Applied statistics and Machine learning in Python with subsurface applications

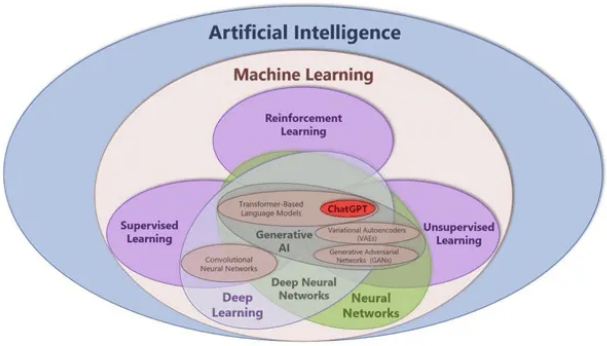
Enrico Riccardi, University of Stavanger

Feb 29, 2024

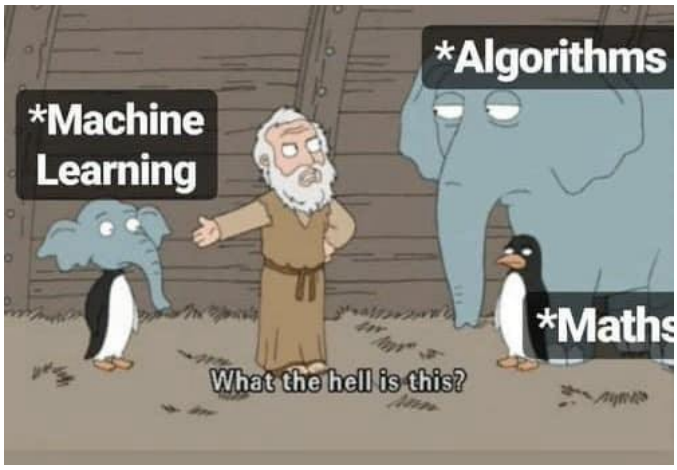
University of Stavanger

Statistics, Machine learning or Artificial intelligence?

What is the main difference between the three fields?



How Machine Learning Started?



Statistics

Let's start from the definition

- Statistics (origin "description of a state/country") is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.
- It is conventional to begin with a statistical population or a statistical model to be studied. Populations can be diverse groups of people or objects such as "all people living in a country" or "every atom composing a crystal".
- Statistics deals with every aspect of data, including the planning of data collection in terms of the design of surveys and experiments.[Wikipedia]

Machine learning

Definitions:

- Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. [IBM]
- Machine learning (ML) is a field of study in artificial intelligence concerned with the development and study of statistical algorithms that can learn from data and generalize to unseen data, and thus perform tasks without explicit instructions. [WIKI]
- Machine learning is a subfield of artificial intelligence that uses algorithms trained on data sets to create models that enable machines to perform tasks that would otherwise only be possible for humans, such as categorizing images, analyzing data, or predicting price fluctuations. [Coursera]

Machine learning

One technical definition

Machine learning is a set of computer based statistical approaches that aim to minimise the loss function to maximise inference accuracy. [Enrico, 5.2.2024]

The loss function is the actual engine in machine learning.

Loss function

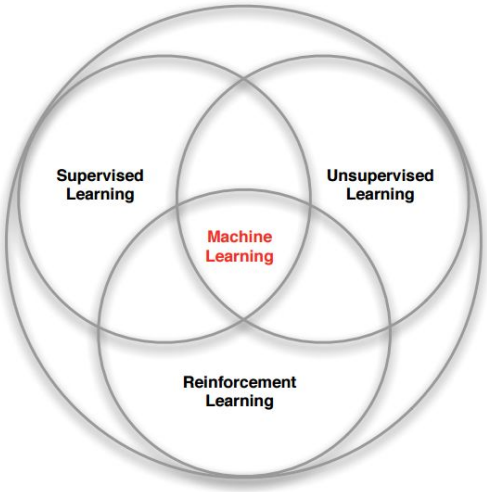
It quantifies the difference between the predicted outputs of a machine learning algorithm and the actual target values.

Artificial intelligences

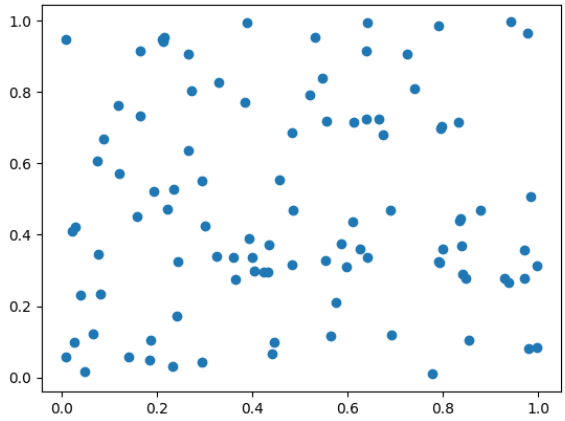
And more definitions:

- Artificial intelligence is the intelligence of machines or software, as opposed to the intelligence of humans or other animals. It is a field of study in computer science that develops and studies intelligent machines. [WIKI]
- Artificial intelligence (AI) is the theory and development of computer systems capable of performing tasks that historically required human intelligence, such as recognizing speech, making decisions, and identifying patterns [Coursera]
- It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable. [IBM]

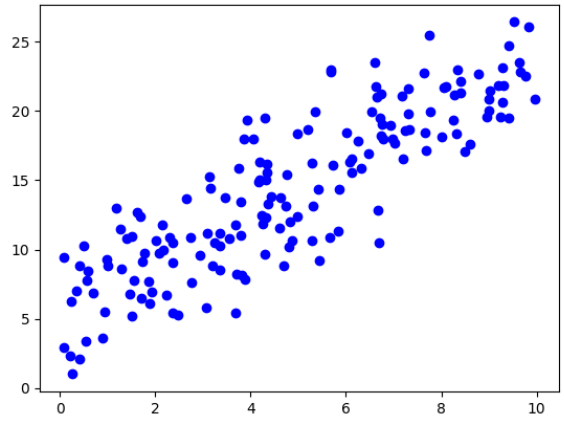
Families of Machine learning



What can we do with that?



What about in this case?



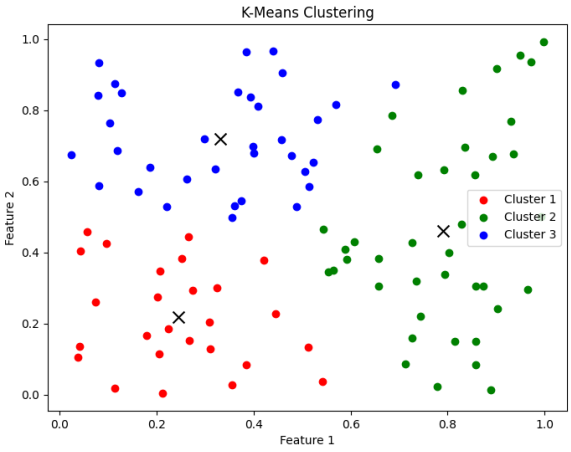
Python Source code 1

```python
import numpy as np
import matplotlib.pyplot as plt

def generate_data(n_random_points, noise=16):
    x = np.random.randn(n_random_points) * noise

    # Add noise
    y += np.random.randn(n_random_points) * noise

    return x, y

# Use the function to generate data
x, y = generate_data(n_random_points=166, noise=3)

# Plot all
plt.scatter(x, y, color='blue', label='Data Points')
plt.show()
```

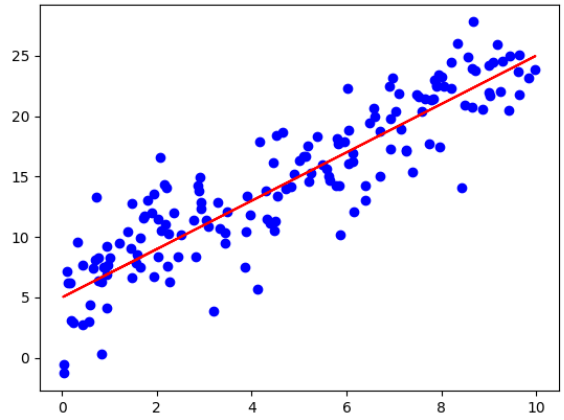
Python Source code 2

```python
import numpy as np
import matplotlib.pyplot as plt

def generate_linear_data(n_random_points, noise=16):
    x = np.random.rand(n_random_points) * 10

    # Make 'perfect' data
    true_slope,  true_intercept = 2, 5
    y = true_slope * x + true_intercept

    # Add noise
    y += np.random.randn(n_random_points)*noise

    return x, y, true_slope, true_intercept

# Use the function to generate data
x, y, true_slope, true_intercept = generate_linear_data(
        n_random_points=166,
        noise=3)

# Plot all
plt.scatter(x, y, color='blue', label='Data Points')
plt.show()
```

Unsupervised learning



Supervised learning



Is the data to decide?

This is why we focus so much on the data type.

The data properties dictate what statistical model can be adopted.

An statistical model has leverages our understanding of the data structure to improve its **predictions** (inference).

The numerical recipe that we used to generate the data is defined the **truth**

Psychology or data science?

Most Machine learning tools are aimed to find the truth. In most cases, we are happy to not find lies.

It is much easier to start from a model (hyphothesis) and collect the data accordingly than the other way around.

Unsupervised learning

Unsupervised learning, a term that resonates with the autonomy of machine intelligence, operates on the principle of identifying patterns and structures in datasets without labelled responses.

This branch of machine learning is distinguished by its lack of explicit guidance, where algorithms are tasked with uncovering hidden structures from unlabeled data.

The most common clustering strategies are :

- filtering
- clustering
- dimensionality reduction
- association learning

Application of unsupervised learning

It is a bit of a holy grail: a computer that finds patterns without guidance. (Yes, it doesn't work, most of the time)

Still, it has been shown efficient for:

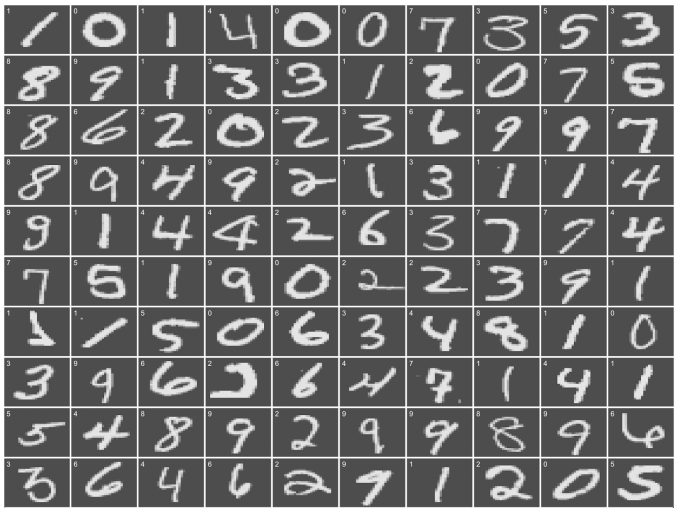
- Computer vision
- Anomaly detection
- Exploratory data analysis

Main challenge

The right result is quite undefined, Uncertain goal.

Consider the famous ML digits challenge:

Uncertain goal



Supervised learning

Supervised learning, a term that implies external intervention (not sure if from a human anymore...), operates on the principle of identifying patterns and structures in datasets with labelled responses.

Data and labels

In a data matrix, one or more coloums are selected as labels (or target, or dependent variables)

The task is to either operate a **regression** or a **classification**

Most common approaches

- *Linear Regression*
- *Logistic Regression*
- *Support Vector Machines (SVM)*
- *Neural Networks*

Weak supervised learning

A less popular type of machine learning problem is when labels are assigned to groups of instances.

The group of instances is called **bag**.

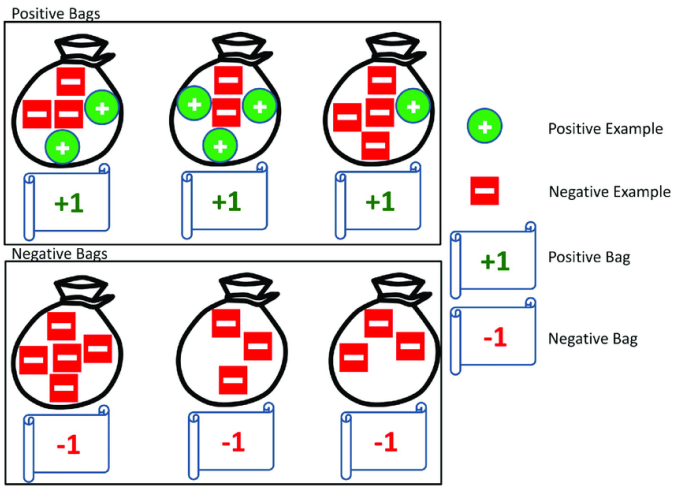
The question is, what is the level of a previously unforeseen bag?

This data structure and question type request a hybrid treatment between supervised and supervised learning.

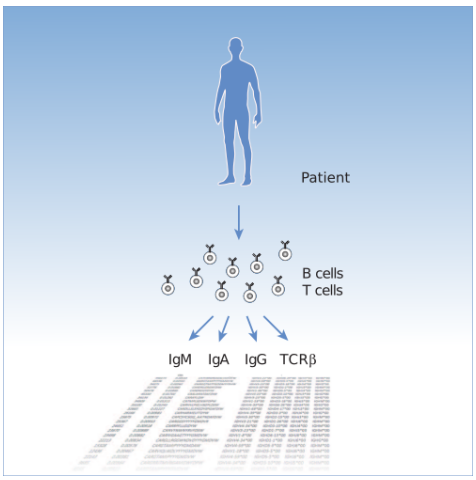
Multiple instance learing

Multiple instances are needed to learn (quite clear name)

Weak Supervised learning



Weak Supervised learning



Reinforcement learning

Finally, there is a further approach.

Reinforcement learning (RL)

It aims to train an intelligent agent to take actions in a dynamic environment in order to maximise the cumulative reward.

It learns from outcomes and decides which action to take next. After each action, the algorithm receives feedback that helps it determine whether the choice it made was correct, neutral or incorrect.

It is a self-teaching system that essentially learns by trial and error.

It is a dependable tool for automated decision making. The model outcome is the identification of the best set of 'actions' (e.g. chess).

What ML can do in GEO?

As discussed, Machine learning can be used in a large set of fields. In particular, in geo-applications one can:

1. Handle (too) large dataset
2. Automate long repetitive and low value tasks
3. Support decision making
4. Implement expert assesment
5. Provide visualization and analysis
6. Correct models with data (e.g. geosteering)

Application examples

It all starts from the question!

Is data fit for the task?

1. Prompt anamolous operation conditions (safe drilling).

Frequency? Resolution? Interpretability? Variable types?

1. Feature detection/guided interpretation.

Significance? Labeling strategy? Correlated data? Statistical interactions?

General Work Flow

Also in geo applications, a predominant quantity of time or efforts shall be dedicated to data retrival, preparation, interpretation.

1. Data preparation
2. Model selection
3. Model training and tuning
4. Model testing, inference and validation
5. Result intepretation

Data preparation

The hands on operation begin with the data!

1. metadata preparation
2. dataset preparation
3. filtering
4. cure data

Metadata

Data without metadata are just numbers: useless.

Metadata quality is a necessary pre-condition

What is metadata?

A set of data that described and fives information about other data [wiki]

Black box

Machine learning offers a (super) large set of statistical based tools that are, in essence, a black box.

Use with CAUTION!

As every black box, they are not fully reliable.

Here some guide:

1. Use the simplest approach!
2. Consider approach quality
3. Compare your results with benchmarks.
4. Avoid not reporducible studies.
5. Prioritize intepretability above computational speed and accuracy.