MACHINE LEARNING

Josiane.mothe@irit.fr

https://www.irit.fr/~Josiane.Mothe/welcome_a.html

Introduction

- If you use or reuse this document or part of it; you need to give the credit to
  « J. Mothe (2018) » or
  « J. Mothe (2018) Introduction to Machine Learning *in* FabSpace Bootcamp »
- Objectives:
  - What *machine learning* covers?
  - Basics in machine learning
- Plan
  - Introduction
  - Data representation for ML
  - ML approaches
  - Applications
  - Books and MOOCs

Your turn !

I. INTRODUCTION

Machine Learning

- What does *Machine Learning* cover ?
- What is it linked to?

Machine Learning

Machine Learning

- Supervised : Train / test
- Predictive analysis

Machine Learning

Data Mining

- Un-supervised
- Descriptive analysis

Machine Learning

- Supervised : Train / test
- Predictive analysis

Machine Learning

Data Mining

- Un-supervised
- Descriptive analysis

Machine Learning

- Supervised : Train / test
- Predictive analysis

- Interpretation
- Graphs
- Synthetic views

Visualisation

Machine Learning

- Data features
- Matrices

Data representation

Data Mining

- Un-supervised
- Descriptive analysis

Machine Learning

- Supervised : Train / test
- Predictive analysis

- Interpretation
- Graphs
- Synthetic views

Visualisation

Machine Learning

- Volume
- Variety
- Velocity
- Veracity

Big Data

- Data features
- Matrices

Data representation

Data Mining

- Un-supervised
- Descriptive analysis

Machine Learning

- Supervised : Train / test
- Predictive analysis

- Interpretation
- Graphs
- Synthetic views

Visualisation

Machine Learning

- Is part of a more general process

Data
Big data

Data representation

Data mining
Machine learning

Data visualisation
Interpretation
Results

Machine Learning

- Machine learning designs and studies **algorithms** that can **learn from data** and **make predictions on data**

Machine Learning

Training / Learning

- Annotated data : Examples for which we know the decision

SUN

BALLON

Testing / Predicting

- Non-annotated data : for which we want the decision

?

Machine Learning

Training / Learning

- Annotated data : Examples for which we know the decision

Ind	Weight	Height	Hair	Class
I1	20	1.12	Blond	Child
I2	75	1.80	Brown	Adult
I3	80	1.74	Brown	Adult
I4	18	0.80	Brown	Child

Testing / Predicting

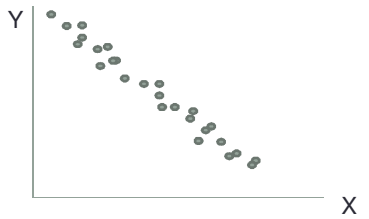
- Non-annotated data : for which we want the decision

Ind	Weight	Height	Hair	Class
I8	15	0.87	Blond	?

Machine Learning

Training / Learning

- Annotated data : Examples for which we know the decision



Testing / Predicting

- Non-annotated data : for which we want the decision

X=5
Y=?

II. DATA REPRESENTATION FOR MACHINE LEARNING

- Indexing
- Information Extraction

Data representation

- Quality
- Cleaning, sampling, completeness

- Individuals / Features - Variables
- Matrices or Vectors

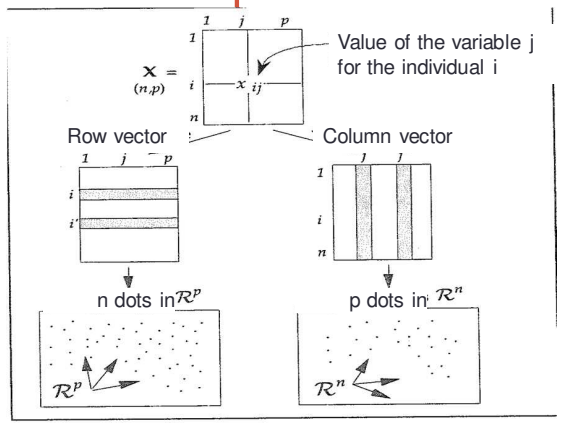
Data representation

- Depends on the data type
- Text – non structured
  - Indexing = term extraction
  - Vectors
- Images
  - Histograms of colors
  - CNN
- Structured data
  - Vectors: characteristics of each individuals
  - Matrices

Forms of data representation

- Vectors or matrices
  - For each text document the terms it is composed of
  - For each person her characteristics or feature values
  - For each individual the value of each variable

Variable & information representation



adapted from Lebart et al.

Types of variables / Features

- Quantitative Variable

Types of variables / Features

- Quantitative Variable

  Numerical feature

  Arithmetic's can be applied with meaningful results.
- Examples

  Age

  Temperature
- Continuous vs discrete
  - Discrete: variable with a finite number of values
  - Continuous: variable with an infinite number of values

Types of variables / Features

- Qualitative Variable
  Non-numerical feature
  Arithmetic's is not meaningful
- Examples
  Color of eyes
  Level of satisfaction

  - Nominal: categories with no order
  - Ordinal: categories can be ordered/ranked

Type of variables

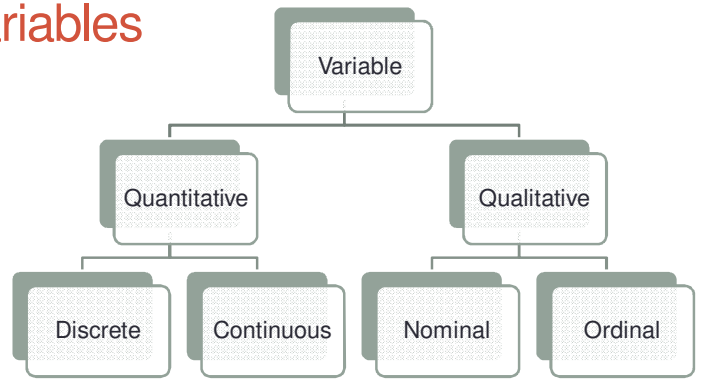
- Number of failures of a system
- Time needed to process a task
- Color of eyes
- Answer to the question: « do you think the system answers correctly »?

Type of variables

- Number of failures of a system
- Time needed to process a task
- Color of eyes
- Answer to the question: « do you think the system answers correctly »?

- Quantitative & discrete
- Quantitative & continuous
- Qualitative & nominal
- Qualitative & ordinal

Variables



Inspired from ebrunelle.ep.profweb.qc.ca/MQ/Chapitre2.pdf

III. MACHINE LEARNING APPROACHES

Machine learning approaches

Descriptive analysis

- Analysis of the data
  - No train/test
  - Extract the model from the data

Predictive analysis

- Analysis of the training set
  - Extract the model
- Use the model on the test or data for which we need a prediction

Machine learning approaches

Descriptive analysis

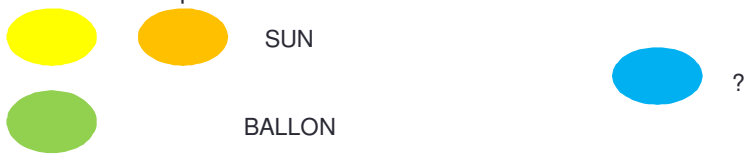
- Analysis of the data
  - No train/test
  - Extract the model from the data
- Once the model is extracted, it can be considered as a trained one and used for new data: the border is not that easy.

Predictive analysis

- Analysis of the training set
  - Extract the model
- Use the model on the test or data for which we need a prediction

Data quality

- Completeness of the training data set
  - Lack of examples



Data quality

- Completeness of the features
  - Lack of values in training or testing

Ind	Weight	Height	Hair	Class
I1		1.12	Blond	Child
I2	75	1.80	Brown	Adult
I3	80	1.74		Adult
I4	18		Brown	Child

Ind	Weight	Height	Hair	Class
I8		1.30	Blond	?

Over-fitting

SUN

BALLON

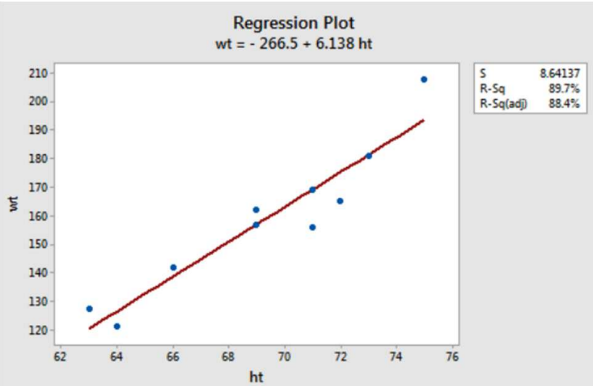
A few supervised methods

Simple/Multiple linear regression

- Simple: Relationships between two continuous (quantitative) variables
  - One (simple) variable is regarded as the **predictor**/explanatory/independant variable
  - The other variable is regarded as the **response**/outcome/dependent variable
- Extract the trend / relationship that may exist between the predictor and the response
- Linear

Simple/Multiple linear regression

- fitted using the least squares (minimizes *the sum of squared residuals)* approach (or other)



https://onlinecourses.science.psu.edu/stat501/node/252

Simple/Multiple linear regression

- More?
  - https://onlinecourses.science.psu.edu/stat501/
    - Simple linear regression
    - Multiple linear regression
    - Evaluation
  - http://www.r-tutor.com/elementary-statistics

Decision Tree

- The model predicts the value of a target (response) variable by learning simple decision rules inferred from the data features.
- Both categorical and continuous input/predictor and output/response variables

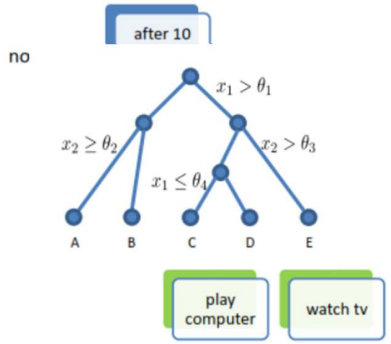


Illustration of Decision Tree

https://medium.com/@haydar_ai/learning-data-science-day-12-decision-tree-fa83499f89fd

Decision Trees

- CART (Classification And Regression Tree)
  - Binary tree
  - A greedy approach is used to divide the space : recursive binary splitting.
  - All the values are lined up and different split points are tried and tested using a cost function. The split with the lowest cost is selected
  - Regression predictive modeling: the cost function is the sum squared error across all training samples

https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/

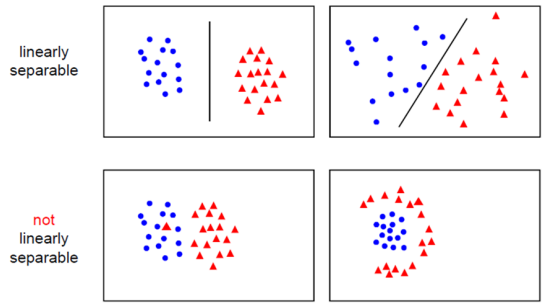
Decision Tree

- More?
- http://scikit-learn.org/stable/modules/tree.html
- https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning
- https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/

SVM

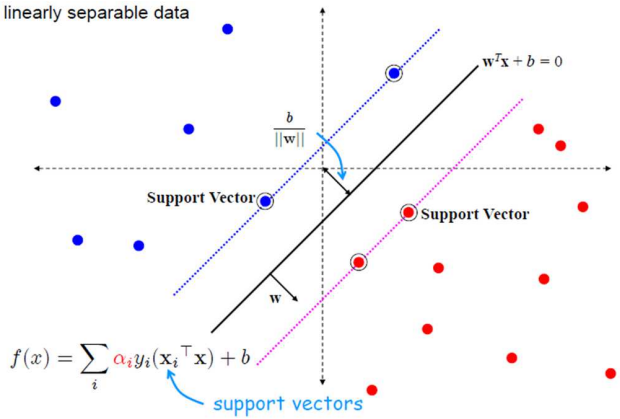
- Support Vector Machine

Binary classification
(multi-classes too)



www.robots.ox.ac.uk/~az/lectures/ml/lect2.pdf

SVM



www.robots.ox.ac.uk/~az/lectures/ml/lect2.pdf

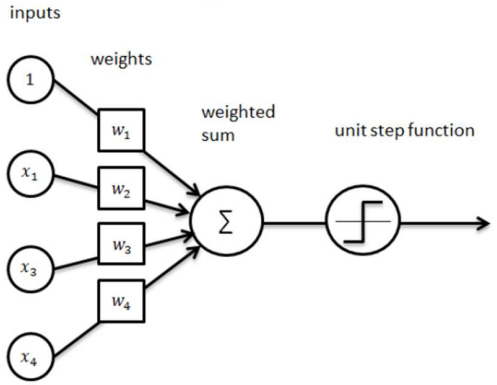
SVM

- More?
  - *www.robots.ox.ac.uk/~az/lectures/ml/lect2.pdf (slides)*
  - *cs229.stanford.edu/notes/cs229-notes3.pdf(textual)*
  - https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-034-artificial-intelligence-fall-2010/lecture-videos/lecture-16-learning-support-vector-machines/ (video)

From NN to Deep Learning

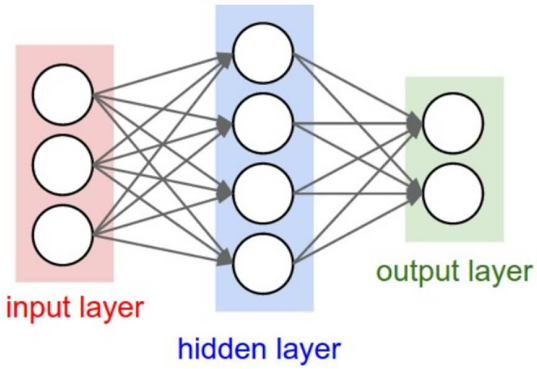
- Perceptron

- Unique neuron
- Learns the connection weights
- Using examples
  Input
  Expected output



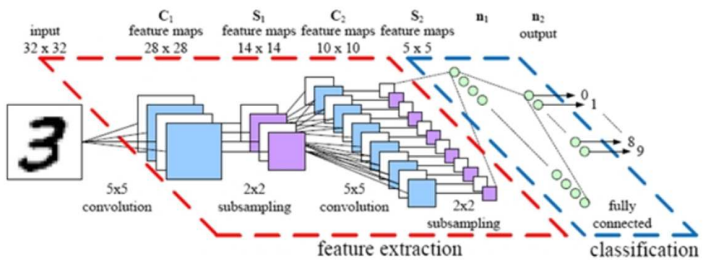
From NN to Deep Learning

- Multi layer NN



From NN to Deep Learning

- Convolutional NN



From NN to Deep Learning

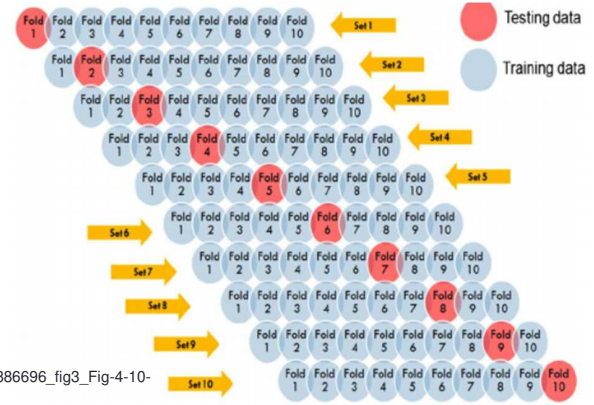
- Want to know more?
  - Udacity (Google) https://www.udacity.com/course/deep-learning--ud730
  - Coding http://course.fast.ai/
  - Tensorflow library: https://www.tensorflow.org/
  - Keras: https://elitedatascience.com/keras-tutorial-deep-learning-in-python

Evaluation in supervised methods

- Cross validation
  - A training set (to calculate the model)
  - A testing set (to check the results)
  - [A validation set]
- How to choose the training/testing sets?
  - 80 / 20

Evaluation in supervised methods

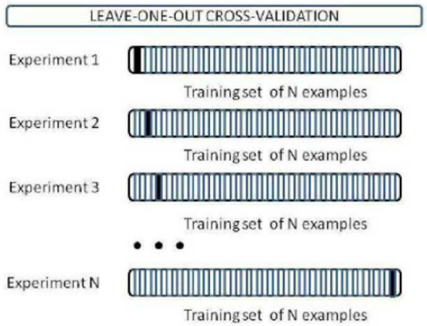
- 10-folds cross validation



https://www.researchgate.net/figure/239386696_fig3_Fig-4-10-fold-cross-validation-procedure

Evaluation in supervised methods

- Leave-one out



http://www.nosimpler.me/machine-learning/

A few non-supervised methods

Agglomerative clustering

Algorithm:

1 Step: Initialisation

- Initial classes = n individuals or objects
- Calculate the matrix of distances between each pairs of objects

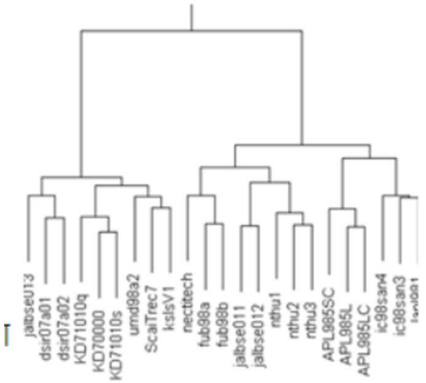
2 Step: Iteratively process the following:

- Group together the closest 2 elements (objects or groups) according to the agregation criteria
- Update the matrix of distance by replacing the two grouped elements by a new element and by calculating its distance with the other elements

End of the iteration: no element to group (a single class)

Agglomerative clustering

- Results in a hierarchy of partitions
- Dendrogram or tree that can be cut at different levels



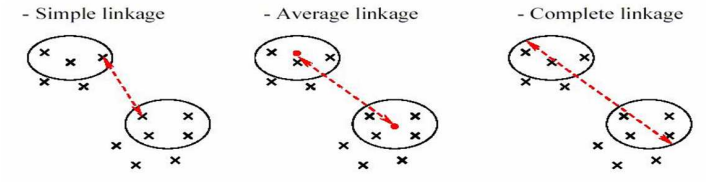
Aggomerative clustering

- Needs a distance

Nom	Paramètre	Fonction
distance de Manhattan	1-distance	$\sum_{i=1}^{n} \lvert x_i - y_i \rvert$
distance euclidienne	2-distance	$\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$
distance de Minkowski	p-distance	$\sqrt[p]{\sum_{i=1}^{n} \lvert x_i - y_i \rvert^p}$

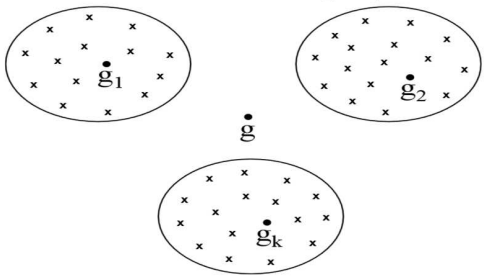
Agglomerative clustering

- Needs an aggregation criteria

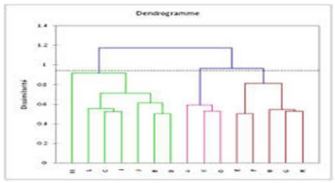


Agglomerative clustering

- Ward criteria
  - minimum of the intra-cluster inertia or
  - maximum of the inter-group inertia



Inter-group inertia: mean of the squared distances from the center of gravity



K-means and variants

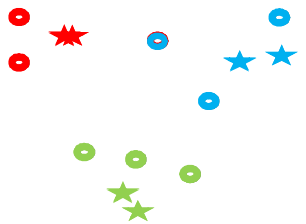
- Intitialisation:
  - Choose Q group centroïds
    Cluster the objects (closest centroid)
    ➔ First partition
  - Repeat
    Calculate the new centroïds considering the objects in the cluster
    Recluster the objects

Until either no changes, or a certain number of iterations, or …

K-means and variants

- Parameters - variants
  - Choice of the initial Q centroïds
  - Distance « *closest* »
  - New centroïds

Red becomes Bleu

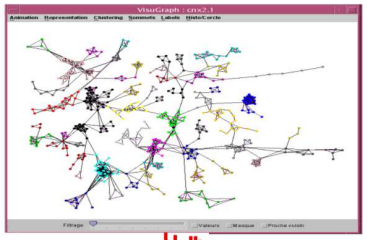


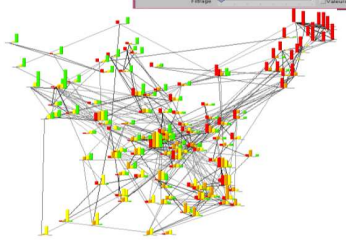
Others

- Factorial analysis
  - Principal Component Analysis
  - Correspondance Analysis
  - Search for the best axis in order to visualize a N dimensional space into a P dimensional space P << N
  - Based on singular value decomposition
  - Graphical visualization of the results

Others

- Graph analysis





IV. WANT TO KNOW MORE?

MOOCs
Books
Practical works

MOOCs

- Machine Learning : http://online.stanford.edu/course/machine-learning or https://www.coursera.org/learn/machine-learning

This course provides a broad introduction to machine learning, datamining, and statistical pattern recognition. Topics include: (i) Supervised learning (parametric/non-parametric algorithms, support vector machines, kernels, neural networks). (ii) Unsupervised learning (clustering, dimensionality reduction, recommender systems, deep learning). (iii) Best practices in machine learning (bias/variance theory; innovation process in machine learning and AI). The course will also draw from numerous case studies and applications, so that you'll also learn how to apply learning algorithms to building smart robots (perception, control), text understanding (web search, anti-spam), computer vision, medical informatics, audio, database mining, and other areas.

https://www.youtube.com/watch?v=UzxYlbK2c7E

Andrew Ng, Co-founder, Coursera; Adjunct Professor, Stanford University; formerly head of Baidu AI Group/Google Brain

MOOCs

- Machine Learning : https://lagunita.stanford.edu/courses/HumanitiesSciences/StatLearning/Winter2016/about

This is an introductory-level course in supervised learning, with a focus on regression and classification methods. The syllabus includes: linear and polynomial regression, logistic regression and linear discriminant analysis; cross-validation and the bootstrap, model selection and regularization methods (ridge and lasso); nonlinear models, splines and generalized additive models; tree-based methods, random forests and boosting; support-vector machines. Some unsupervised learning methods are discussed: principal components and clustering (k-means and hierarchical).

MOOCs

- Deep learning: https://www.udacity.com/course/deep-learning--ud730

We'll show you how to train and optimize basic neural networks, convolutional neural networks, and long short term memory networks. Complete learning systems in TensorFlow will be introduced via projects and assignments. You will learn to solve new classes of problems that were once thought prohibitively challenging, and come to better appreciate the complex nature of human intelligence as you solve these same problems effortlessly using deep learning methods.

Vincent Vanhoucke, Principal Scientist at Google

Practice ML with R

- https://www.datacamp.com/community/tutorials/machine-learning-in-r (Karlijn Willems, 2015)

This small tutorial is meant to introduce you to the basics of machine learning in R: it will show you how to use R to work with KNN

- https://machinelearningmastery.com/machine-learning-in-r-step-by-step/ (Jason Brownlee, 2016)

In this step-by-step tutorial you will:

1. Download and install R and get the most useful package for machine learning in R.
2. Load a dataset and understand it's structure using statistical summaries and data visualization.
3. Create 5 machine learning models, pick the best and build confidence that the accuracy is reliable.

Practice ML with Python

- https://machinelearningmastery.com/machine-learning-in-python-step-by-step/ (Jason Brownlee, 2016)

In this step-by-step tutorial you will:

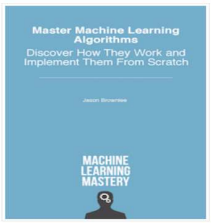
1. Download and install Python SciPy and get the most useful package for machine learning in Python.
2. Load a dataset and understand it's structure using statistical summaries and data visualization.
3. Create 6 machine learning models, pick the best and build confidence that the accuracy is reliable.

Books

Machine Learning Algorithms

#1 BEST SELLER

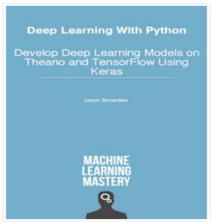
(ideal for Beginner Level)



Deep Learning With Python

TOP SELLER

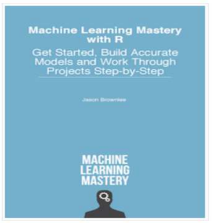
(ideal for Advanced Level)



Machine Learning With R

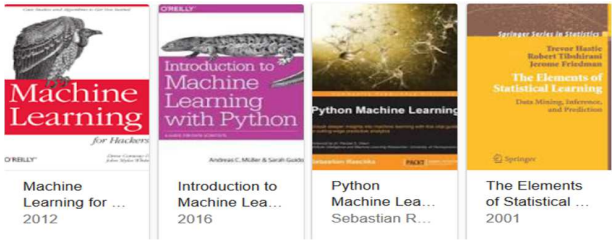
TOP SELLER

(ideal for Intermediate Level)



Books

https://github.com/josephmisiti/awesome-machine-learning/blob/master/books.md



Machine Learning for ... 2012

Introduction to Machine Lea... 2016

Python Machine Lea... Sebastian R...

The Elements of Statistical ... 2001