

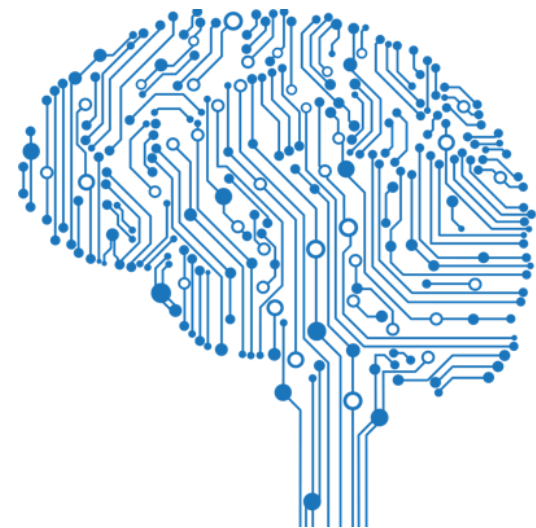
Introduction to Machine Learning

From raw data to predictive models

Yordan Darakchiev

Technical Trainer

iordan93@gmail.com





sli.do

#MachineLearning

Table of Contents

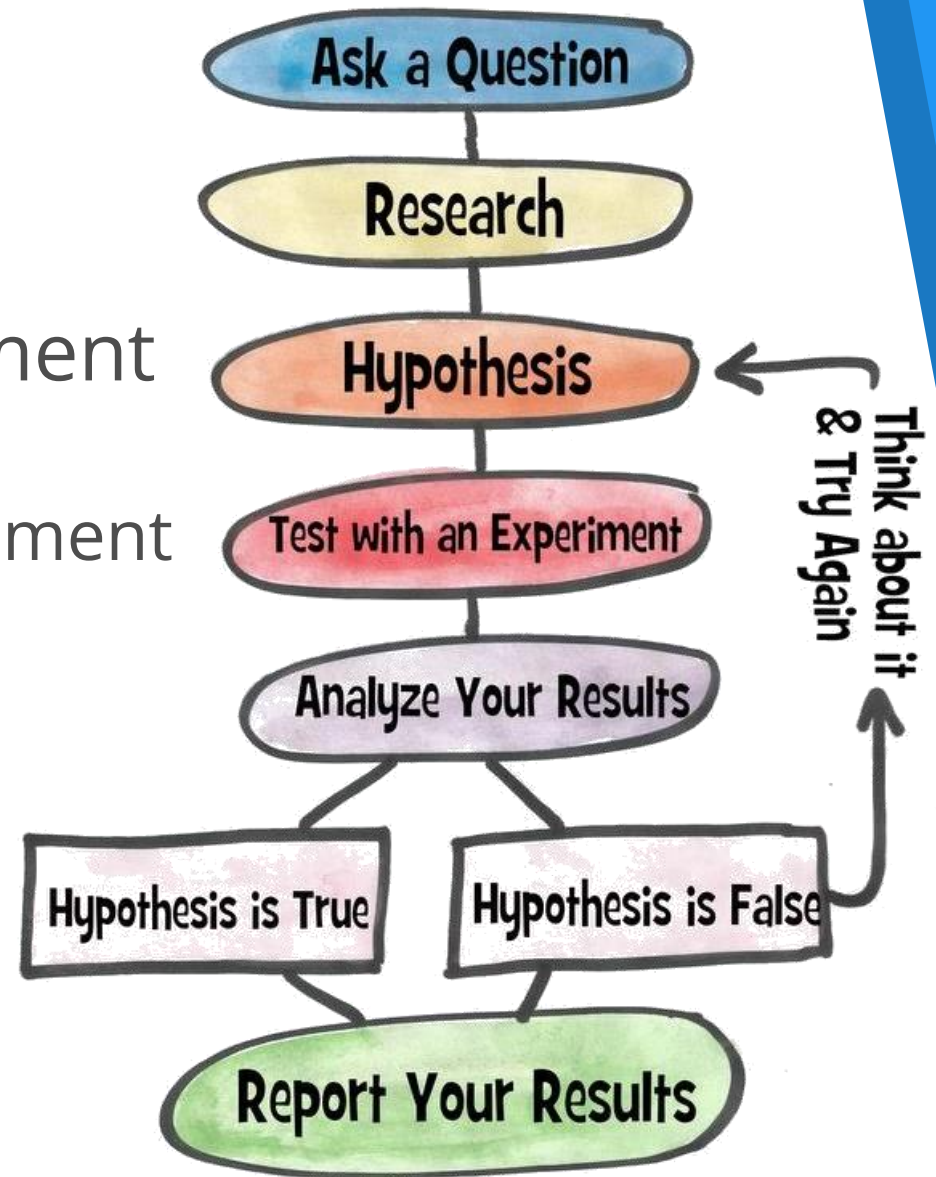
- The scientific method – overview
 - Knowledge discovery from data
- Machine learning
 - Basic concepts
 - Algorithms (models) overview
- Getting, preparing and exploring data
 - Review
- Machine learning process overview

The Scientific Method

How to work with data...
the right way

The Scientific Method Steps

- Ask a question
- Do a research
- Form a hypothesis
- Test the hypothesis with an experiment
 - Experiment works \Rightarrow Analyze the data
 - Experiment doesn't work \Rightarrow Fix experiment
- Results align with hypothesis \Rightarrow OK
- Results don't align with hypothesis \Rightarrow new question, new hypothesis
- Communicate the results



OSEMN Model

- Some guidelines on the process to extract meaningful information from data
 - Very similar to the scientific method
 - Can be viewed as a sequential process
 - Or just as some guidelines on how to do research
 - Read as "awesome"
1. **O**btain data
 2. **S**crub data
 3. **E**xplore data
 4. **M**odel data
 5. **iN**terpret the results

Applied Machine Learning Process

- This allows us to do our job faster and more reliably

1. Problem definition

- Make sure the problem is well-defined and that you're solving the right problem

2. Data analysis

- Get familiar with the available data

3. Data preparation

- Get the data ready for modelling

4. Algorithm evaluation

- Test and compare algorithms

5. Result improvement

- Use results to create better models (e.g. fine-tuning, ensembles)

6. Result presentation

- Describe the problem and solution to non-specialists





Machine Learning

Fundamental concepts

Machine Learning

- We described a general process
 - We didn't explain ML in detail
- *"A computer program is said to learn from experience E with respect to some task T and some performance measure P , if its performance on T , as measured by P , improves with experience E ." – Tom Mitchell, Carnegie Mellon University*
- More simply, **making computers learn from data**
 - And observing them getting better and better
 - Results: **computers do things that they weren't explicitly told**
- The field is vast (and expanding)
 - There are many sub-fields, variations and algorithms
 - ... but the basis is still the same

Types of Machine Learning Algorithms

■ Supervised learning

- We train the program on previously known (labelled) data
- After training, we expect it to make predictions on new data
- Examples: regression, classification

■ Unsupervised learning

- We leave the program to find patterns in data
- Examples: clustering analysis, dimensionality reduction

■ Reinforcement learning

- A form of unsupervised learning
- The program learns continuously
- Examples: learning to play a game by observing other players, learning to drive a car

Algorithms by Task

- **Statistical algorithms**
- **Regression** – predicting a continuous variable
- **Classification** – predicting class labels
- **Clustering** – finding compact groups of data points
- **Dimensionality reduction** – simplifying the input data
- **Recommendation** – suggest items for users
- **Optimization** – minimize / maximize a target function
- **Testing and improvement algorithms** – helper algorithms to select, fine-tune and optimize other ML algorithms
- ... and more :)

Getting and Preparing Data

**Review: Preparing raw data
for modelling**

Common Libraries

- In Python, we use libraries to perform common operations
- **scikit-learn** – machine learning models
- **pandas** – working with data
 - Reading, tidying, cleaning, preparation
- **numpy** and **scipy** – numerical and scientific libraries
 - Contain a ton of useful functions for performing research
- **matplotlib** – plotting and data visualization
- There are many more we'd like to use but these are the most commonly used ones

Getting and Preparing Data

- [10 Minutes to pandas](#)
- [Pandas Cheat Sheet](#)
- [Full docs](#)
- Tidy up the data
- Preprocess the data w.r.t. the task at hand
- Explore the data
 - Exploratory data analysis
 - Don't forget to make graphs
- Create meaningful features
 - Feature {selection, extraction, engineering}
- Example: Titanic dataset

Example: Preparing Data for Modelling

- Most models require two additional steps
 - **Convert categorical variables** into **indicator variables**

```
dataset = pd.get_dummies(dataset)
```
 - **Normalize values** if needed (e.g., scale all variables from 0 to 1 using min-max scaling, or use Z-scores)
- Perform other model-specific transformations
 - E.g., your model may not work well with highly imbalanced data (when you look for anomalies)
- If possible, prepare several versions of the dataset
 - To see how a transformation affects model performance
- **Describe and document the entire process!**
 - Don't forget the rules for reproducible research

Summary

- The scientific method – overview
 - Knowledge discovery from data
- Machine learning
 - Basic concepts
 - Algorithms (models) overview
- Getting, preparing and exploring data
 - Review
- Machine learning process overview

The image features a white background with two blue decorative bars. The top bar is a solid blue strip. The bottom bar is a gradient blue strip that transitions from a lighter blue on the left to a darker blue on the right. The word "Questions?" is centered in a blue, sans-serif font.

Questions?