.471	1.654	0.513
...	...	...	...	...

The data analysis process

Selection, cleaning and transformation (III)

Example: Image recognition



(Source)

Pixel1	Pixel2	Pixel3	⋯	Pixel784
o	o	o	⋯	o
⋯	⋯	⋯	⋯	⋯
o	o	o	⋯	o

The data analysis process

Selection, cleaning and transformation (IV)

Example: Text classification (bag-of-words representation)

1. Original text

```
(1) John likes to watch movies. Mary likes movies too.
(2) John also likes to watch football games.
```

2. Build list

```
(1) "John","likes","to","watch","movies","Mary","likes","movies","too"
(2) "John", also","likes","to","watch","football","games"
```

3. Build dictionary

```
(1) {"John":1,"likes":2,"to":1,"watch":1,"movies":2,"Mary":1,"too":1};
(2) {"John":1,"also":1,"likes":1,"to":1,"watch":1,"football":1,"games":1};
```

John	likes	to	watch	movies	Mary	too	also	games	⋯
1	2	1	1	2	1	1	0	0	⋯
1	1	1	1	0	0	0	1	1	⋯

The data analysis process

Selection, cleaning and transformation (V)

Preprocessing tasks

- Handle outliers (remove or leave them)
- Sample data (in case there are too much)
- Handle missing values
- Remove irrelevant or redundant features (feature selection)
    - For instance, attributes “social class” and “salary” contain highly correlated information
- Compute new attributes (feature engineering)
    - For instance, compute “population density” from “area” and “population”
- Transform attributes
    - Discretization, normalization, numerization, ...

The data analysis process

Machine Learning

Goal: Train an algorithm to perform a task

- As result, we obtain a model (or classifier or predictor depending on the context)

Machine Learning training methods (or ML tasks)

- Supervised learning: **classification** and regression
- Unsupervised learning: **clustering**, association, **dimensionality reduction** and anomality detection
- Reinforcement learning
- Many others

No Free-Lunch Theorem

No learning algorithm is a priori guaranteed to work better
More info: (D. Wolpert, 1996)

The data analysis process

Learning evaluation (I)

We do need to evaluate the trained model

- Models should perform well on new data

A naïve and wrong approach. Why is it wrong?

1. Train the model
2. Use the model to predict labels
3. Compute accuracy comparing predicted labels with known labels

Solution: Training and validation datasets

- **Training set**: Data used to train the models. Usually 70 %
- **Validation set**: Data used to validate the models. Usually 30 %
- Problems: Bias and loose of relevant data (serious in small datasets)

The data analysis process

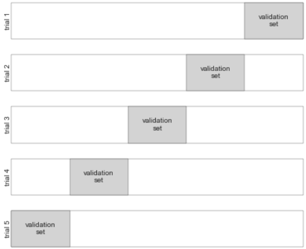
Learning evaluation (II)

Crossvalidation

1. Divide dataset in folds
2. Take one fold for validation
3. Train with the other folds
4. Validate and compute performance
5. Take another fold and repeat until finish
6. Average performance measures

Usually we use 10 folds

- 10-fold cross validation (or 10-CV)



(Source)

The data analysis process
Learning evaluation (III)

Select a measure to evaluate learning

- Proper measures depends on the problem

Classification learning measures

- Accuracy: Ratio of correct predictions
- F-Measure
- Confusion matrix
- ROC curve

Regression learning measures

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- R^2

Validation error must be taken, always, on the validation set

Confusion matrix

Actual class \ Predicted class	Class A	Class B	Class C
Class A	100	0	10
Class B	10	80	10
Class C	30	0	70

(Source)

The data analysis process

Model exploitation

Model explotation depends on the objectives

- In Data Science, the model is interpreted and a report wroten
    - Formal report, bussiness intelligence dashboard, ...
- In Machine Learning, the model is integrated into a software system
    - Web application, app, robot controller, ...

The model may need maintenance

Types of Machine Learning systems

Overview

We can classify ML systems based on several (non-exclusive) criteria

- Whether or not they are trained with human supervision
  - Supervised, unsupervised, semisupervised and Reinforcement Learning
- Whether or not they can learn incrementally
  - Online vs. batch learning
- Whether they compare new data to known data
  - Instance-based vs. model-based learning
- The purpose of the system
  - Predictice models vs. explicative models
- The goal of the system
  - Discriminative models vs. generative models

We focus on supervised and unsupervised model-based discriminative batch algorithms.

Types of Machine Learning systems

Supervised learning (I)

In supervised learning input data comes along with the desired output

- Usually human beings label the output (named labels)

f_1	f_2	$\cdots$	f_n	y
$a_{1,1}$	$a_{2,1}$	$\cdots$	$a_{n,1}$	y_1
$a_{1,2}$	$a_{2,2}$	$\cdots$	$a_{n,2}$	y_2
$a_{1,3}$	$a_{2,3}$	$\cdots$	$a_{n,3}$	y_3
$a_{1,4}$	$a_{2,4}$	$\cdots$	$a_{n,4}$	y_4
$a_{1,5}$	$a_{2,5}$	$\cdots$	$a_{n,5}$	y_5

Two main tasks in supervised learning

- **Classification** if y is a categorical attribute. Target attribute named class
- **Regression** if y is numerical

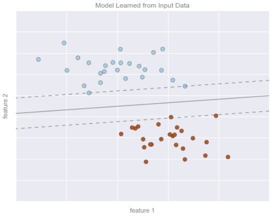
Advanced supervised learning tasks

- Semi-supervised learning, weakly supervised learning and multilabel classification

Types of Machine Learning systems

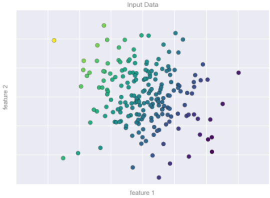
Supervised learning (II)

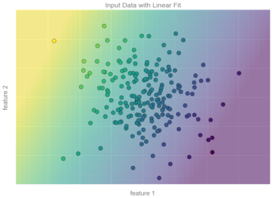
Classification



(Source)

Regression





(Source)

Types of Machine Learning systems

Supervised learning (III)

Important *classification* algorithms:

- k-Nearest Neighbors
- Support Vector Machines (SVMs)
- Decision Trees
    - ID3, C4.5 (J48), ...
- Rules
    - PART, CN2, AQ, ...
- Random Forests
- Bayesian Networks
- Neural Networks
- Ensambles

Important *regression* algorithms:

- Linear Regression
- Logistic Regression
- Symbolic Regression
- Regression trees
    - LM3 (M5), ...
- Neural Networks

Types of Machine Learning systems

Supervised learning: Classification (I)

Example: Bank credit risk management

IDC	Years	Euros	Salary	Own house	Defaulter accounts	Returns credit
101	15	60000	2200	Yes	2	No
102	2	30000	3500	Yes	0	Yes
103	9	9000	1700	Yes	1	No
104	15	18000	1900	No	0	Yes
105	10	24000	2100	No	0	No
...	...	...	...	...	...	...

Objective: Predict if a customer would return a credit or not

Types of Machine Learning Systems

Supervised learning: Classification (II)

Años	Euros	Salario	Casa propia	Cuentas morosas	Crédito
10	50000	3000	Si	0	??

Años	Euros	Salario	Casa propia	Cuentas morosas	Crédito
15	60000	2200	Si	2	No
2	30000	3500	Si	0	Si
9	9000	1700	Si	1	No
15	18000	1900	No	0	Si
10	24000	2100	No	0	No
	...	...	...	...	...

Algoritmo ML

IF CM >0 **THEN** NO

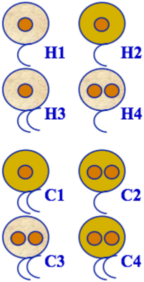
IF CM =0 **Y** S>2500 **THEN SI**

Crédito = Si

Types of Machine Learning systems

Supervised learning: Classification (III)

Example: Cancerous cells prediction

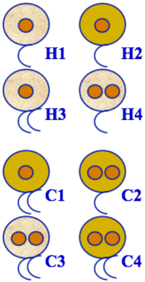


ID	Colour	nuclei	tails	*class*
H1	light	1	1	*healthy*
H2	dark	1	1	*healthy*
H3	light	1	2	*healthy*
H4	light	2	1	*healthy*
C1	dark	1	2	*cancer*
C2	dark	2	1	*cancer*
C3	light	2	2	*cancer*
C4	dark	2	2	*cancer*

Types of Machine Learning systems

Supervised learning: Classification (IV)

Example: Cancerous cells prediction



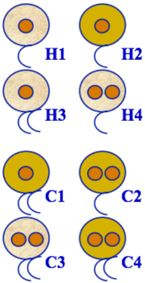
Decision rules

```
if colour = light and nuclei = 1
then cell = healthy ^^I
^^I^^I^^I
if nuclei = 2 and colour = dark
then cell = cancerours

(and 4 rules more)
```

Types of Machine Learning systems

Supervised learning: Classification (V)

Example: Cancerous cells prediction



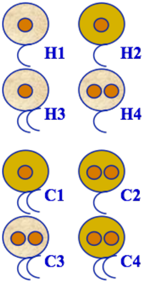
Hierarchical decision rules

```
if colour = light and nuclei = 1
then cell = healthy   ^^I
^^I^^I^^I
else
    if nuclei = 2 and colour = dark
    then cell = cancerous

  else
        if tails = 1
        then cell = healthy

        else cell = cancerous
```

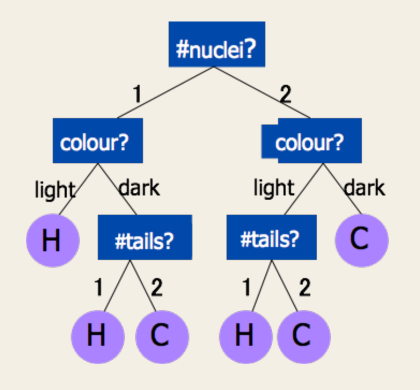
Types of Machine Learning systems

Supervised learning: Classification (VI)

Example: Cancerous cells prediction



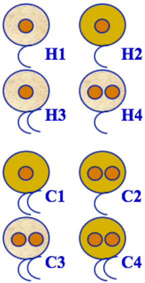
Decision tree



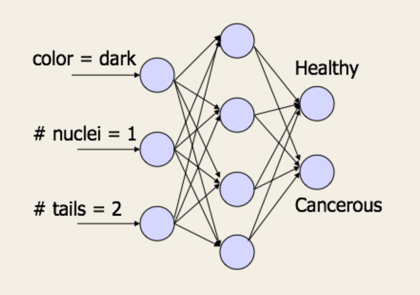
Types of Machine Learning systems

Supervised learning: Classification (VII)

Example: Cancerous cells prediction



Neural network



Types of Machine Learning systems

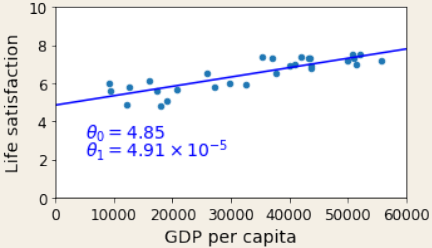
Supervised learning: Regression (I)

Example: Does money make people happier? (example from (Géron, 2017))

Country	GDP	*LS*
Hungary	12,240	4.9
Korea	27,195	5.8
France	37,675	6.5
Australia	50,962	7.3
USA	55,805	7.2

LS =Life satisfaction

Linear regression



$$\mathit{life\_satisfaction} = \theta_0 + \theta_1 \times \mathrm{GDP\_per\_capita}$$

Types of Machine Learning systems

Unsupervised learning

In unsupervised learning there are no labels

f_1	f_2	f_3	$\cdots$	f_n
$a_{1,1}$	$a_{2,1}$	$a_{3,1}$	$\cdots$	$a_{n,1}$
$a_{1,2}$	$a_{2,2}$	$a_{3,2}$	$\cdots$	$a_{n,2}$
$a_{1,3}$	$a_{2,3}$	$a_{3,3}$	$\cdots$	$a_{n,3}$
$a_{1,4}$	$a_{2,4}$	$a_{3,4}$	$\cdots$	$a_{n,4}$
$a_{1,5}$	$a_{2,5}$	$a_{3,5}$	$\cdots$	$a_{n,5}$

Tasks in unsupervised learning

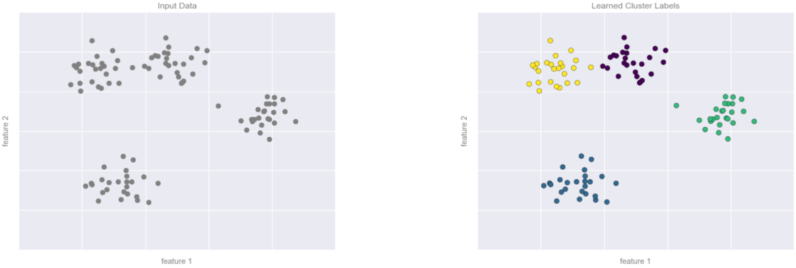
- Clustering
- Association rules
- Dimensionality reduction
- Anomality detection

Types of Machine Learning systems

Unsupervised learning: Clustering (I)

Clustering is a set of techniques that identify groups of data (clusters)

- Algorithms: K-means, db-scan, Gaussian Mixture Models (GMM), Expectation Maximization (EM), ...

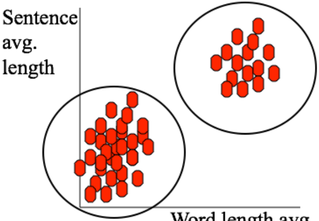


(Source)

Types of Machine Learning systems

Unsupervised learning: Clustering (II)

Example: Cluster word-sentence length in a books corpus



Clusters interpretation

- Long words and sentences: Philosophy?
- Short words and sentences: Novel?

Types of Machine Learning systems

Unsupervised learning: Clustering (III)

Example: Human resources department wants to know their employees profiles

Salary	Married	Car	Child.	Rent/owner	Syndicated	Leaves	Sen.	Sex
1000	Yes	No	0	Rent	No	7	15	M
2000	No	Yes	1	Rent	Yes	3	3	F
1500	Yes	Yes	2	Owner	Yes	5	10	M
3000	Yes	Yes	1	Rent	No	15	7	F
1000	Yes	Yes	0	Owner	Yes	1	6	M

Types of Machine Learning systems

Unsupervised learning: Clustering (IV)

	Group 1	Group 2	Group 3
Salary	1535	1428	1233
Married	77 %	98 %	0 %
Car	82 %	1 %	5 %
Child.	0.05	0.3	2.3
Rent/owner	99 %	75 %	17 %
Syndicated	80 %	0 %	67 %
Leaves	8.3	2.3	5.1
Seniority	8.7	8	8.1
Sex (M/F)	61 %	25 %	83 %

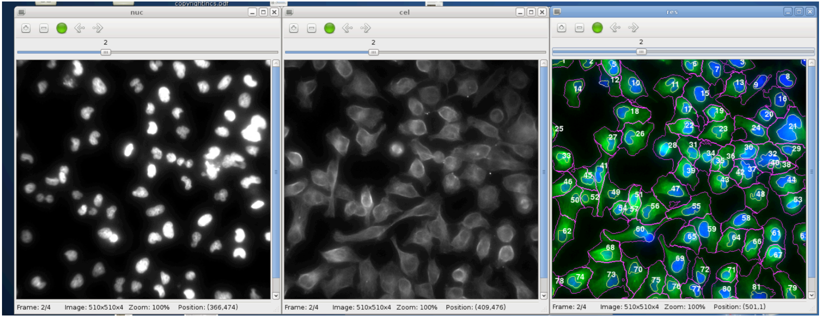
Analysis:

- Group 1: No children, with rented house. Low syndication. Many sick leaves.
- Group 2: No children, with car. High syndication. Low sick leaves. Usually women and rent.
- Group 3: With children, married, with car. Usually owners men. Low syndication.

Types of Machine Learning systems

Unsupervised learning: Clustering (V)

Example: Cells number count



Types of Machine Learning systems

Unsupervised learning: Association rules (I)

Association rules seek relations among attributes

f_1	f_2	f_3	$\cdots$	f_n
$a_{1,1}$	$a_{2,1}$	$a_{3,1}$	$\cdots$	$a_{n,1}$
$a_{1,2}$	$a_{2,2}$	$a_{3,2}$	$\cdots$	$a_{n,2}$
$a_{1,3}$	$a_{2,3}$	$a_{3,3}$	$\cdots$	$a_{n,3}$
$a_{1,4}$	$a_{2,4}$	$a_{3,4}$	$\cdots$	$a_{n,4}$
$a_{1,5}$	$a_{2,5}$	$a_{3,5}$	$\cdots$	$a_{n,5}$

Main association algorithms

- Apriori, Eclat, GP-growth

Algorithm output

- Rules
- Confidence: How often the rule is true
- Support: How often the rule applies

Types of Machine Learning systems

Unsupervised learning: Association rules (II)

Example: Market basket analysis

- A supermarket wants to gather information about its clients shopping behaviour

Objective

- Identify complementary items
- Enhance product placement

Id	Eggs	Oil	Diapers	Wine	Milk	Butter	Salmon	Lettuce	...
1	Yes	No	No	Yes	No	Yes	Yes	Yes	...
2	No	Yes	No	No	Yes	No	No	Yes	...
3	No	No	Yes	No	Yes	No	No	No	...
4	No	Yes	Yes	No	Yes	No	No	No	...
5	Yes	Yes	No	No	No	Yes	No	Yes	...
6	Yes	No	No	Yes	Yes	Yes	Yes	No	...
7	No	No	No	No	No	No	No	No	...
8	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	...
...	...	...	...	...	...	...	...	...	...

Types of Machine Learning systems

Unsupervised learning: Association rules (IV)

Association rules

```
if diapers=yes
then milk=yes (100%, 37%)

if eggs=yes
then oil=yes (50%, 25%)

if wine=yes
then lettuce=yes (33%, 12%)
```

where (confidence, support)

Types of Machine Learning systems

Unsupervised learning: Dimensionality reduction (I)

Dimensionality reduction transforms data into more convenient representations

- Reduce data dimensionality
- Visualize multidimensional data

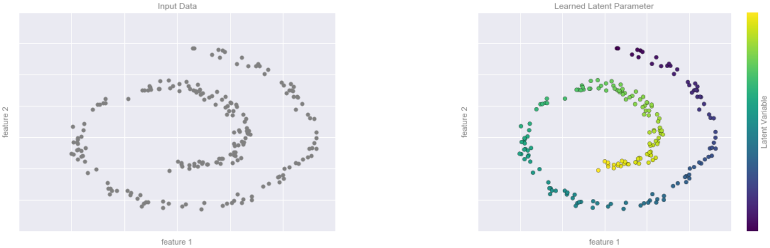
Main algorithms

- Isomap
- Principal Components Analysis (PCA)
- T-distributed Stochastic Neighbor Embedding (t-SNE)

Types of Machine Learning systems

Unsupervised learning: Dimensionality reduction (II)

Example: Isomap

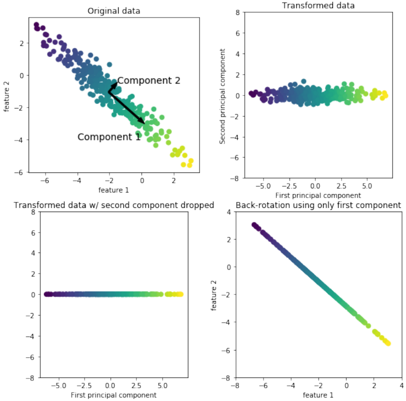


(Source)

Types of Machine Learning systems

Unsupervised learning: Dimensionality reduction (III)

Example: PCA



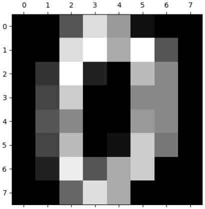
(Source)

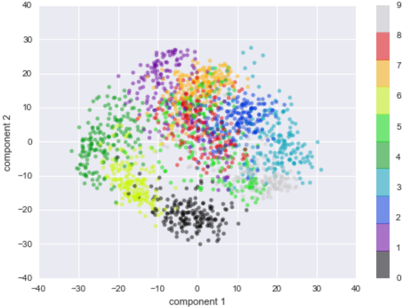
Types of Machine Learning systems

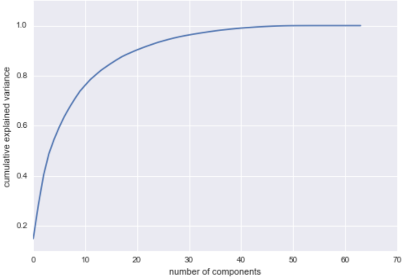
Unsupervised learning: Dimensionality reduction (IV)

Example: Hand-written digits recognition

- Images of hand-written digits
- 8x8 images (64 dimensions)
- 10 digits
- Classification problem





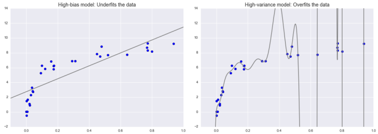


Main challenges of Machine Learning

Under and overfitting

Underfitting: Does not learn

- Topology too simple
- The model does not fit data
- Solution:
    - Increase model complexity



(Source)

Overfitting: Memorizes samples

- Topology too complex
- Very serious concern in ML
- The model does not generalize data
- Model fails when exposed to new data
- Solutions:
    - Reduce model complexity
    - Increase dataset
    - Apply regularization

Main challenges of Machine Learning

The curse of dimensionality

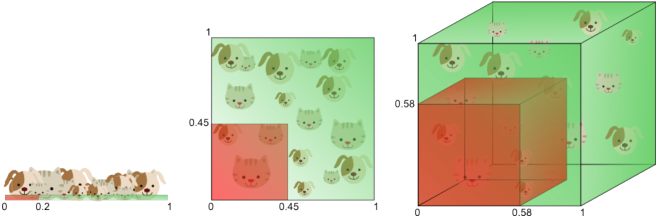
ML algorithms are statistical by nature

- Count frecuency of observations in regions

Fewer observations per region as dimensionality increases

- Data become sparser
- Need of more data to keep patterns
- Increased overfitting risk

Goal: Reduce dimensionality as much as possible

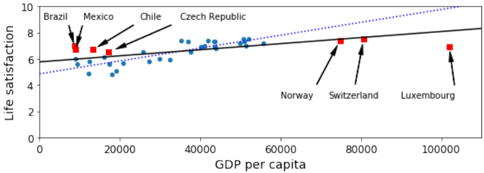


(Source)

Main challenges of Machine Learning

Other challenges

- Insufficient data
  - Given enough data, algorithms tend to similar performance
  - Remember: ML is data-centric
- Non representative training data
- Poor quality data
- Irrelevant features
- Unbalanced datasets



(Source)

Case studies

Case study 1: Bank propensity model

Client

- Bank

Business problem

- Identify those clients prone to buy a service

Data

- Available on several databases
- Historical data on service adquisition available

Propose a solution to:

- Data adquisition
- ML task
- Predictive or explicative model
- Model explotation
- Model maintenance

Case studies

Case study 2: Social media compaign impact

Client

- Car manufacturer

Business problem

- Real-time analysis of a campaign impact in Twitter
- Answer if people have a positive reaction to the campaign

Data

- None

Propose a solution to:

- Data adquisition
- ML task
- Predictive or explicative model
- Model explotation
- Model maintenance

Case studies

Case study 3: Hubble FGS-3 servo failure prediction

Client

- NASA

Business problem

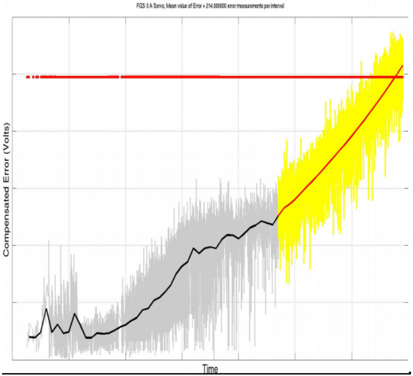
- Predict Hubble FGS-3 servo failure

Data

- Compensated error telemetry
- Servo will fail if compensated error exceeds a threshold

Propose a solution to:

- ML task
- Predictive or explicative model
- Model explotation
- Model maintenance



(Source of this study case)

Case studies

Case study 4: Fall detection with triaxial accelerometer

Client

- Technological start-up

Business problem

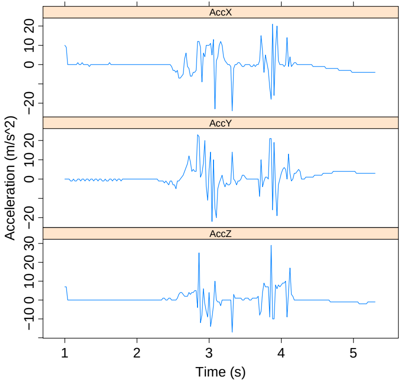
- Detect falls with a smartwatch
- Improve elderly people attention

Data

- None

Propose a solution to:

- Data adquisition
- ML task
- Data preprocessing
- Model explotation
- Model maintenance



(More info)

Case studies

Case study 5: Fall detection with sound

Client

- Technological start-up

Business problem

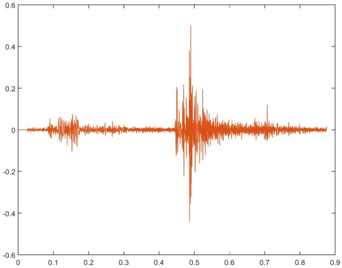
- Detect falls with sound
- Improve elderly people attention

Data

- None

Propose a solution to:

- Data adquisition
- ML task
- Data preprocessing
- Model explotation
- Model maintenance



Energy Mean	Energy Std
Number of Zeros Mean	Number of Zeros Std
Spectral Flux Mean	Spectral Flux Std
Roll off Factor Mean	Roll off Factor Std
Spectral centroid Mean	Spectral Centroid Std

(More info)

Case studies

Case study 6: NASA JPL BioSleeve

Client

- NASA JPL Advanced Robotics Group

Business problem

- Recognize hand gestures (more info)

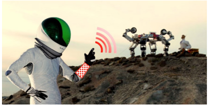
Data

- None

Propose a solution to:

- Data adquisition
- ML task

(Source)



(Source)

Wolf, Michael T., et al. *Decoding static and dynamic arm and hand gestures from the JPL BioSleeve.* IEEE Aerospace Conference. IEEE, 2013.

(Solution) (Results)

Case studies

Case study 7: UAV terrain classification

Client

- NASA JPL Advanced Robotics Group

Business problem

- Recognize terrain type for automatic UAV landing
- (Video)

Data

- UAV down-looking camera
- No dataset available

Propose a solution to:

- Data adquisition
- ML task
- Feature extraction