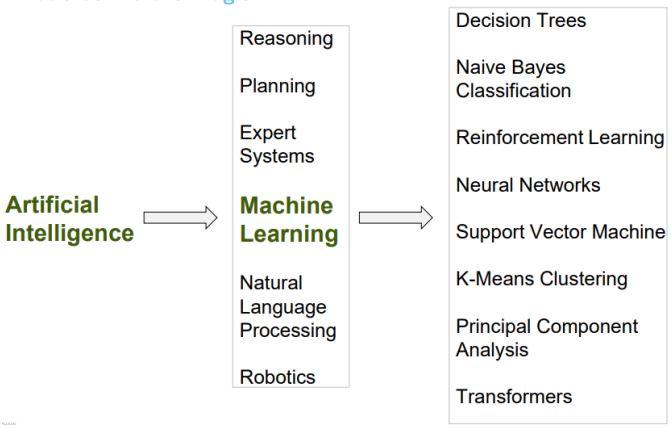
Machine Learning and Data Mining

Jil Zerndt
FS 2025

01 - Intro

What's behind the Magic?



Overview

Data Mining

- Discovering patterns in large data sets
- Extraction of patterns and knowledge from large amounts of data

Machine Learning

- Study of algorithms and statistical models without using explicit instructions, relying on patterns and inference instead
- Branch of AI devoted to developing and understanding methods that "learn", i.e. that leverage fata to make predictions or decisions (act like humans) without being explicity programmed to do so

Model is a logical, mathematical or probabilistic relationship between several variables

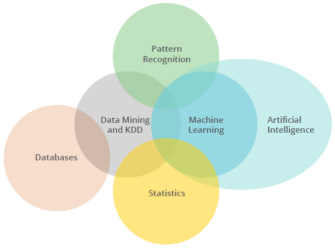
Learning/Training in ML employs adaptive models, which are configured and parameterised based on the training data.

Deep Learning

- Subset of machine learning where artificial neural networks (**Deep Neural Networks**), algorithms inspired by the human brain, learn from large amounts of data
- Uses multiple layers to progressively extract higher level features (attributes) from the raw input

Reinforcement Learning (Trial and Error)

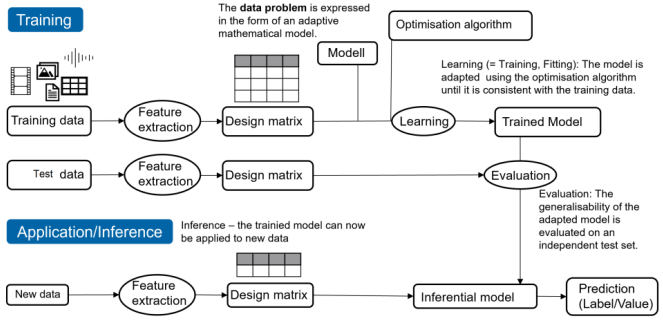
- Concerned with how software agents take actions in an environment in order to maximize some notion of cumulative reward

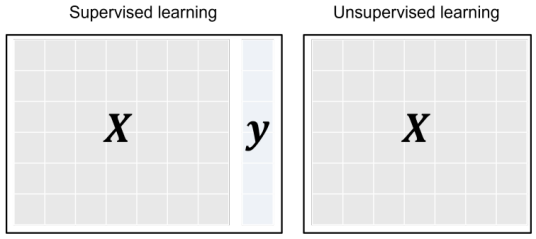


Machine Learning - Types

Supervised Learning

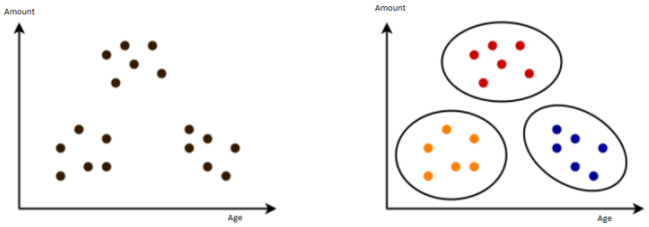
- The algorithm learns from labeled training data, and makes predictions (class, value) on unseen data
- Example: Classification, Regression (see script for math shit)





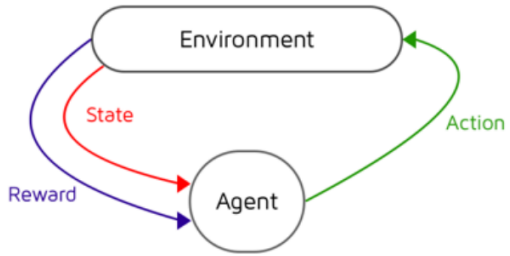
Unsupervised Learning

- The algorithm learns from unlabeled data, and determines data patterns / groupings / clusters
- Example: Clustering, Association, Hierarchical



Reinforcement Learning

- The algorithm learns to perform an action from experience
- Example: Game playing, Robotics



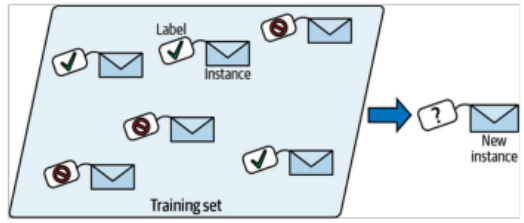
Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups.

Classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. E.g. correctly classifying (assigning a label to) an email as spam or not spam.

Classification

Target variable y: categorical

$$y_m \in \{C_1, C_2, \ldots, C_K\}$$

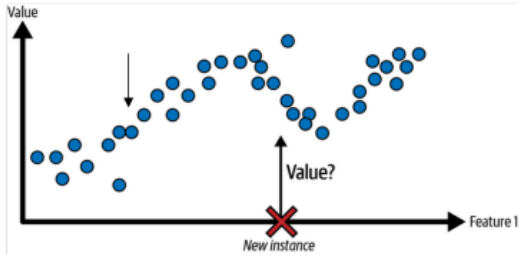


Regression is the problem of predicting and forecasting a concrete number based on a set of data containing observations whose category is known.

Regression

Target variable y: numerical - continuous

$$y_m \in \mathbb{R}$$

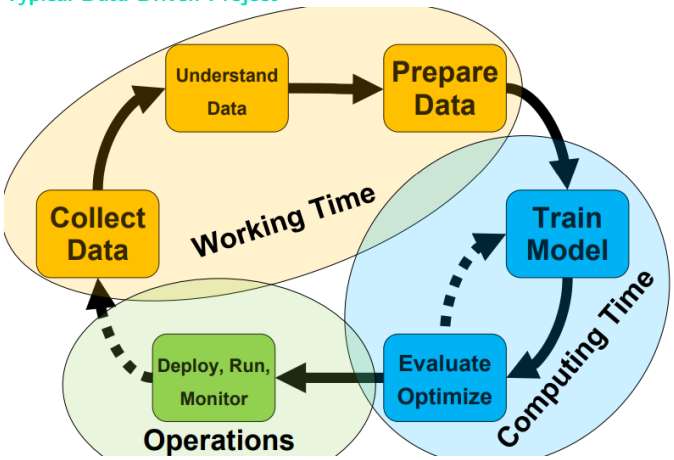


Data Preprocessing

Learning Objectives:
- Understand fundamental importance of data preprocessing
- Know basic algorithms for data cleaning, (near) duplicate detection and filling missing values

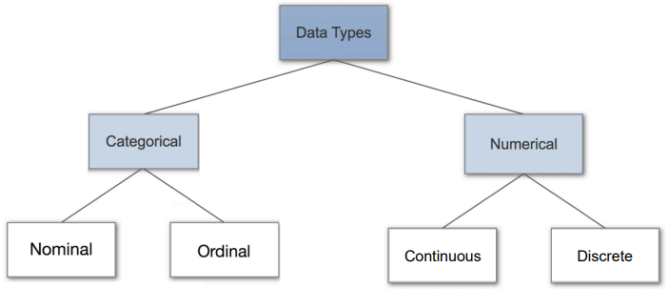
Data

Typical Data Driven Project



Data has many sources, e.g.: sensor, survey, simulation, social media, textual, financial, multimedia, ERP systems data, etc.
Independent of the data source, each data point has a data type

Data Types



Nominal Data
- Nominal scales are used for **labelling** variables, without any quantitative value
- No numerical significance
- Nominal data has no order
- Scales could simply be called labels
- Examples: gender, hair colour, race, marital status

Ordinal Data
- Represents **discrete and ordered** units
- Nearly the same as nominal data, but **order matters**
- No distance between the different categories
- Examples: military rank, star rating, education level

Discrete Numeric Data
- Represents items that can be **counted**
- Values may go from 0, 1, 2, on to infinity (making it countably infinite)
- Examples: number of persons in a room, number of "headsïn 60 coin flips, time elapsed in minutes

Continuous Numeric Data
- Also known as **interval data**
- Often measurements
- Possible values **cannot be counted** and can only be described using intervals on the real number line
- Examples: temperature, weight, height, time, ...

Overview Data Types
Categorial Data
- Nominal: no order, Scale ("labels") → e.g. hair colour, gender
- Ordinal: ordered → e.g. military rank, star rating

Numerical Data (ordered)
- Discrete: countable, ratio → e.g. number of persons in a room
- Continuous: interval, numeric scale → e.g. temperature, weight

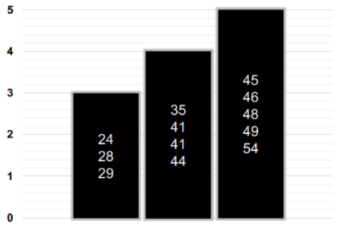
Data Cleaning is the process of improving the data quality by removing or improving incorrect or improperly formatted data.

(near) duplicate detection is the process of identifying and removing or merging duplicate data points.
- compare attributes of the tuple
- compare content of the attributes

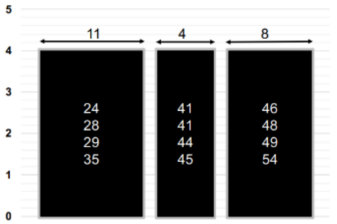
Filling missing values is the process of replacing missing values with substituted values.
- ignore tuple
- fill in missing value manually
- use global constant such as ünknownör 1"
- use attribute mean, median, mode
- use most probable value

Noisy data is data with errors or outliers.
- Binning: divide the range of attribute values into bins
- Regression: smooth data by fitting the data into a function
- Clustering: detect and remove outliers

Binning
Equal Width Binning:
- Divide the range into N intervals of equal size (= width)
- $width = \frac{max-min}{N}$, $bin_i = [min + i \cdot width, min + (i+1) \cdot width]$



Equal Depth/Frequency Binning:
- Divide the range into N intervals with equal number of data points/records (= depth/frequency)
- $depth = \frac{N}{n}$, $bin_i = [data_{(i-1)\cdot depth}, data_{i \cdot depth}]$



Gradient Descent

Polynomial Regression