Neural
Academy

Master in Data Science - Lesson 2

Introduction to Machine Learning

What is Machine Learning ?

Neural Academy

What is Machine Learning?

Definition

“Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed”

Arthur L. Samuel, AI pioneer, 1959

What is Machine Learning?

Definition

“Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed”

Arthur L. Samuel, AI pioneer, 1959

Without ML

I have to explicitly encode computer behavior

With ML

Computers learn without direct involvement of programmer

What is Machine Learning?

What is a model?

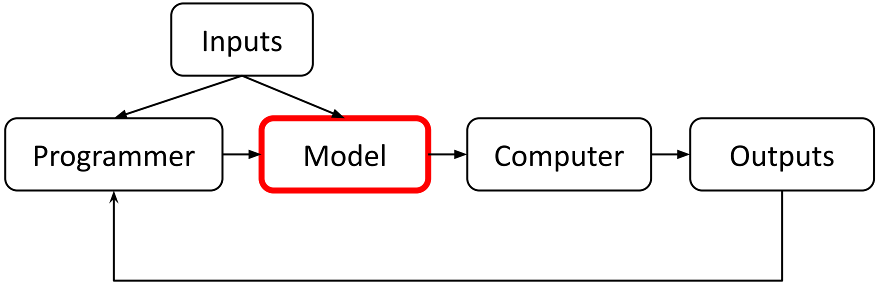
- In **physics** a *model*, it is a representation (simplified) of a system
  - It is much simpler and idealized than a real system
  - Eg. the shape of the Earth not actually a sphere, but we might treat it as one if we are designing a globe.

- In **machine learning**
  - A model is the output of a machine learning algorithm run on data.
  - definition by Tom Mitchell:
    - *"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."*

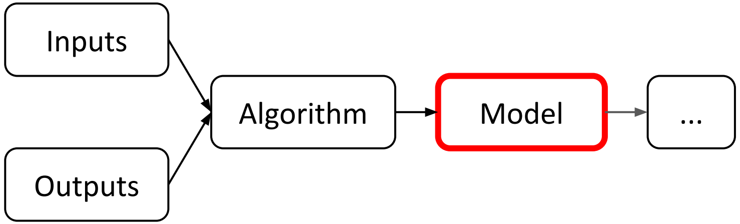
What is Machine Learning?

The paradigm shift

The Traditional Programming Paradigm

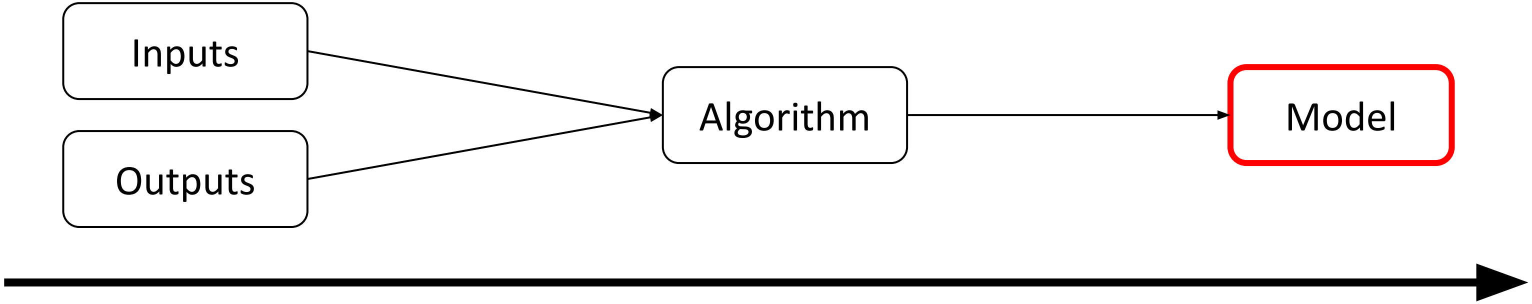


Machine Learning



What is Machine Learning?

The new paradigm



Inputs

Outputs

Algorithm

Model

Data

Gets better with more training examples and bigger computers

Algorithm

Statistics
Optimization
Computer science

Program

Given examples (*training data*), make a machine learn how to predict on new samples, or discover patterns in data.

What is Machine Learning?

Model for predictive tasks

- I want to predict *Event E*.
  - Ideally, I would like to measure it, but because I can’t measure it directly, the next best thing I can do is to predict it
  - Eg. As real estate agent I want to estimate (predict) if I sell this house (yes/no)

- I can observe **some** properties related to the Event E.
  - Eg. I can observe m^2, number of bedrooms, garage yes/no, garden yes/no, etc..
  - Eg. I can’t measure house foundation, client’s real budget, neighbour friendliness

What is Machine Learning?

Model for predictive tasks

x1	x2	x3	x4	x5	x6	x7	y
							0
							0
					X		0
							0
							1
							1
							1
							1

m^2, number of bedrooms, garage yes/no, garden yes/no

house foundation, client’s real budget, neighbour friendliness

1 = yes
0 = no

- I can’t directly measure **y**
- Using **x1**, **x2**, **x3**, **x4**, **x5**, **x6**, **x7** I would perfectly predict **y**
- I can observe (and measure) **x1**, **x2**, **x3**, **x4**
- I can’t observe **x5**, **x6**, **x7**
- I use **x1**, **x2**, **x3**, **x4** to predict **y** (with some errors)

What is Machine Learning?

Variation is information (monovariate)

x	y
3	0
3	0
3	0
3	0
3	1
3	1
3	1
3	1

x gives no information to predict **y**

x	y
3	0
3	0
5	0
5	0
5	1
5	1
5	1
5	1

x gives some information to predict **y**

x	y
3	0
3	0
3	0
3	0
5	1
5	1
5	1
5	1

x gives perfect information to predict **y**

What is Machine Learning?

Variation is information (multivariate)

x1	x2	y
5	3	0
5	3	0
5	3	0
5	3	0
5	3	1
5	3	1
5	3	1
5	3	1

x1 and x2 give no information to predict **y**

x1	x2	y
2	1	0
2	3	0
8	6	0
8	6	0
8	6	1
2	4	1
2	5	1
2	5	1

x1 and x2 give some information to predict **y**

x1	x2	y
2	1	0
2	3	0
8	7	0
8	6	0
2	6	1
2	4	1
2	5	1
2	9	1

x1 and x2 give some information to predict **y**
(x1 = 2 & x2 > 3 ⟶ 1, else 0)

What is Machine Learning?

Structured vs unstructured data

Structured Data

Set of numbers and labels organized in a tabular format

Databases
Spreadsheets

Semi-structured Data

Loosely organization with categories and meta tags

Emails by Sent / Inbox
Tweets and hashtags
Folders by topic

Unstructured Data

Unorganized data

Media posts, Emails
Video, Images
Speech, Sounds

Machine Learning Categories

Neural Academy

The 3 main Machine Learning categories

Supervised Learning

Labeled data
Direct feedback
Predict outcome
Predict future

Unsupervised Learning

No labels
No feedbacks
Find hidden structures

Reinforcement Learning

Decision process
Reward system
Learn series of action

The 3 main Machine Learning categories

Supervised Learning

Supervised Learning

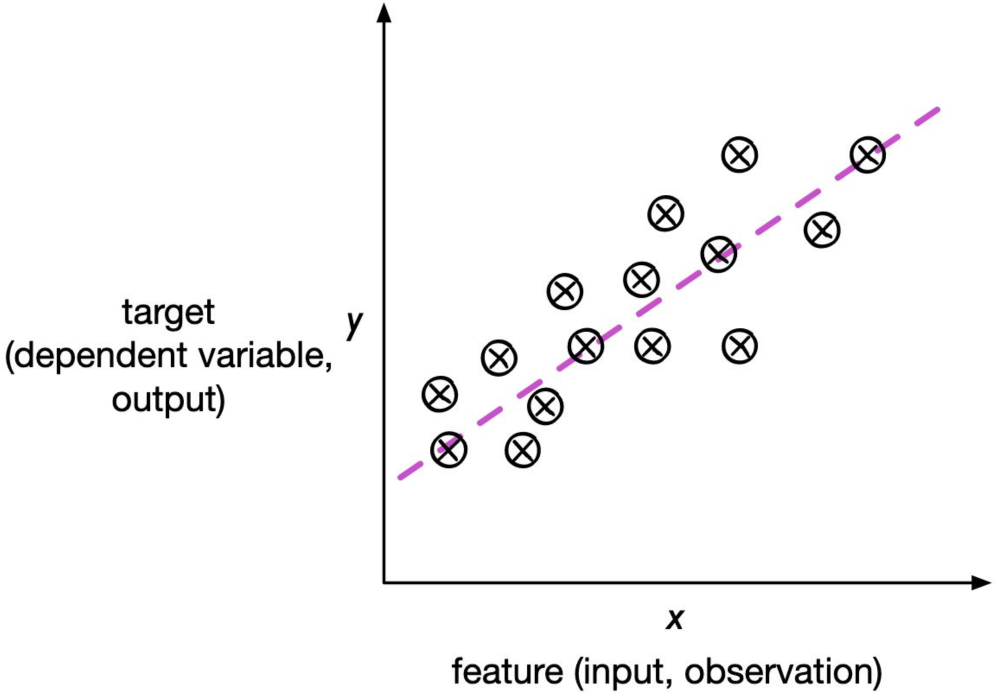
Labeled data
Direct feedback
Predict outcome
Predict future

The 3 main Machine Learning categories > Supervised Learning

Regression

Numerical values

x	y
2	5
4	6
7	10
3	5
..	..



Source: Raschka and Mirjalily (2019). *Python Machine Learning, 3rd Edition*

The 3 main Machine Learning categories > Supervised Learning

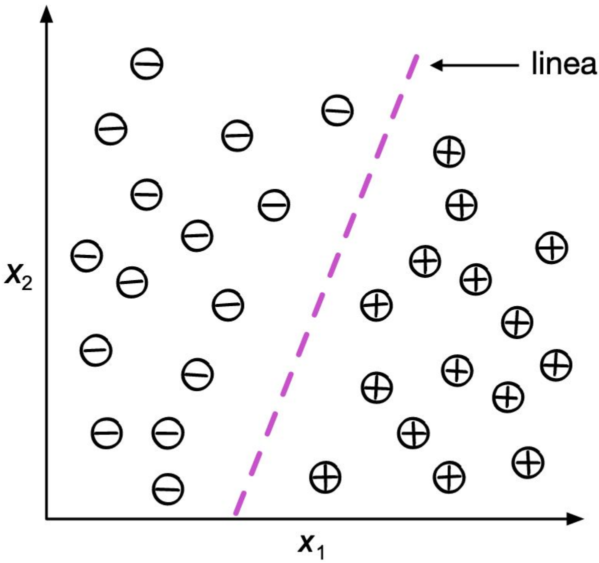
Classification

Categorical values

x1	x2	y
2	5	-
4	10	-
7	-2	+
3	8	+
..	..	..

Binary classification example with two *features* ("independent" variables, predictors)

What are the class labels (y's)?



linear decision boundary

Source: Raschka and Mirjalily (2019). *Python Machine Learning, 3rd Edition*

The 3 main Machine Learning categories > Supervised Learning

Ranking

- Correct order matter

HARRISON FORD
BLADE RUNNER
THE FINAL CUT

>

IV
STAR WARS
A NEW HOPE
The classic adventure that started it all
Formerly titled Star Wars: From the Adventures of Luke Skywalker
George Lucas

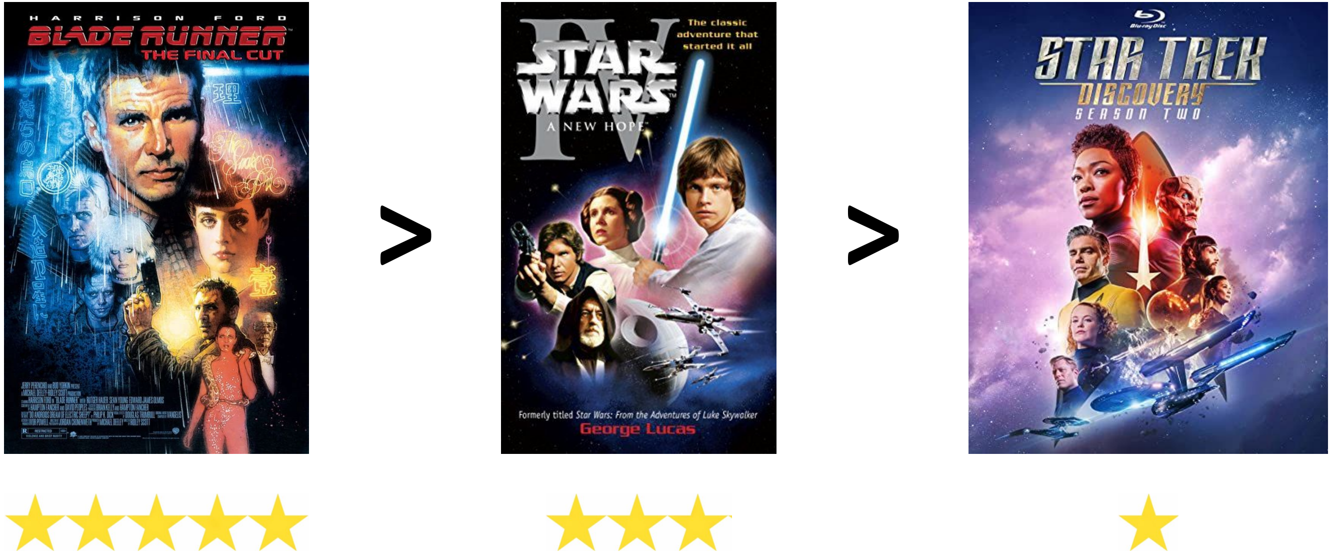
>

Blu-ray Disc
STAR TREK
DISCOVERY
SEASON TWO

The 3 main Machine Learning categories > Supervised Learning

Ordinal Regression

- Correct label matter
  - Eg. Modeling of human levels of preference (from "very poor" to "excellent")



The 3 main Machine Learning categories

Unsupervised Learning

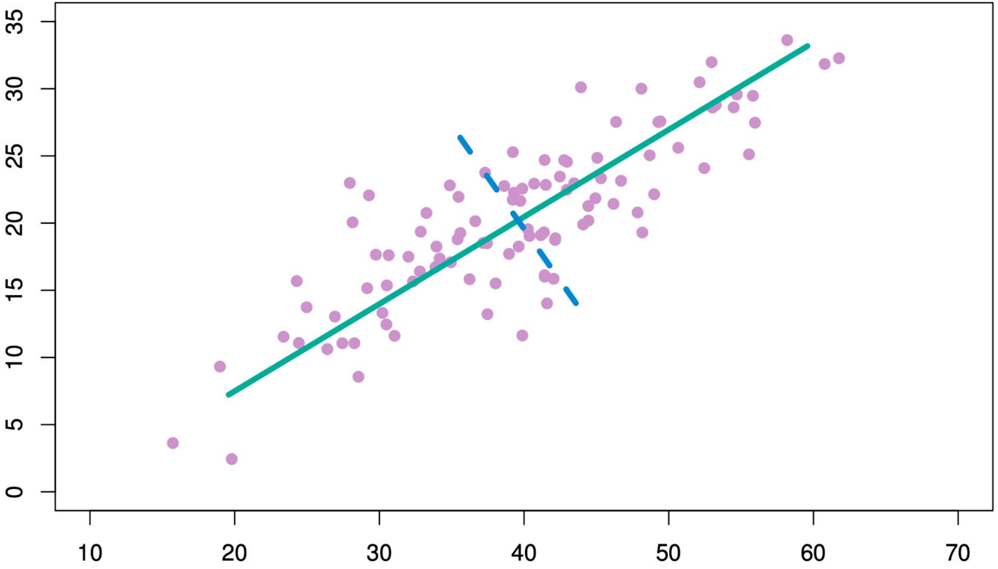
Unsupervised Learning

No labels
No feedbacks
Find hidden structures

Dimensionality Reduction - PCA

Principal Component Analysis (PCA)

x1	x2
2	5
4	10
7	-2
3	8
..	..

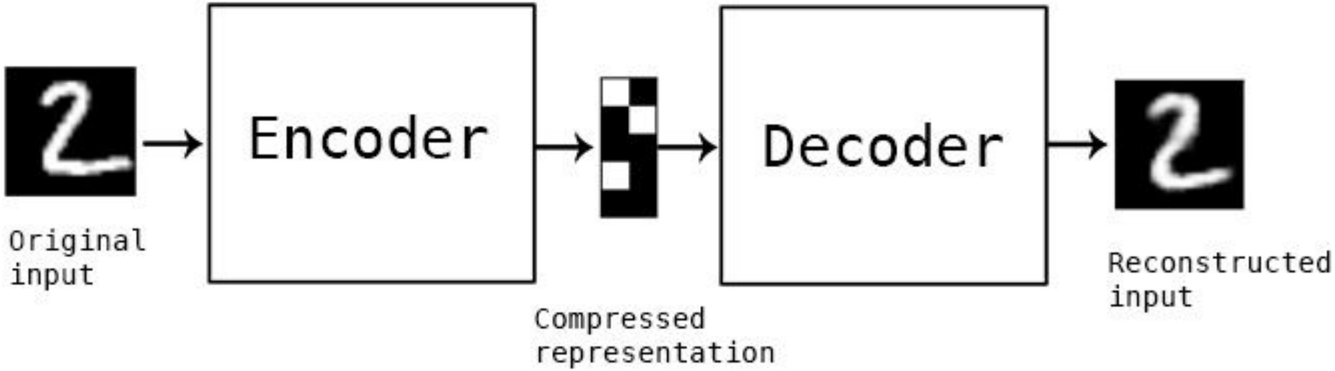


- Maximize the variance of projection along each component

- Minimize the reconstruction error (of the original data)

The 3 main Machine Learning categories > Unsupervised Learning

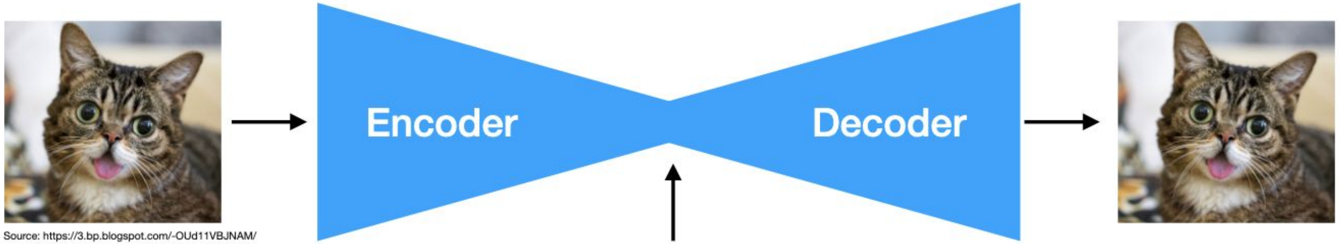
Dimensionality Reduction - Autoencoders



https://blog.keras.io/building-autoencoders-in-keras.html

The 3 main Machine Learning categories > Unsupervised Learning

Dimensionality Reduction - Autoencoders



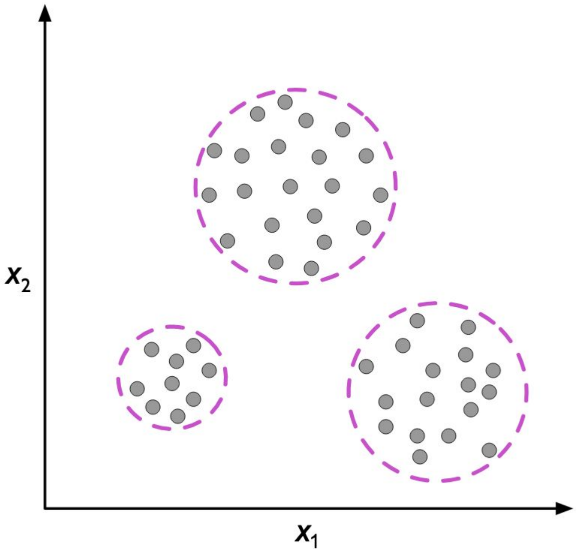
Source: https://3.bp.blogspot.com/-OUd11VBJNAM/VsFacR_YhBI/AAAAAAAABh0/ZKfKAnRj3x0/s1600/cannot%2Bresist.jpg

latent representation/
feature embedding

The 3 main Machine Learning categories > Unsupervised Learning

Clustering

Assigning group memberships to unlabelled examples (instances, data points)



Source: Raschka and Mirjalily (2019). *Python Machine Learning, 3rd Edition*

The 3 main Machine Learning categories

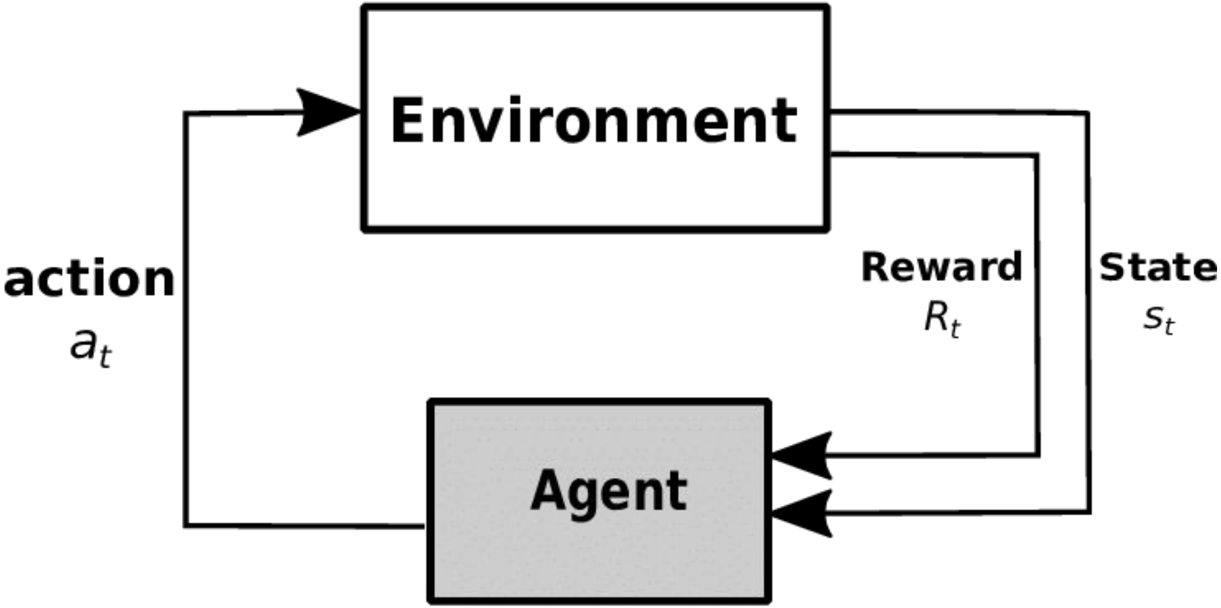
Reinforcement Learning

Reinforcement Learning

Decision process
Reward system
Learn series of actions

Reinforcement Learning

Overview



Reinforcement Learning

Lane following task

https://neptune.ai/blog/reinforcement-learning-applications

Terminology & Notation

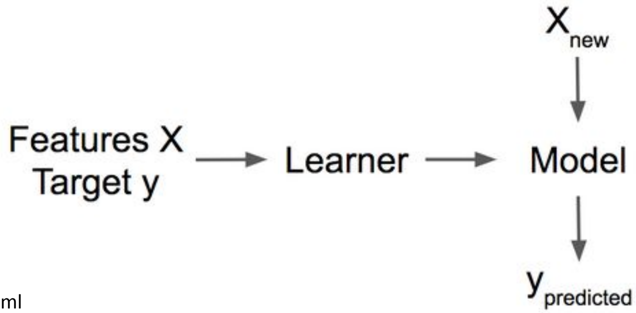
Neural Academy

Terminology & Notation

Learner & Model

- A **Learner** or **Machine Learning Algorithm** is the program used to learn a machine learning model from data. Another name is "inducer" (e.g. "tree inducer").

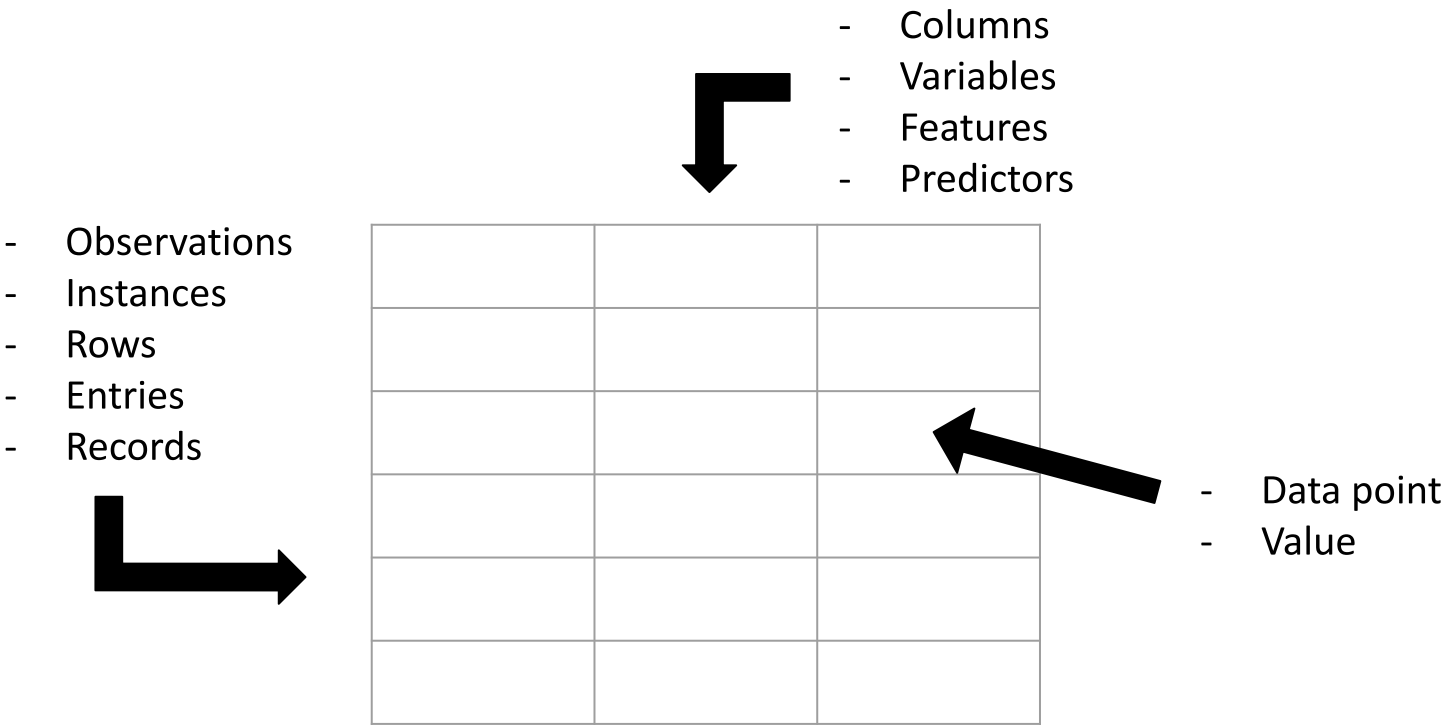
- A **Machine Learning Model** is the learned program that maps inputs to predictions. This can be a set of weights for a linear model or for a neural network. Other names for the rather unspecific word "model" are "predictor" or - depending on the task - "classifier" or "regression model". In formulas, the trained machine learning model is called f^ or f^(x).



https://christophm.github.io/interpretable-ml-book/terminology.html

Terminology & Notation

Structured data



- Columns
- Variables
- Features
- Predictors

- Observations
- Instances
- Rows
- Entries
- Records

- Data point
- Value

Terminology & Notation

Terminology

- **Supervised learning**: learn function to map input x (features) to output y (targets)
- **Structured data**: databases, spreadsheets/csv files
- **Unstructured data**: features like image pixels, audio signals, text sentences
- **Training example**, synonymous to observation, training record, training instance, training sample (in some contexts, sample refers to a collection of training examples)
- **Feature**, synonymous to predictor, variable, independent variable, input, attribute, covariate
- **Target**, synonymous to outcome, ground truth, output, response variable, dependent variable, (class) label (in classification)
- **Output /Prediction**, use this to distinguish from targets; here, means output from the model

Terminology & Notation

Supervised learning notation

"training examples"

Training set: $\mathscr{D} = \{\langle \mathbf{x}^{[i]}, y^{[i]} \rangle, i = 1, \dots , n\},$

Unknown function: $f(\mathbf{x}) = y$

Hypothesis: $h(\mathbf{x}) = \hat{y}$

sometimes t or o

Classification

Regression

$$h : \mathbb{R}^m \rightarrow \mathscr{Y}, \quad \mathscr{Y} = \{1,\dots,k\}$$

$$h : \mathbb{R}^m \rightarrow \mathbb{R}$$

Terminology & Notation

Data representation

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

Feature vector

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}$$

Design Matrix

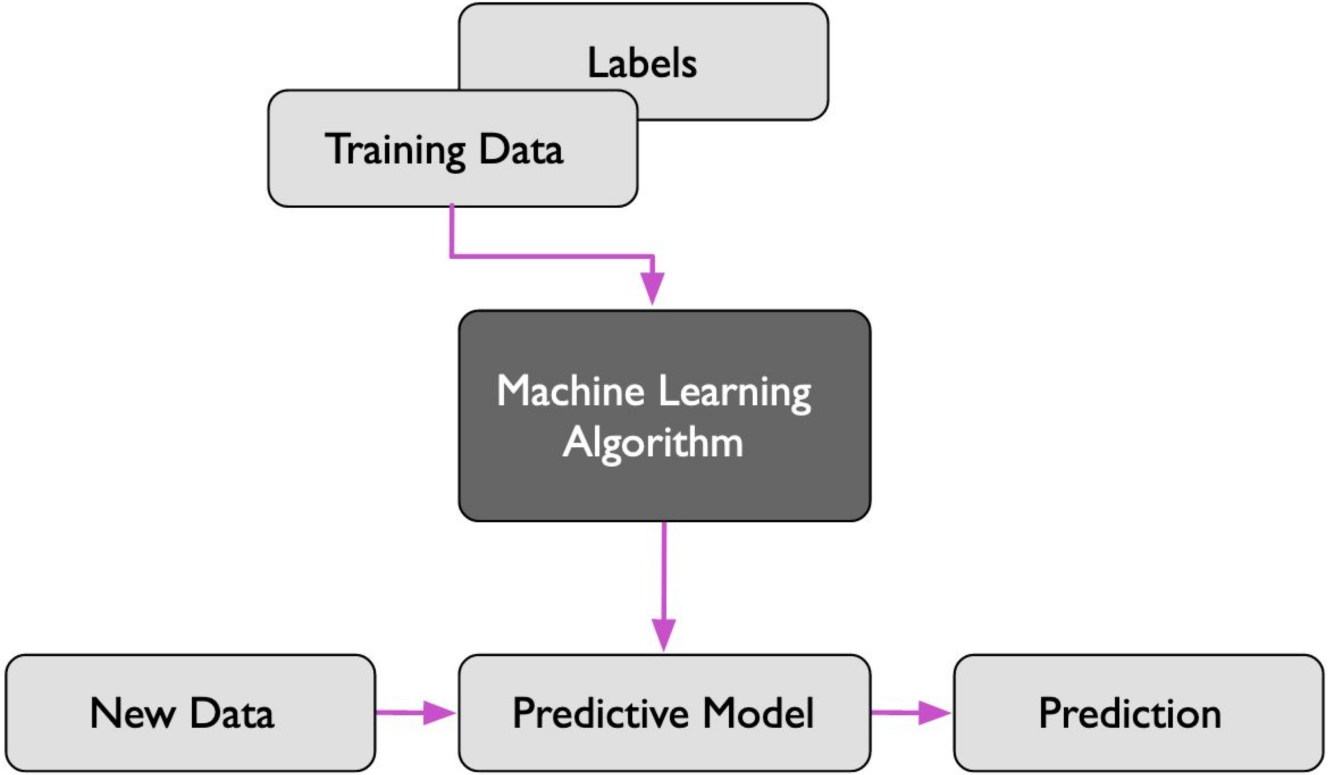
$$\mathbf{X} = \begin{bmatrix} x_1^{[1]} & x_2^{[1]} & \cdots & x_m^{[1]} \\ x_1^{[2]} & x_2^{[2]} & \cdots & x_m^{[2]} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{[n]} & x_2^{[n]} & \cdots & x_m^{[n]} \end{bmatrix}$$

Design Matrix

Machine Learning Pipeline

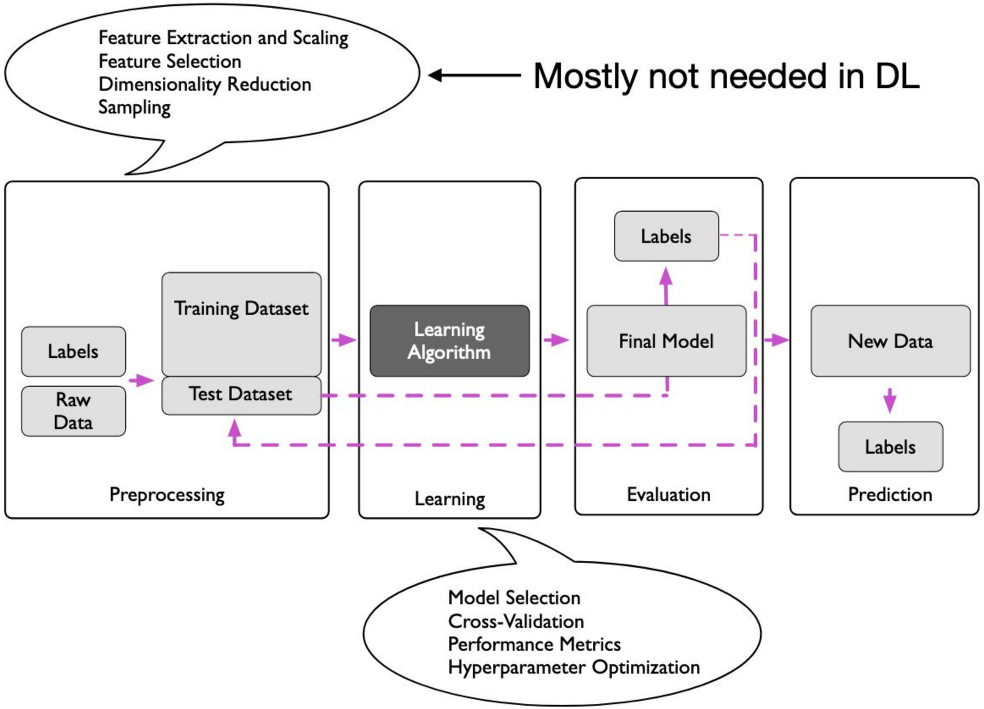
Machine Learning Pipeline

Supervised Learning Workflow



Machine Learning Pipeline

Supervised Learning Workflow (more details)



Feature Extraction and Scaling
Feature Selection
Dimensionality Reduction
Sampling

Mostly not needed in DL

Labels

Raw Data

Training Dataset

Test Dataset

Preprocessing

Learning Algorithm

Learning

Labels

Final Model

Evaluation

New Data

Labels

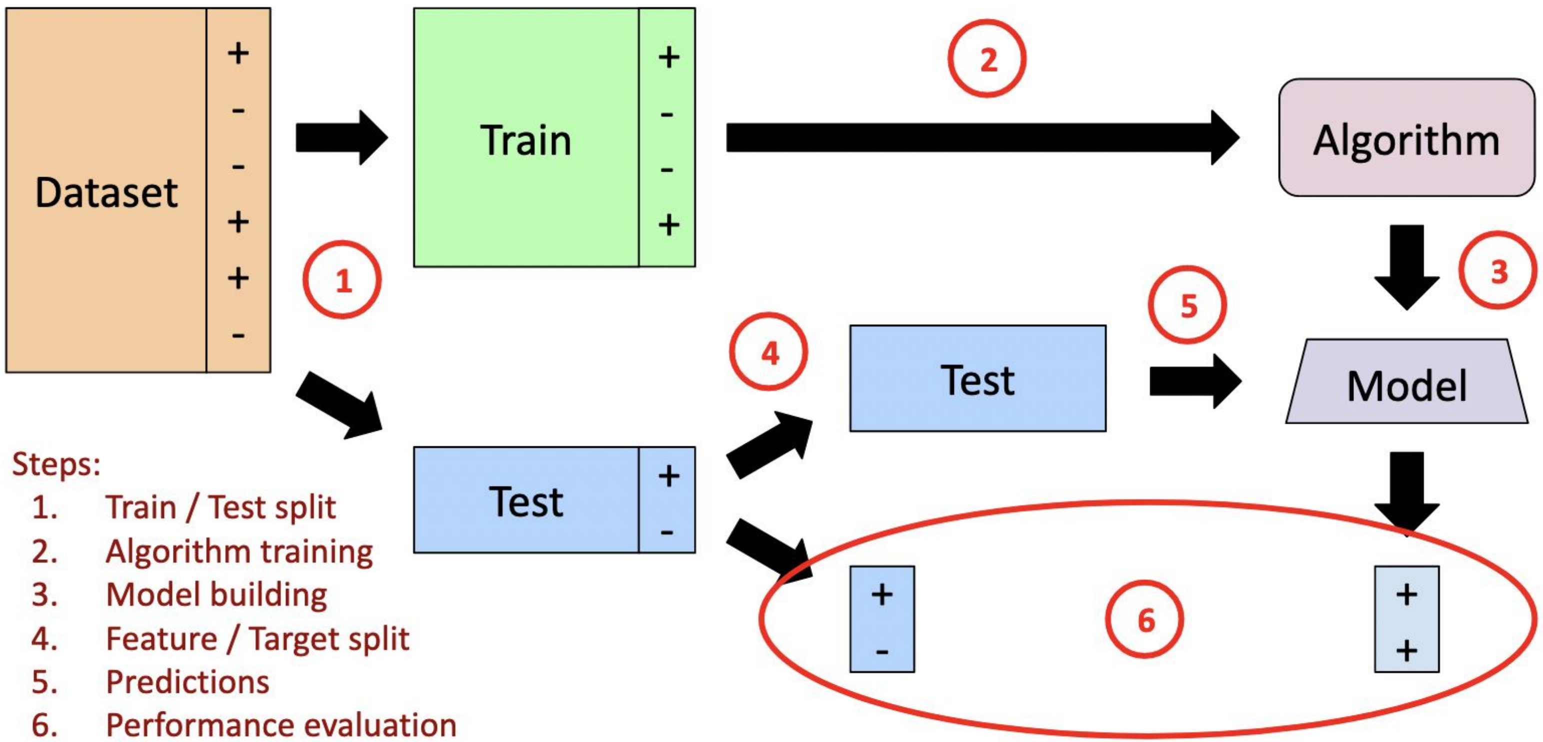
Prediction

Model Selection
Cross-Validation
Performance Metrics
Hyperparameter Optimization

Source: Raschka and Mirjalily (2019). *Python Machine Learning, 3rd Edition*

Machine Learning Pipeline

Training and Test dataset

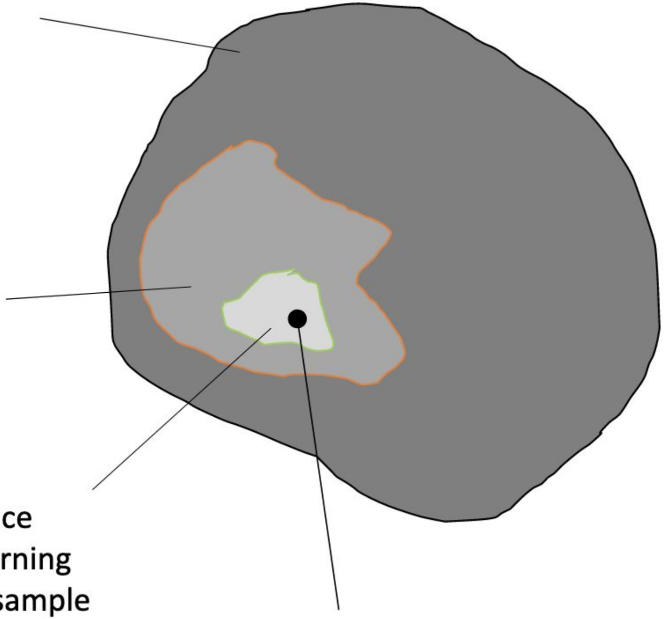


Steps:

1. Train / Test split
2. Algorithm training
3. Model building
4. Feature / Target split
5. Predictions
6. Performance evaluation

Machine Learning Pipeline

Hypothesis space



Entire hypothesis space

Hypothesis space
a particular learning
algorithm category
has access to

Hypothesis space
a particular learning
algorithm can sample

Particular hypothesis
(i.e., a model/classifier)

Machine Learning Pipeline

5 steps to address Machine Learning problems

1. Define problem to solve
2. Get data
3. Choose machine learning algorithm
4. Choose optimization metric for the model
    - Cost function / Loss function
5. Choose evaluation metric(s)

Machine Learning Pipeline

Learning

Learning = Representation + Evaluation + Optimization

(Pedro Domingos, A Few Useful Things to Know about Machine Learning

https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf)

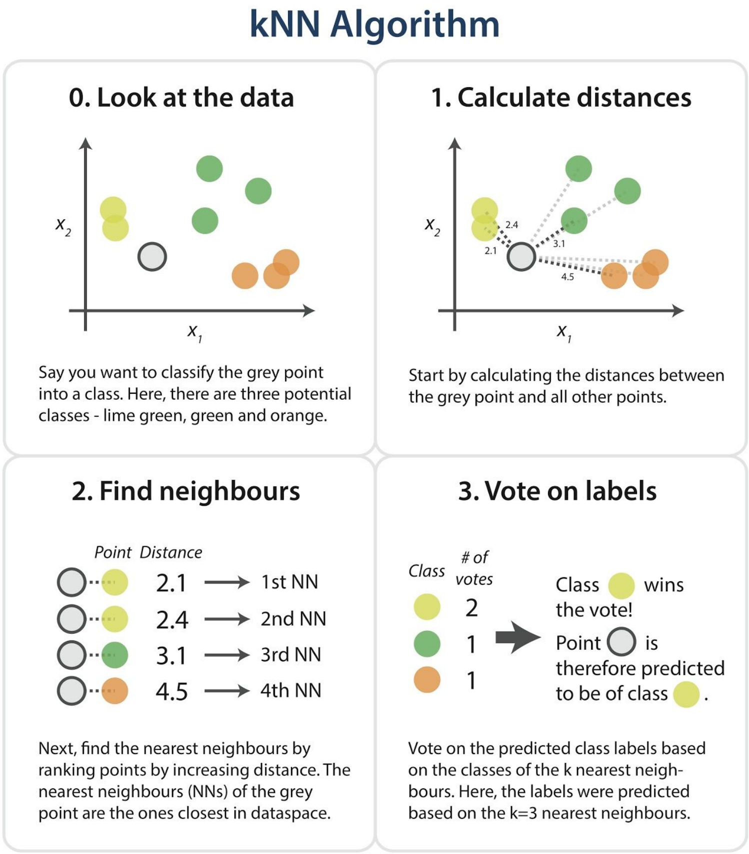
Machine Learning Pipeline

The first model: kNN

- **k-nearest neighbour (kNN)** classifies new instances by grouping them together with the most similar cases

- kNN is a type of supervised machine learning (though somewhat confusingly, in kNN there is no explicit training phase; see lazy learning)

- The kNN task can be broken down into 3 main tasks (see image)



kNN Algorithm

0. Look at the data

Say you want to classify the grey point into a class. Here, there are three potential classes - lime green, green and orange.

1. Calculate distances

Start by calculating the distances between the grey point and all other points.

2. Find neighbours

Point	*Distance*	
◯··🟡	2.1	→ 1st NN
◯··🟡	2.4	→ 2nd NN
◯··🟢	3.1	→ 3rd NN
◯··🟠	4.5	→ 4th NN

Next, find the nearest neighbours by ranking points by increasing distance. The nearest neighbours (NNs) of the grey point are the ones closest in dataspace.

3. Vote on labels

Class	*# of votes*
🟡	2
🟢	1
🟠	1

Class 🟡 wins the vote!

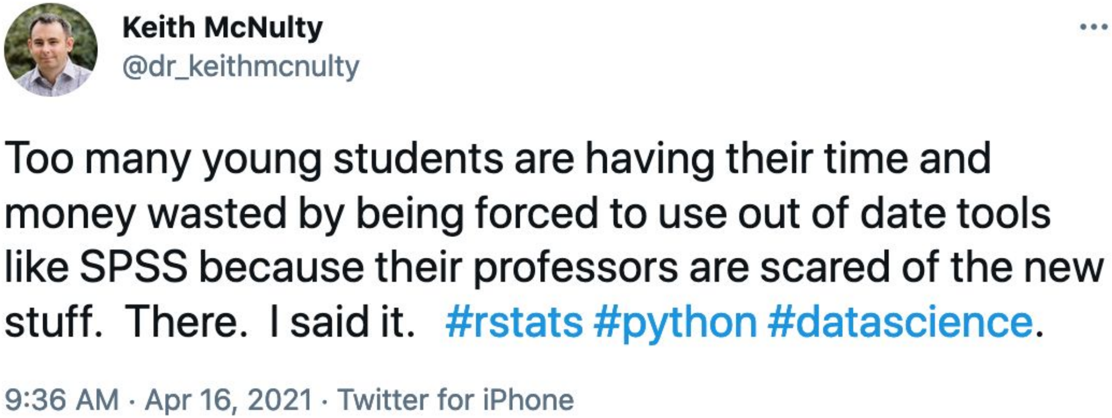
Point ◯ is therefore predicted to be of class 🟡.

Vote on the predicted class labels based on the classes of the k nearest neighbours. Here, the labels were predicted based on the k=3 nearest neighbours.

Conclusion

Conclusion

ML is a field in constant evolution



Keith McNulty
@dr_keithmcnulty

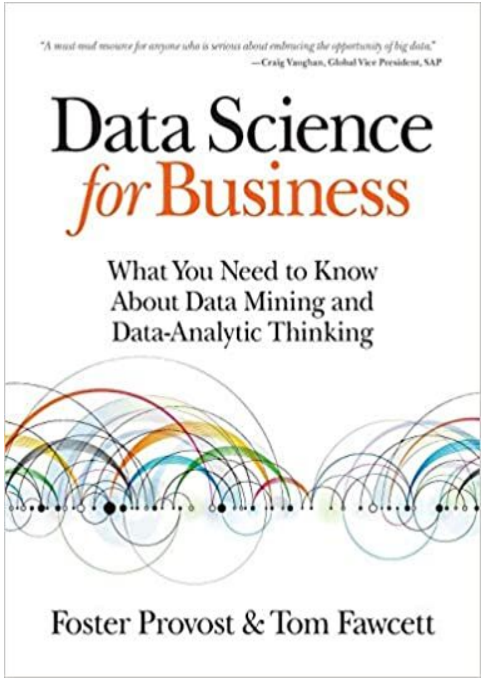
Too many young students are having their time and money wasted by being forced to use out of date tools like SPSS because their professors are scared of the new stuff. There. I said it.   #rstats #python #datascience.

9:36 AM · Apr 16, 2021 · Twitter for iPhone

https://twitter.com/dr_keithmcnulty/status/1382870396408627202

Conclusion

Suggested readings

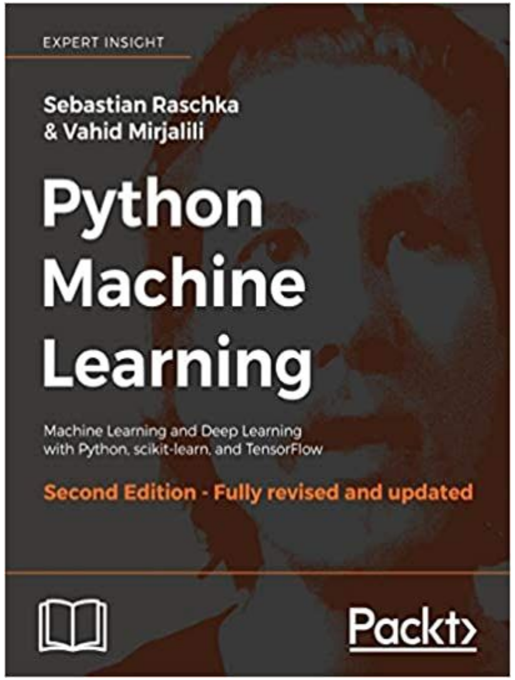


"A must read resource for anyone who is serious about embracing the opportunity of big data."
—Craig Vaughan, Global Vice President, SAP

Data Science *for* Business

What You Need to Know About Data Mining and Data-Analytic Thinking

Foster Provost & Tom Fawcett



EXPERT INSIGHT

Sebastian Raschka & Vahid Mirjalili

Python Machine Learning

Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow

Second Edition - Fully revised and updated

Packt>

review articles

DOI:10.1145/2347736.2347755

Tapping into the "folk knowledge" needed to advance machine learning applications.

BY PEDRO DOMINGOS

A Few Useful Things to Know About Machine Learning

MACHINE LEARNING SYSTEMS automatically learn programs from data. This is often a very attractive alternative to manually constructing them, and in the last decade the use of machine learning has spread rapidly throughout computer science and beyond. Machine learning is used in Web search, spam filters, recommender systems, ad placement, credit scoring, fraud detection, stock trading, drug design, and many other applications. A recent report from the McKinsey Global Institute asserts that machine learning (a.k.a. data mining or predictive analytics) will be the driver of the next big wave of innovation.[15] Several fine textbooks are available to interested practitioners and researchers (for example, Mitchell[16] and Witten et al.[24]). However, much of the "folk knowledge" that is needed to successfully develop machine learning applications is not readily available in them. As a result, many machine learning projects take much longer than necessary or wind up producing less-than-ideal results. Yet much of this folk knowledge is fairly easy to communicate. This is the purpose of this article.

» key insights

- Machine learning algorithms can figure out how to perform important tasks by generalizing from examples. This is often feasible and cost-effective where manual programming is not. As more data becomes available, more ambitious problems can be tackled.
- Machine learning is widely used in computer science and other fields. However, developing successful machine learning applications requires a substantial amount of "black art" that is difficult to find in textbooks.
- This article summarizes 12 key lessons that machine learning researchers and practitioners have learned. These include pitfalls to avoid, important issues to focus on, and answers to common questions.

78 COMMUNICATIONS OF THE ACM | OCTOBER 2012 | VOL. 55 | NO. 10

Conclusion

Suggested videos

Ted talk

Conclusion

Sources

- Sebastian Raschka - STAT 453: Intro to Deep Learning - 2020 slides
- Implementing your own k-nearest neighbour algorithm using Python
- The missing semester https://missing.csail.mit.edu/2020/
- https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html#sklearn.neighbors.KNeighborsClassifier

Buono Studio!

Contatti:

claudio.reggiani@neural.academy

agata.manco@neuralacademy.it
carlotta.reggioli@neuralacademy.it