What is Machine Learning?

TM Quest

A Machine Learning Overview

Example of a Machine Learning Problem

Example

Take in an image of a handwritten numbers and output the number.

We know: The pixel placements where there is color.

We want: The actual number it represents.

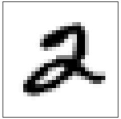


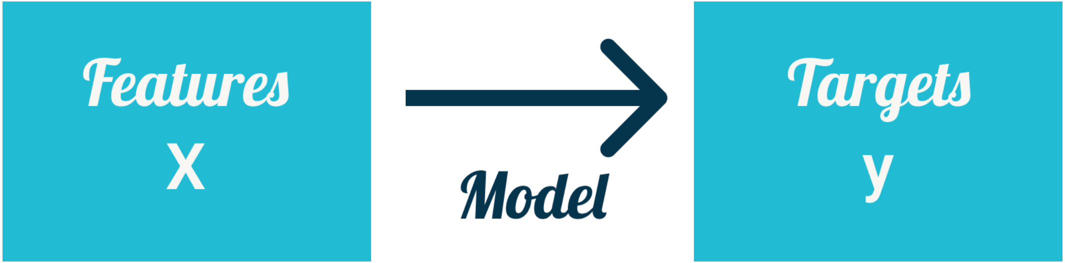
Figure: The model should give out 2 in this case

What is a Machine Learning Model?

- An algorithm that takes in input(s) to predict output(s). The algorithm should ”learn” from the data to become more accurate.

Definition

The input(s) are called features (or predictors) and the output(s) are called targets.



Features

X

Model

Targets

y

More Examples of Machine Learning Problems

Example

- Take in patient data and output disease probability.

  Features: Age, sex, blood tests, bmi, etc.
  Target: A number indicating the disease probability.

- Classify penguins into penguin species.

  Features: Beak length, height, colors, fur thickness, etc.
  Target: The species of the penguin in question.

- Separate consumers based on consumer data.

  Features: Time of shopping, previous purchases, country of residence, etc.
  Goal: Get a better understanding of the consumer.

ML Terminology

Supervised vs. Unsupervised Machine Learning

Definition

- Supervised machine learning problems are problems where one is given a set of features and corresponding targets.
- We make the model ”learn” how to deduce the targets from the features through a process called training.

Definition

- Unsupervised machine learning problems are problems where one is given a set of features. The targets are unknown.
- We will only consider unsupervised machine learning much later in the course, so don’t worry!

Regression vs. Classification Problems

In supervised machine learning, we separate regression and classification problems.

Definition

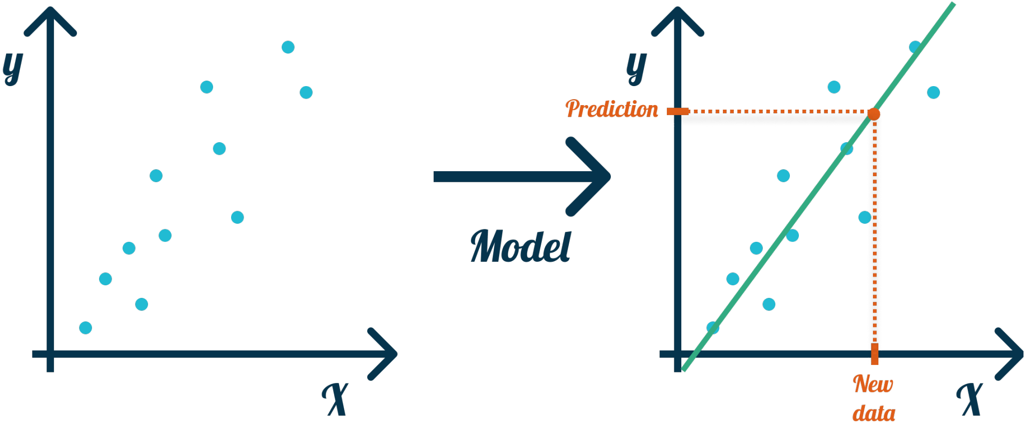
If we are trying to predict a target which is:

- A continuous number, then the problem is called a regression problem.
- A category (such as red, green, and blue), then the problem is called a classification problem.

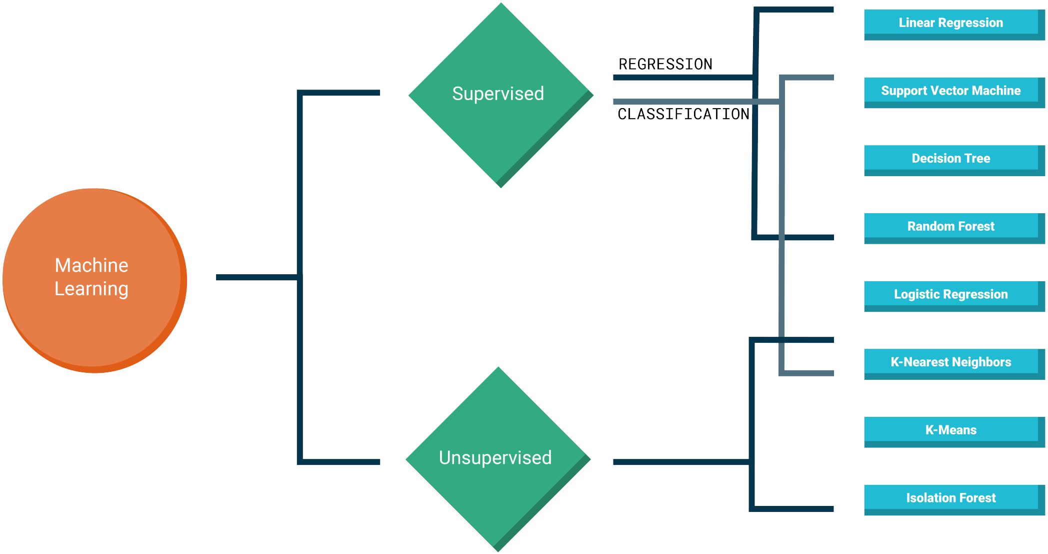
Example

- Hand-written number detection -> Supervised classification problem!
- Disease probability -> Supervised regression problem!
- Penguin species classification -> Supervised classification problem!
- Consumer understanding -> Unsupervised clustering problem!

Illustration of a Regression Problem

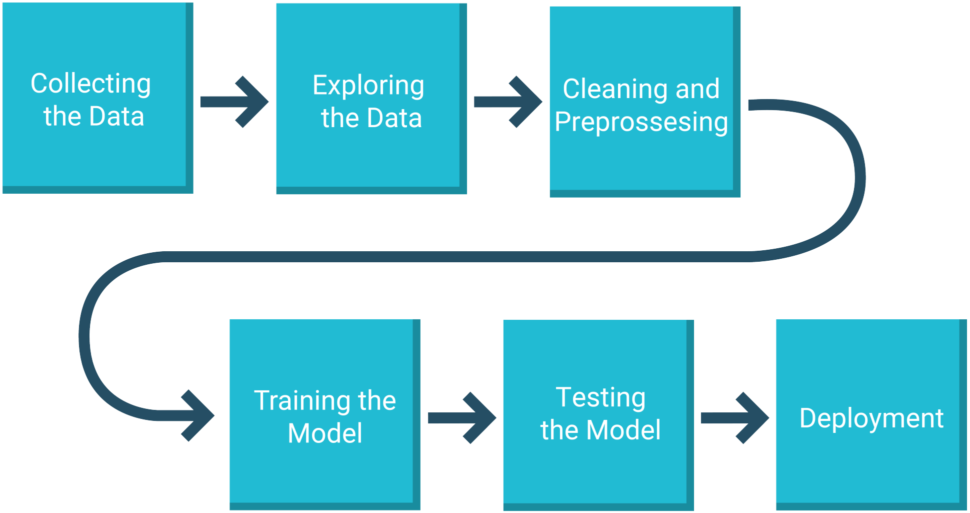


A Few Examples for the Future!



Anatomy of a ML Project

Anatomy of a ML Project



Collecting the Data

Exploring the Data

Cleaning and Preprossesing

Training the Model

Testing the Model

Deployment

Step 1: Collecting the Data

Not part of this course.

Where to Get the Data?

- A Survey
- Web-scraping
- Log-data
- Sensor Data
- Government Statistics

What we Hope the Data is:

- Accurate—The data is not wrongly reported
- Complete—No/little missing data
- Informative—The data contains features which are relevant for the task

Step 2: Exploring the Data

Goal: Get a better understanding of the data.

Ways to Understand the Data

- Ask how the data was collected
- Understand what the features/targets represent
- Speak to domain experts about the relevance of the features
- Plot the data and gauge relationships
- Find correlations between features and targets

We will learn more about this point in this module!

Step 3: Cleaning and Preprocessing

Definition

- Cleaning a dataset is to make the data uniform and in the correct format.
- Preprocessing is everything you do after collecting the data and before training the model.

Example

- In a survey, you ask people in France and the USA to report their height.
  - France uses the metric system (the only right way).
  - USA use the imperial system (an outdated relic of the past).
- Fill in (or drop) missing data.
- Scaling the data so that the different features are similar in size.

Step 4: Training the Model

After we have cleaned the data, we need to train the model. What does this mean?

What is Training?

- A machine learning model uses the available data to learn the connections between the features and the target.
- Each machine learning model learns the connections in the data in a different way.
- Requesting the model to learn the connections between the data is called training the model.

We will learn more about training in the next module!

Step 5: Testing the Model

So, you've built an amazing (or so you think) machine learning model. Great! But how do you know that it is really good? We need to test the model.

What is Testing?

- We test a machine learning model by seeing how it performs on new data.
- There are several different criteria (metrics) we can investigate when testing a model.
- Evaluating the model on sufficiently many new data values with a chosen metric is referred to as testing the model.

We will learn more about testing in the next module!

Step 6: Deployment

Not part of this course.

Now is your time to shine and show the world (or your team) what you have produced!

What to do With Your Model?

- Embed your model into a web-page or a widget.
- Serve your model as an API so that others can use it with ease!
- Automate the process of running your model in the background.
- Draw conclusions about the relations in your data by using the model.
- Track your model and make sure that it still performs well when the data changes.

Introducing Scikit-Learn

What is Scikit-Learn?

- Science kit learn
- Used for Machine Learning in Python
- Webpage: https://scikit-learn.org/stable/index.html

```
# One mostly import spesific machine learning models
    from sklearn
from sklearn.linear_model import LinearRegression
```