University of BRISTOL

INTRODUCTION TO MACHINE LEARNING FUNDAMENTAL

BRMS PGR Python Training

Zhaozhen Xu, zhaozhen.xu [at] bristol.ac.uk

11.11.2024

bristol.ac.uk

This morning…

- Basic concepts of Machine Learning
- Examples in Health Data Science
- Apply Machine Learning in Python

What is Machine Learning?

What is Machine Learning?

Loss: Difference between prediction and ground truth

Training:
Adjust the Model

Training Data:
Examples for the model to learn

Model

Prediction

? CAT

Ground Truth (label):correct or true information used as a reference

What is Machine Learning?

- Focus on algorithms that help computer learn from the data without a human telling the computer exactly what to do

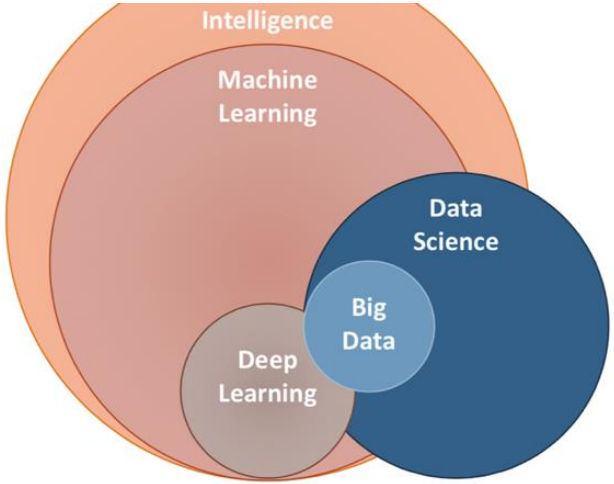
- Using the right features to build the right models that achieve the right tasks
  - Task: a problem to solve
  - Model: machine learning algorithms
  - Feature: abstract ‘language’ used to define an object (e.g. pixel of the picture, biomarker, demographics, etc.)

Artificial Intelligence, Machine Learning, Data Science

- Artificial Intelligence (AI): the goal is to enable computers/machines to perform human-like tasks and simulate human behaviour

- Machine Learning: Subset of AI, tries to solve a specific problem and make predictions using data

- Data Science: find pattern and draw insight from data (may use Machine Learning)



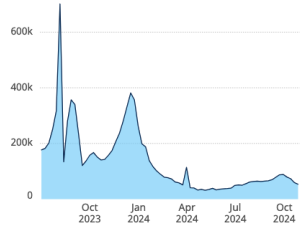
Source: https://ai.plainenglish.io/data-science-vs-artificial-intelligence-vs-machine-learning-vs-deep-learning-50d3718d51e5

Applications in Health Science

- **Medical Imaging Diagnosis: Image data**
  - Analyse images from X-rays, MRIs, CT scans, …
  - Detect tumours, fractures, infections, …

- **Disease Risk Prediction: Numerical data, Categorical Data**
  - Identify genetic mutations, or biomarkers
  - Predict a person's risk of developing specific disease

- **Information Extraction for Electronic Health Records (EHRs): Text data**
  - Process unstructured data (e.g. text) within EHRs
  - Extract family histories, historical diagnoses, …



- **Pandemic Modelling: Time-series data**
  - Analyse epidemiological data (e.g. case count, test rate), demographic, healthcare system data, …
  - Forecast the case of infection in the next 30 days

Applications in Health Science

- **Disease Diagnosis Assistance**
- **Treatment Plans**
  - Suggest prescription, surgical plan,…
- **Drug Discovery and Development**
  - Target Identification
  - Drug Repurposing
- **Medical Literature Mining**

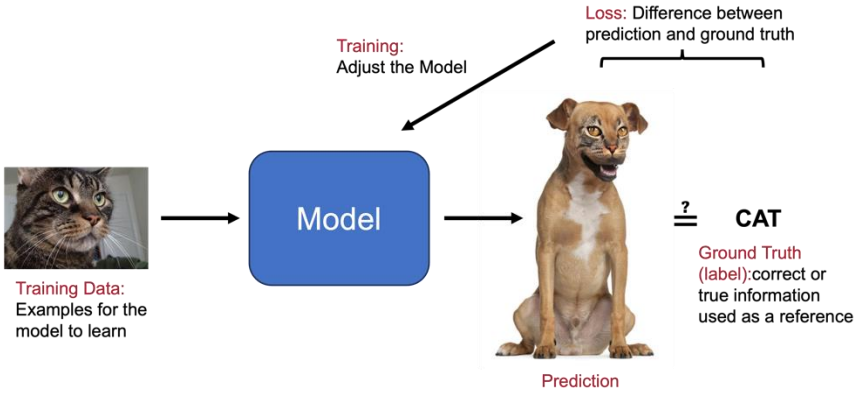
…

Supervised Learning

A model is trained on a labelled dataset.

Each training example has both input data (features) and a known output label (target).

- Classification
- Regression



Supervised Learning: Classification

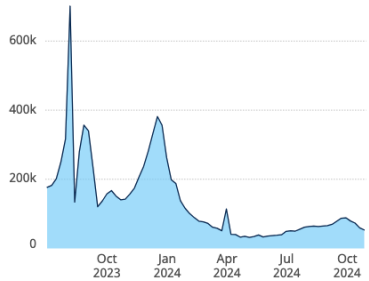
- Classify data into predefined categories or classes
  - Binary Classification (e.g. cat or dog)
  - Multi-Class Classification (e.g. cat, dog, tiger, or lion)

- Classification Algorithms:
  - Support Vector Machine (SVM)
  - Naïve Bayes
  - XGBoost
  - …
- Examples:
  - Early Classification of Sepsis in ICU Patients
  - Identifying Pneumonia in Chest X-rays

Supervised Learning: Regression

- Linear Regression
  - Calculate the effect size between two variables (e.g. body mass index and blood pressure)
  - In machine learning, we focus on predicting. Assuming two variables have linear relationship, we use the independent variable A (input) to predict the value of the dependent variable B (output)

- Regression Algorithms:
  - Decision Tree
  - Support Vector Regression
  - Random Forest
  - …



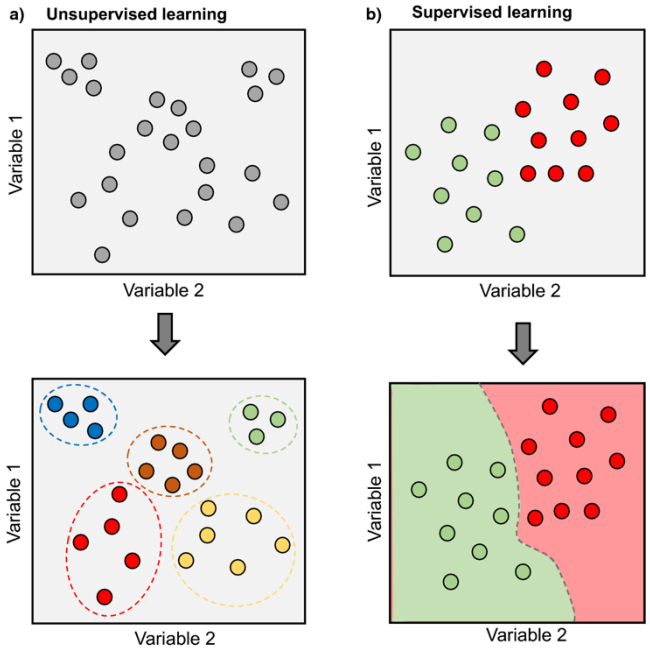
- Examples:
  - Predicting the progression of diseases such as diabetes or cancer based on patient history and medical data
  - Predicting number of cases for COVID-19

What if we don’t have any labels…

Unsupervised Learning!

A model is trained on data without labelled outputs.

- Discovering Patterns or Grouping Similar Data
- Anomaly Detection: Identifying unusual patterns or outliers in the data
- Reducing Dimensionality: Reduce number of features in the dataset while preserving as much information as possible
  - Principal Component Analysis (PCA)

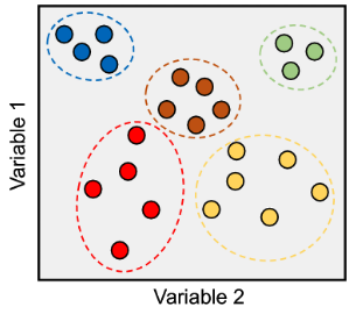


Source: https://evolution-outreach.biomedcentral.com/articles/10.1186/s12052-021-00147-x/figures/3

Unsupervised Learning: Clustering

Clustering algorithms group similar data points into clusters.

Points in the same cluster are more similar to each other than to those in different clusters.

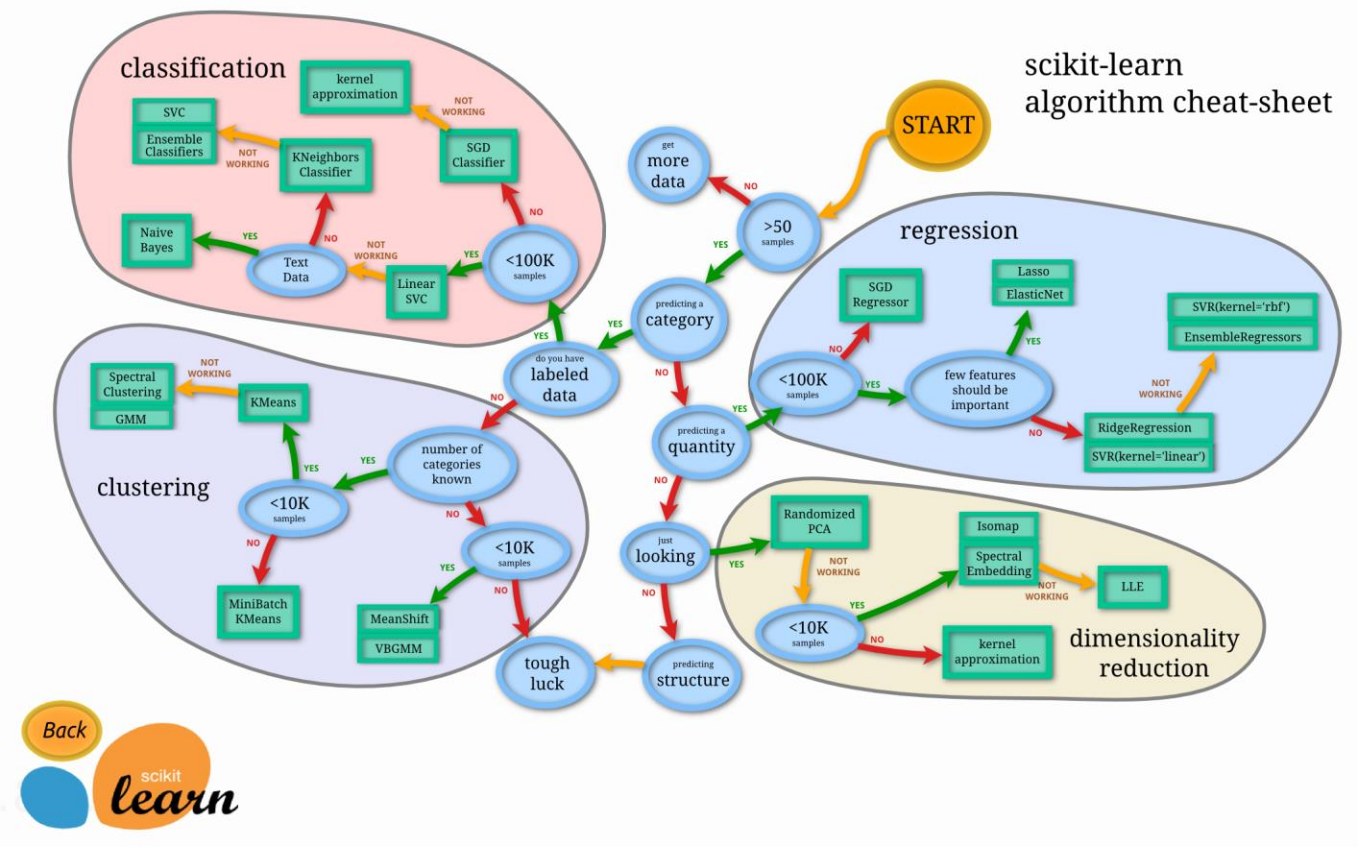


- Cluster Algorithms:
  - K-Means
  - Hierarchical Clustering
  - DBSCAN
  - …
- Examples:
  - Identifying subtypes of complex diseases based on patient data
  - Clustering drugs based on similarity can suggest potential repurposing opportunities
  - Clustering gene expression profiles to identify gene sets with similar expression patterns

Example

- Identifying disease risk using Python and Scikit Learn

How to Choose A Machine Learning Model



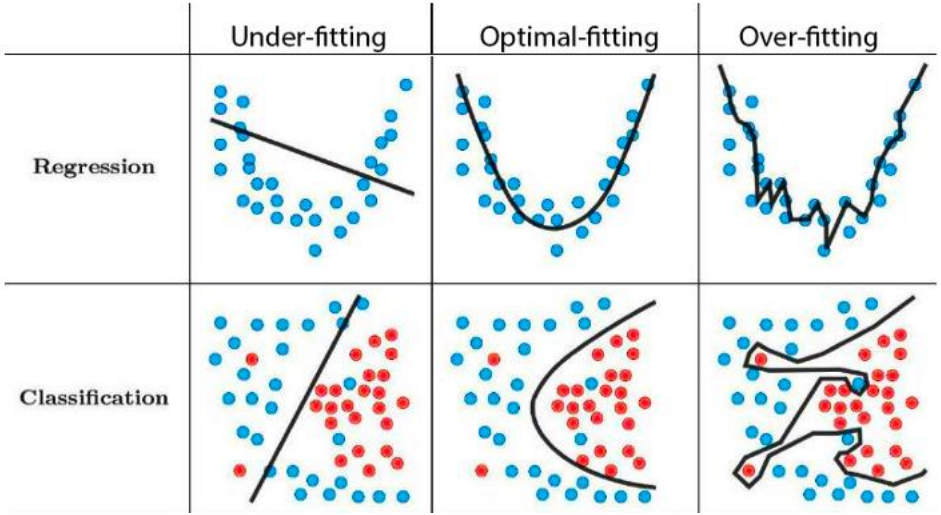
Overfitting and Generalisation

- **Overfitting:**

A model learns all the details, noise, and random fluctuations in the data to an extend that it impacts its performance on new unseen data

- **Generalisation:**

Ability to adapt to new data



100% Accuracy

How to Avoid Overfitting

Good Practice:

1. Separate the data into training set, validation set, and test set
2. Tune the model with training and validation set
3. Evaluate the model with test set

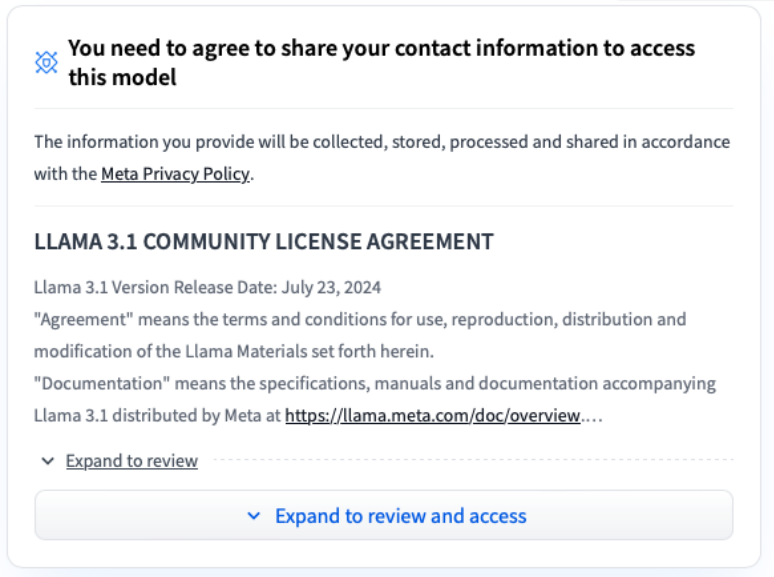
- Use more training data
- Simplify the model
- Cross-Validation

…

Preparation: Apply for Llama Access

- https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

1. Register for Hugging Face
2. Apply for Llama Access
3. Create a Hugging Face tokens (explain in the afternoon)



Reference

- Machine Learning for Everybody: https://youtu.be/i_LwzRVP7bg?si=NwqKp5wqZ3YRwiBQ

- Peter Flach, Machine learning: the art and science of algorithms that make sense of data