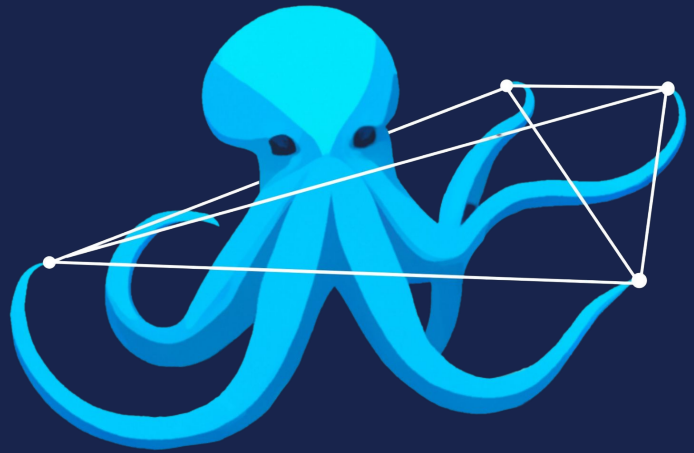
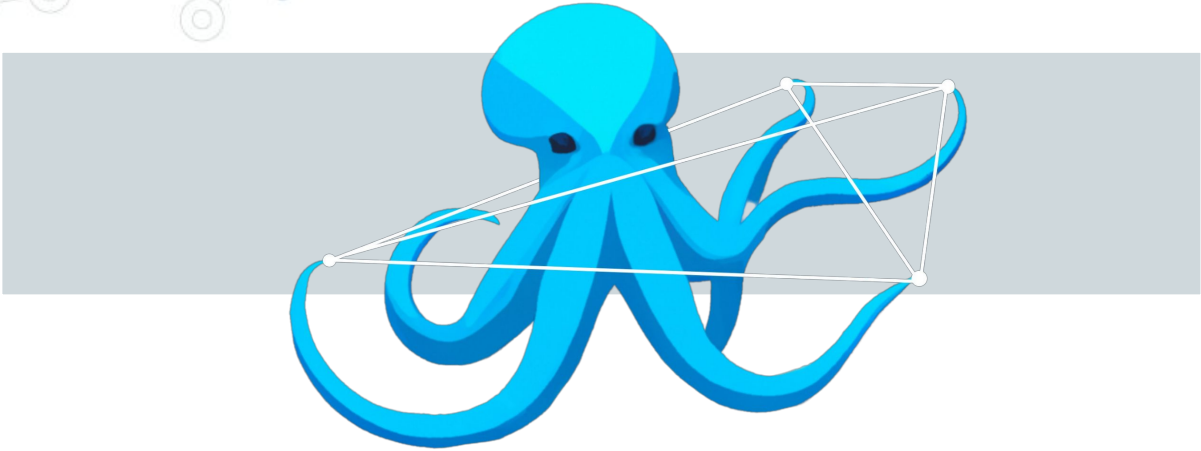


Machine Learning Month

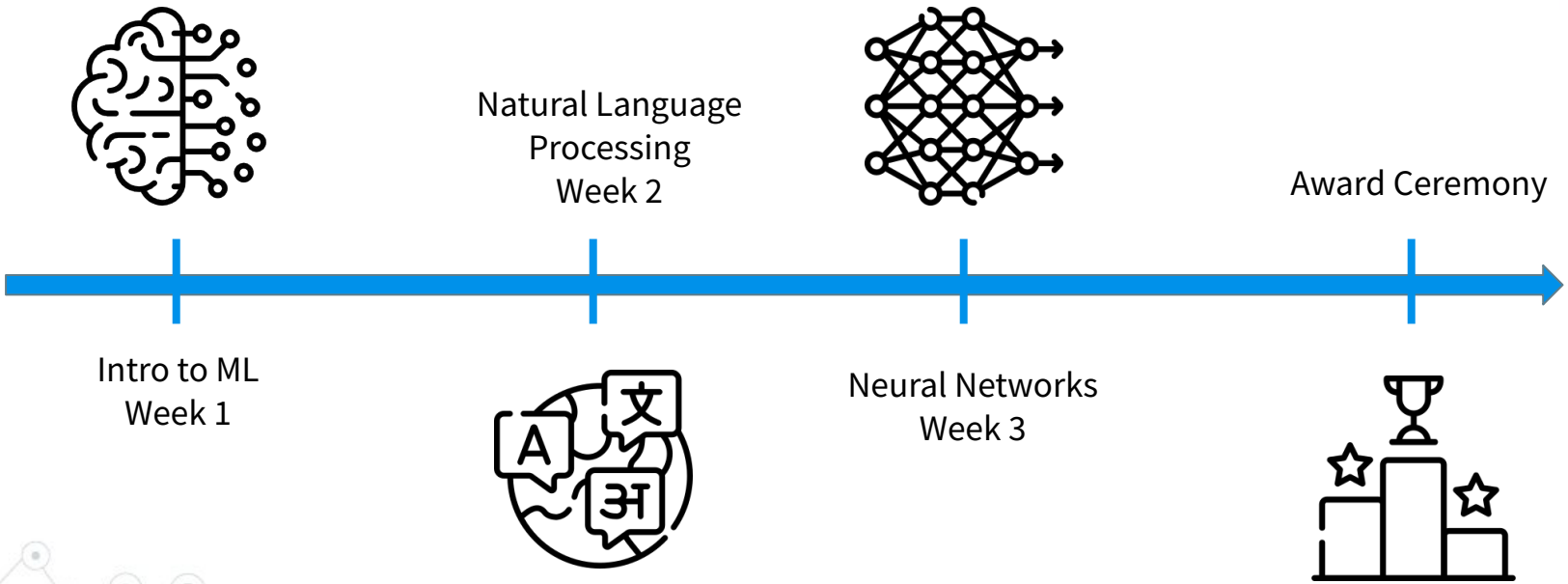
Introduction to ML with Sklearn



Who are we?

Fully Connected Graph

Overview of the Month



Welcome to the First Lecture

Our Objectives:

- Why ML? Types of problems we can solve
- Identify types of data and how to prep them
- Multiple machine learning algorithms
- Practice implementation in Colab
- Prepare for the Kaggle competition

Overview of ML

Will a person like a movie?

Will a person like a movie?

Yes.

Will a person like a movie?

Yes.

Will a person like a movie?

Yes.

Definitions

Some Definitions

- x – the object, it's features
- X – the space of objects
- $y(x)$ – the answer for an object, target feature
- Y – the space of answers

Types of Machine Learning

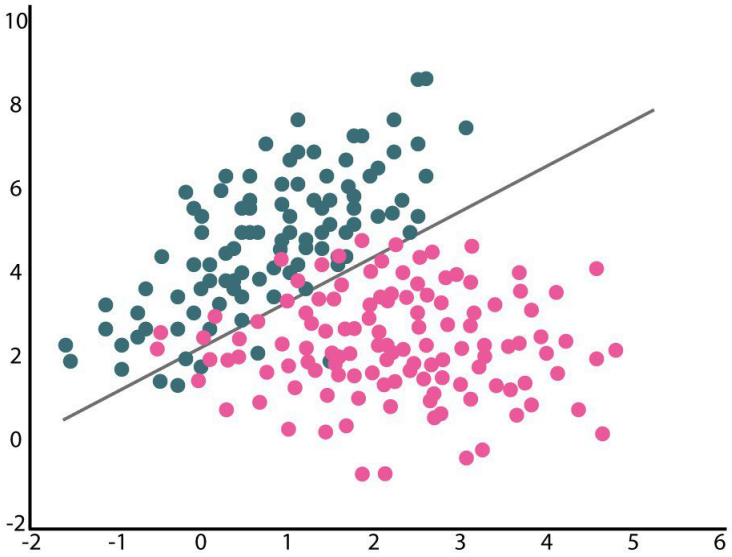
Supervised Learning:

- Data provided is complete
- Useful for prediction and classification

Example of Machine Learning Tasks

- Will a user like a film?
- Will a person return a loan?

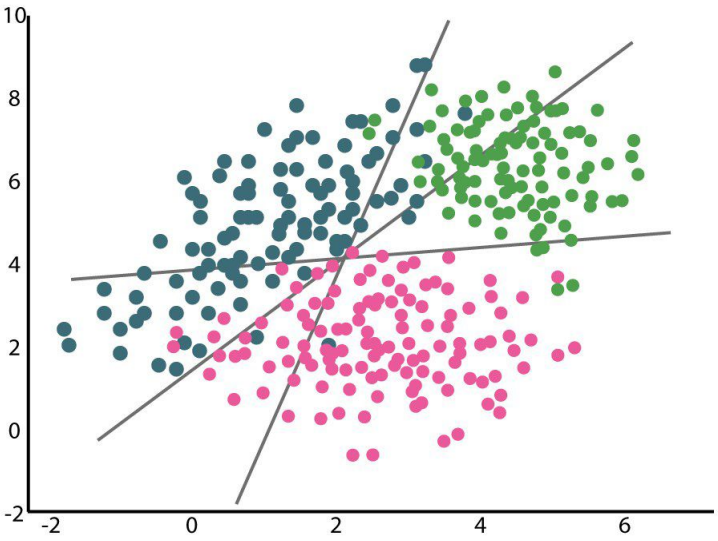
$$Y = \{0, 1\}$$



Binary Classification

Example of Machine Learning Tasks

- What is the topic of an article?
- What sort of an apple is this?
- What type of vehicle is in the image: motorcycle, car or a van?

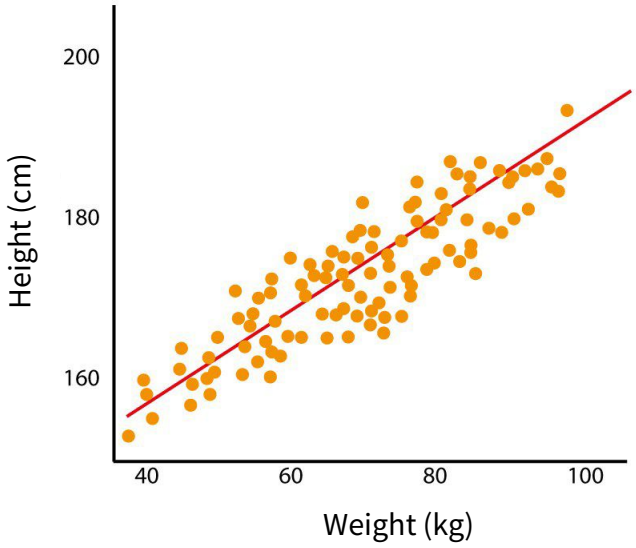


Multi-Class Classification

Example of Machine Learning Tasks

- Weather forecast for tomorrow
- Revenue prediction
- Determining age by photograph

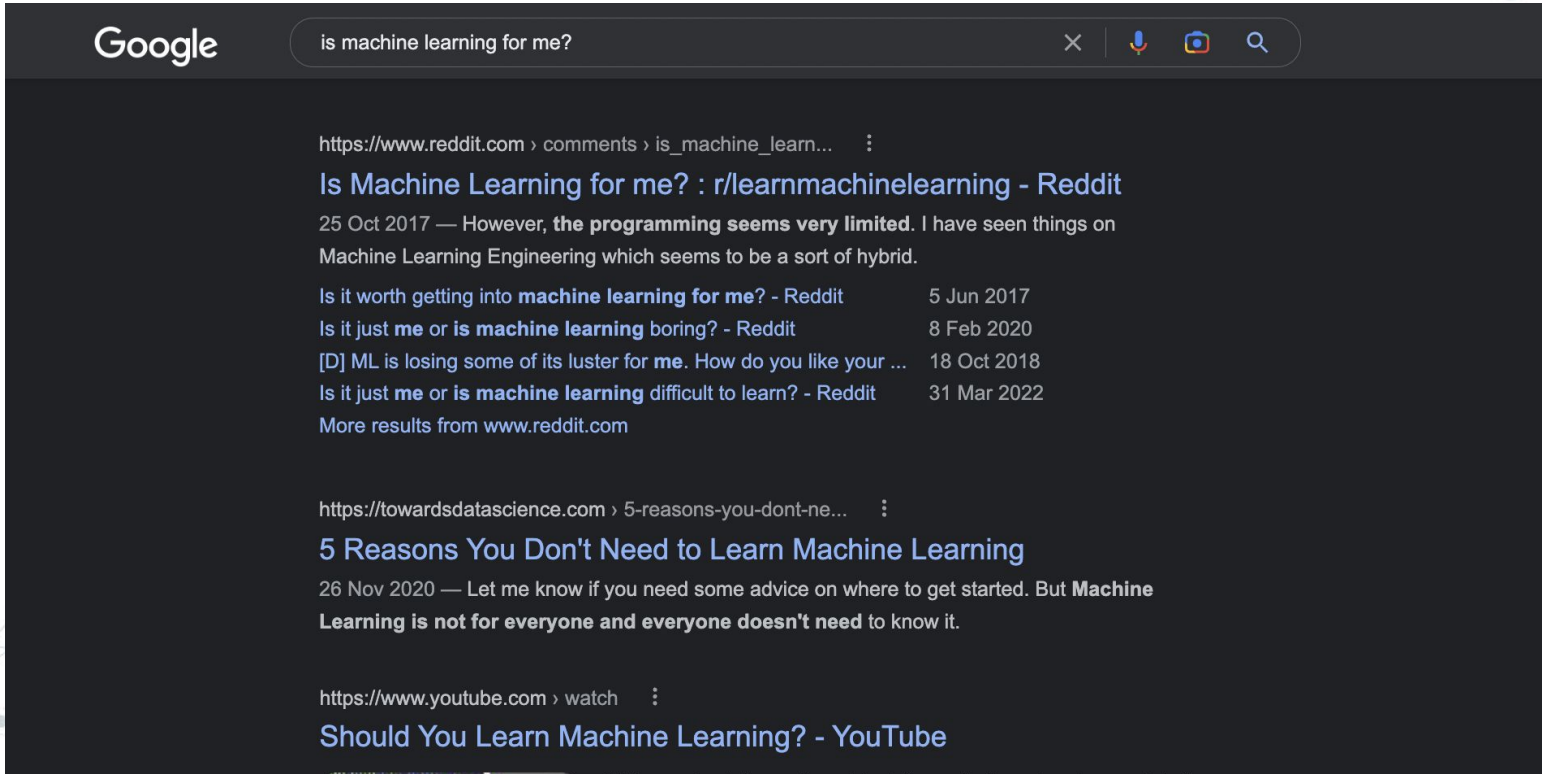
y is real valued



Regression

Example of Machine Learning Tasks

Ranking



Types of Machine Learning

Supervised Learning:

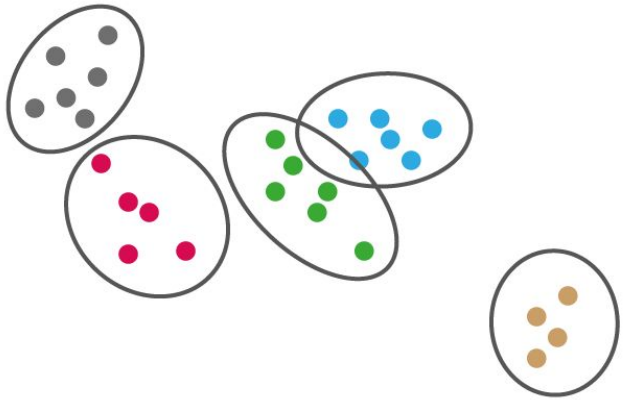
- Data provided is complete
- Useful for prediction and classification

Unsupervised Learning:

- Data is missing the goal value
- Useful for uncovering hidden patterns

Example of Machine Learning Tasks

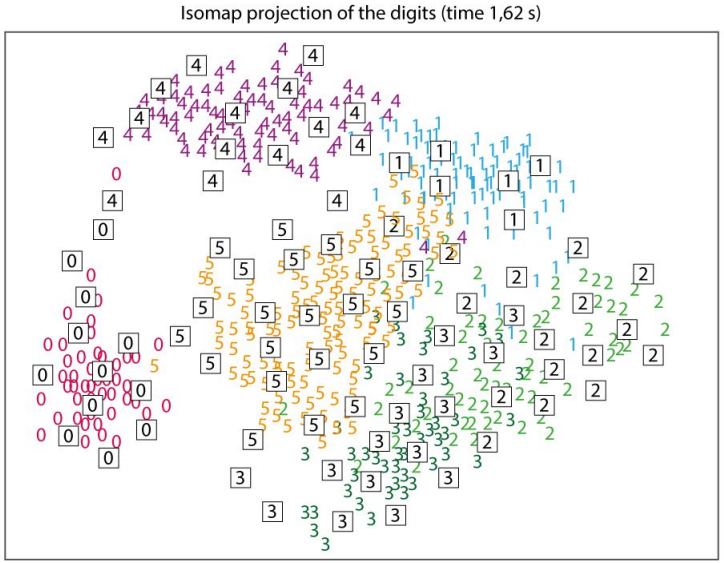
- User segmentation
- Search for similar users in social media
- Search for genes with similar representations



Clustering

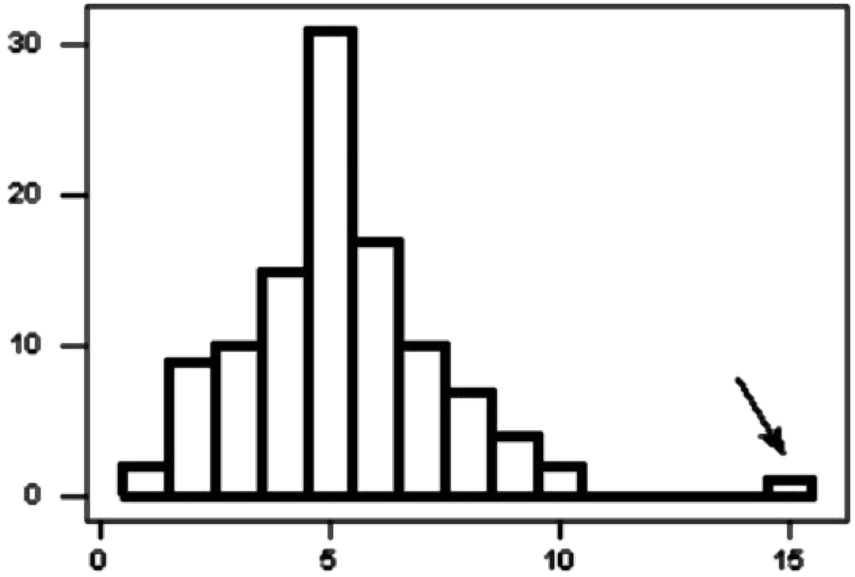
Example of Machine Learning Tasks

- Visualising d-dimensional data in a way that visually shows the structure of data



Visualisation

Example of Machine Learning Tasks



Anomaly detection

Types of Data

Binary features

Can take up one of two values $D_j = \{0, 1\}$

- Is this a cat?
- Did the revenue increase?

Answer yes corresponds to 1, the answer no – 0

Continuous features

Can take up values $D_j = \mathbb{R}$

- Age
- Area of an apartment
- Number of products bought

The last feature corresponds to the set of natural numbers, but it's treated as continuous

Categorical features

They take up values from D_j – an unordered set

- Color of eyes
- City
- Education Level

Categorical features

They take up values from D_j – an unordered set

- Color of eyes
- City
- Education Level (sometimes can have an order)

They are hard to deal with. There are new methods made of how to account for them in machine learning models.

Ranking features

Type of categorical features D_j that can be ordered

- Role in the movie (main, secondary, background)
- Type of populated area (city, town, village)
- Education Level (PhD, Master, Bachelor, Undergrad)

Ranking features

Type of categorical features D_j that can be ordered

- Role in the movie (main, secondary, background)
- Type of populated area (city, town, village)
- Education Level (PhD, Master, Bachelor, Undergrad)

The *distance* between two features doesn't make sense

Features of a dataset

	A	B	C	D	E	F	G
1	id	title	city	postalCode	latitude	longitude	areaSqm
2	0	West-Varkenoor	Rotterdam	3074HN	51.896601	4.514993	14
3	3	Ruiterakker	Assen	9407BG	53.013494	6.561012	16
4	8	Brusselseweg	Maastricht	6217GX	50.860841	5.671673	16
5	10	Donkerslootstra	Rotterdam	3074WL	51.893195	4.516478	25
6	12	Vorselenburgstra	Alphen aan den	2405XJ	52.122335	4.661434	10
7	17	Groenhoven	Amsterdam	1103LW	52.326211	4.976048	19
8	18	Noorderhagen	Enschede	7511EL	52.221643	6.894667	21
9	19	Jaersveltstraat	Rotterdam	3082SJ	51.890481	4.466388	16
10	20	Tongerseweg	Maastricht	6213GB	50.841744	5.670447	17
11	21	Lange Marktstra	Leeuwarden	8911AD	53.197261	5.790455	19
12	22	Guido Gezellest	Eindhoven	5615HL	51.431324	5.475464	16
13	23	Ank van der Moe	Amsterdam	1065LH	52.352244	4.824007	12
14	25	Beatrixstraat	Enschede	7511KL	52.221827	6.902143	20
15	28	Tesselschadestr	Leeuwarden	8913HA	53.198638	5.782587	51

Target feature

Targets refer to the values that we are trying to predict

In the competition, the only target is the rent of a property

rent
500
290
425
600
425
750
240
500
660
412

Break Time

Data Preprocessing

What is Data Preprocessing?

Before passing data to a ML model, it has to be prepared.

- Data normalization
- Missing values
- Handling categorical features

Handling Categorical Data

We can use one hot encoding to represent categorical data.

In this technique, we represent the item as multiple binary values for each possible outcome.

Ex: Suppose we had a category ‘color’ which consisted of “red”, “green” and “blue”

Handling Categorical Data

In one-hot encoding, the data would be represented as:

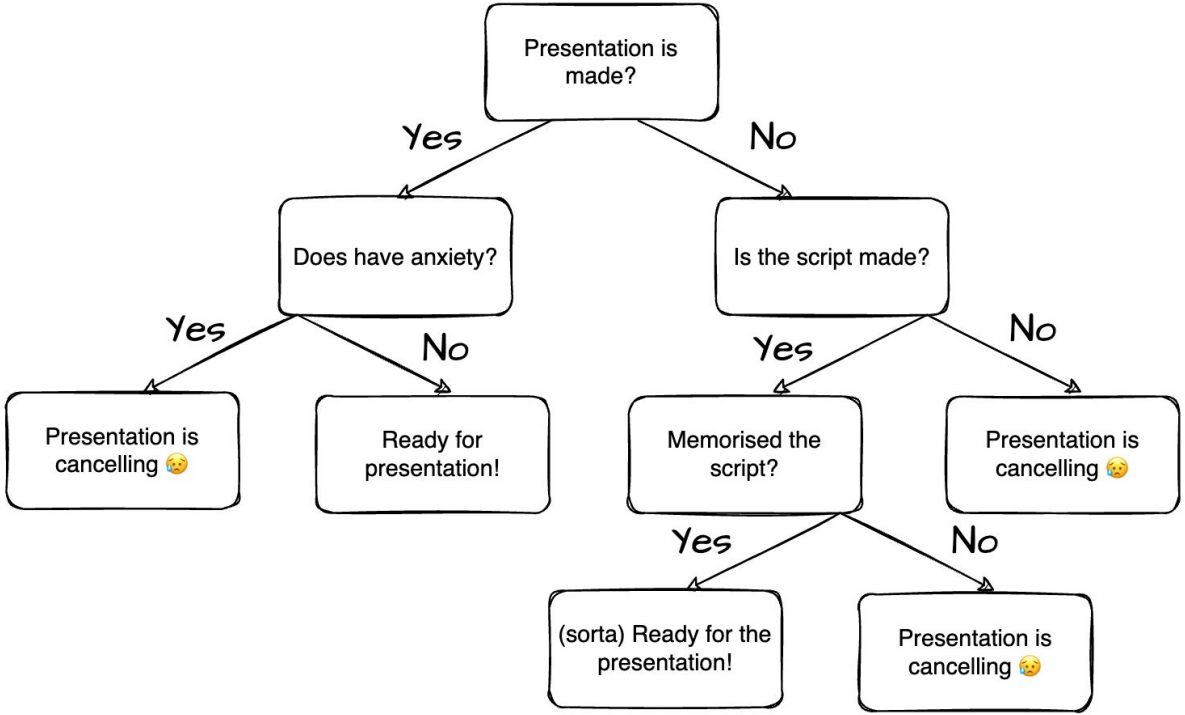
ID	Red	Green	Blue
1	1	0	0
2	0	1	0
3	0	0	1

ML Model

A few ML techniques

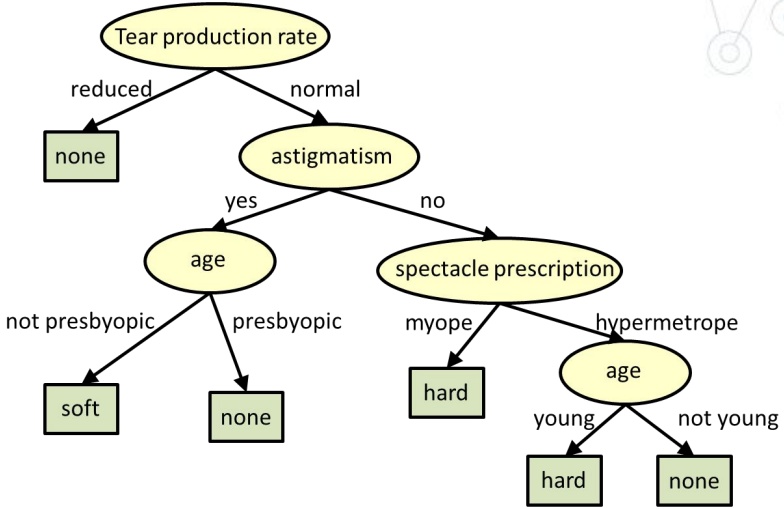
- Decision Trees
- Random Forests
- Linear Regression
- Gradient Boosting
- Neural Networks (Discussed in Lecture 3)

Decision Trees



Decision Trees

- Predicts the dependent variable using inference rules from the given data
- Groups samples with similar values together
- Arrives at a prediction by answering a series of if … else statements

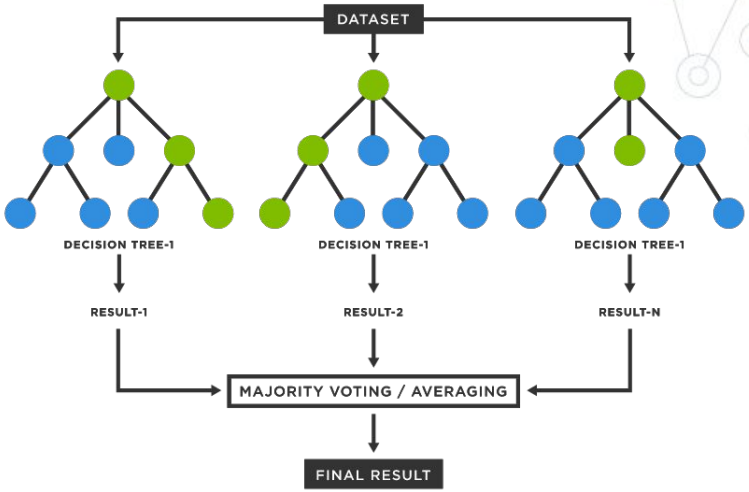


Decision Trees

Strengths	Weaknesses
Can be visualized and explained easily	Prone to overfitting
Can handle both numerical and categorical features	Unstable since even small alterations can change results

Random Forests

- Derivative of Decision Trees
- The dataset is randomly split into several components and a decision tree is created from each of these
- The output value is the value majority voted for



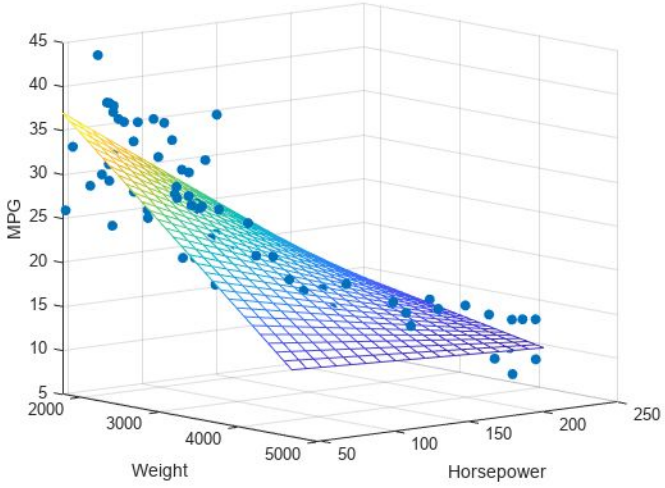
Random Forests

Strengths	Weaknesses
Resistant to overfitting	Loss of Interpretability
Higher accuracy than DTs	

Linear Regression

- Simple model that linearly predicts the goal given features.

- The objective of linear regression is to generate a line of best fit which minimizes the Residual Sum of squares.



$$y(x_1, x_2, \ldots x_n) = c_0 + c_1 x_1 + c_2 x_2 + \ldots + c_n x_n$$

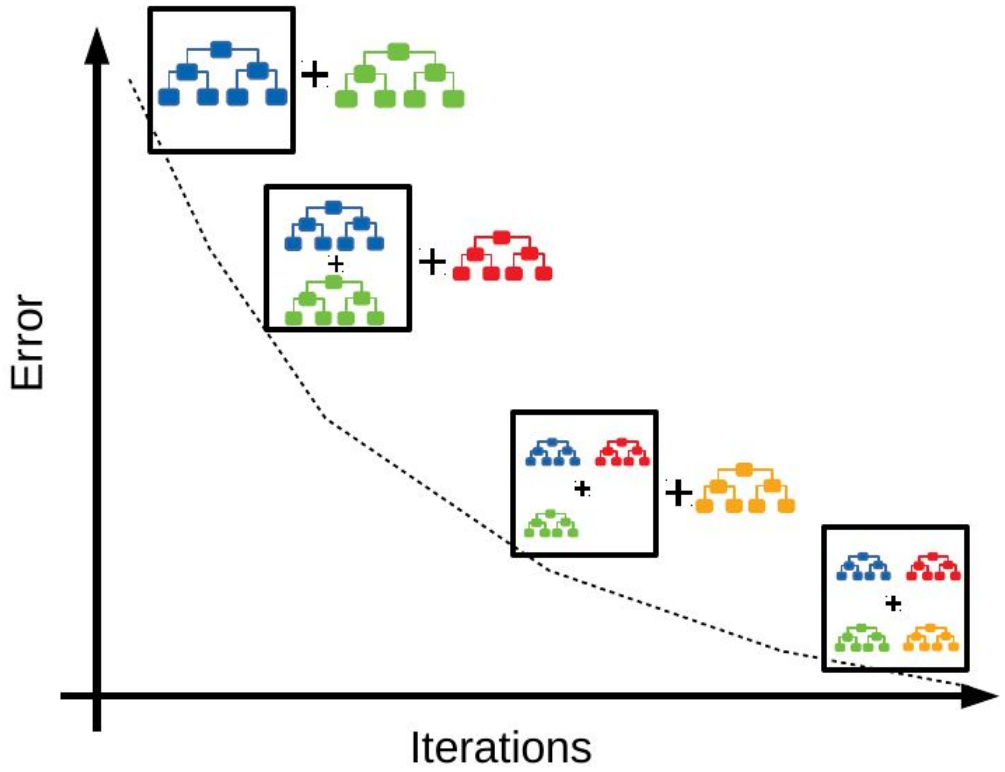
$$y(\mathbf{x}) = c_0 + \mathbf{c}\mathbf{x}$$

Linear Regression

Strengths	Weaknesses
Performs well for linearly separable data	Sensitive to Outliers
Resistant to overfitting due to generalization	Cannot handle categorical values

Gradient Boosting

Gradient Boosting



Gradient Boosting

- Derivatives from decision trees
- Further trains the trees on sections of the data that it has problems with
- Over time, these weaknesses are covered, leading to high accuracy

Gradient Boosting

Strengths	Weaknesses
High Accuracy	Expensive to train
Highly flexible and customizable	Prone to overfit

Practice Session

Installation

Machine Learning is built using an extensive set of libraries.

It can be difficult to get code working locally

For convenience, we can use Google Colab to save time

Using Colab

Jupyter notebooks allow us to execute code while also using markdown to provide comments

Open the link and go to the lecture materials

mlmonth.svcover.nl

The Problem

We have acquired a dataset of the grades of students in Portugal throughout a year. The dataset also contains their demographic data as well.

Using the acquired data, we wish to predict the final grade (G3) of a student.

Tools

Model training with sklearn

- Free ML Library in Python
- Contains useful features such as pre-processing
- Contains implementations of several ML models

Getting Started is Simple

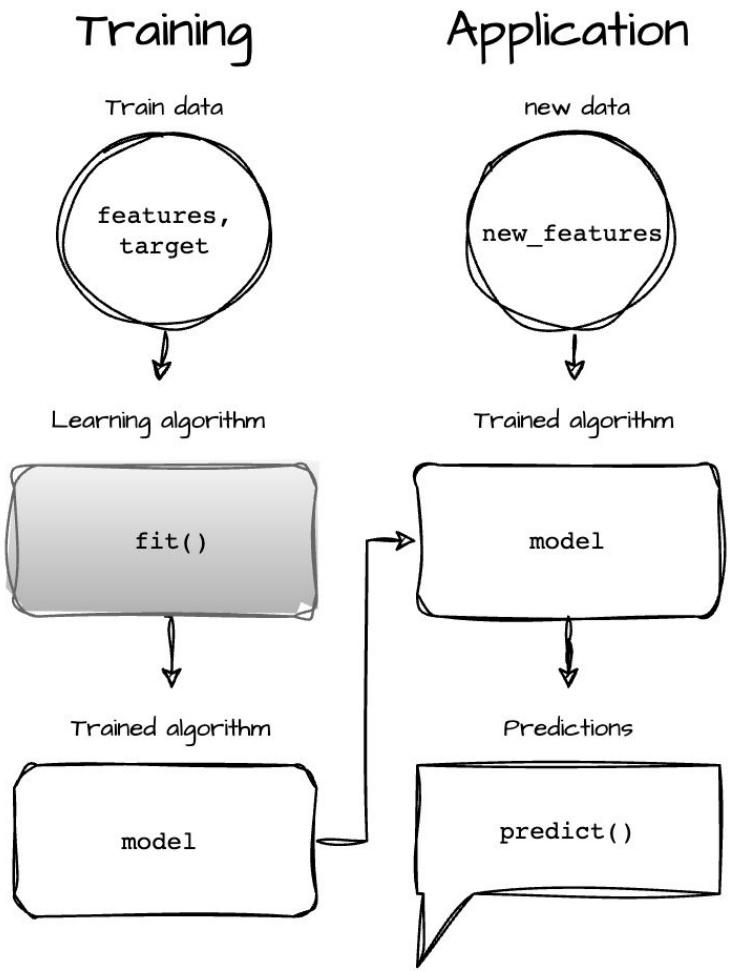
```
pip install sklearn
```

Data Handling with Pandas

Along with sklearn, pandas is helpful in handling data

The pandas dataframe is powerful and allows for convenient data access.

Installation: pip install pandas


Training
Application
Train data
features, target
new data
new_features
Learning algorithm
Trained algorithm
fit()
model
Trained algorithm
Predictions
model
predict()

Kaggle Preparation

Preparation for the Competition

Important things to know:

- Link to Competition
- Read the New to Kaggle section
- Build your model
- Test your model
- Submit your results
- Improve!

November 24

Introduction to Natural Language Processing

Thank you for your attention!

Do you have any more questions? Join our Discord server