02450 Introduction to Machine Learning and Data Mining

Week 1: Introduction

Bjørn Sand Jensen

4 February 2025

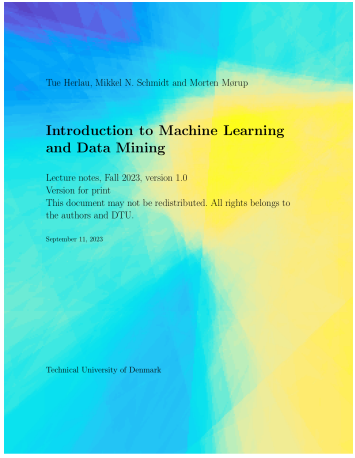
DTU Compute, Technical University of Denmark

Today

Feedback Groups of the day:

Reading material:

Chapter 1, Chapter 2



Lecture Schedule

1. **Introduction**
   4 February: C1,C2

Data: Feature extraction, and visualization

2. Summary statistics, similarity and visualization
   11 February: C4,C7
3. Computational linear algebra and PCA
   18 February: C3
4. Probability and probability densities
   25 February: C5, C6

Supervised learning: Classification and regression

5. Decision trees and linear regression
   4 March: C8, C9 **(Project 1 due 6 March at 17:00)**
6. Overfitting, cross-validation and Nearest Neighbor
   11 March: C10, C12
7. Performance evaluation, Bayes, and Naive Bayes
   18 March: C11, C13
8. Artificial Neural Networks and Bias/Variance
   25 March: C14, C15
9. AUC and ensemble methods
   1 April: C16, C17

Unsupervised learning: Clustering and density estimation

10. K-means and hierarchical clustering
    8 April: C18 **(Project 2 due 10 April at 17:00)**
11. Mixture models and density estimation
    22 April: C19, C20
12. Association mining
    29 April: C21

Recap

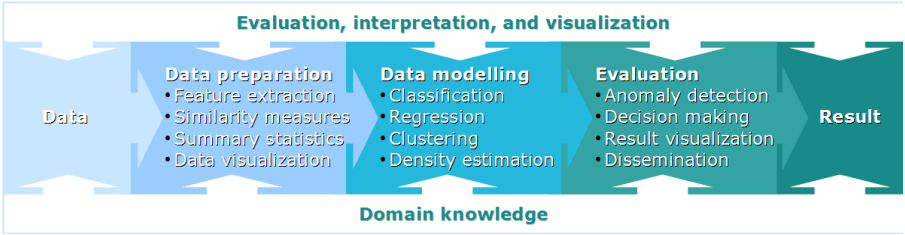
13. Recap and discussion of the exam
    6 May: C1-C21

Online help: Piazza
Videos of lectures: `https://panopto.dtu.dk`
Streaming of lectures: Zoom (link on DTU Learn)

Plan for today:

- Lecture 1 (13:00 – ∼15:00)
  - Introduction
    - What is machine learning?
    - This course and expectations
    - Pre-test
  - Break
  - Overview of the topics
  - Break
  - What is data?
    - Dataset types (tabular, relational, sequential)
    - Attribute types
    - Data quality and issues - missing, noise, corrupted
    - Data in a vector space
- Exercises in your favorite programming language (15:00–17:00)

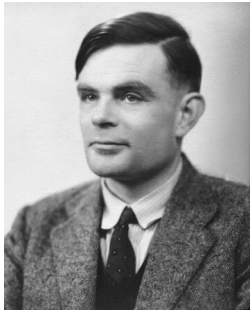
What is machine learning?



Learning Objectives

- Understand what machine learning is and can be used for
- Understand the types of data, their attributes, transformations of attributes and data issues
- Understand the bag-of-words representation as an example of a data transformation

Alan Turing (1946)



Alan Turing
(1912-1954)

- Universal computing
- Proposed machines should learn like children

  We are not in a position to answer if a machine can think because the terms machine and think are undefined. Rather, we should ask if a machine can imitate a human

(the imitation game)

Arthur Samuel (1959)

Arthur Samuel
(1901-1990)

- Samuels wrote a checkers playing program
  - Program played 10000 games against itself
  - Learned value of each board position by considering the resulting score

Machine learning: *(The) field of study that gives computers the ability to learn without being explicitly programmed*

Tom Michell (1999)

A well-posed learning problem: *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at the tasks in T, as measured by P, improves with experience E*

- Checkers example
  - E: Playing 10'000 games
  - T: Playing checkers
  - P: Win or loose

Tom Michell

Quiz 01: Machine learning definition [answer on DTU Learn]


Tom Michell

Recall: A well-posed learning problem is a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at the tasks in T, as measured by P, improves with experience E.

Suppose a program watches as you label images of eyes as containing evidence of cataracts (clouding of the lens) or not, thereby learning to diagnose new examples. What is the experience E?

A The number of correctly diagnosed patients

B A database containing images of eyes with their labels

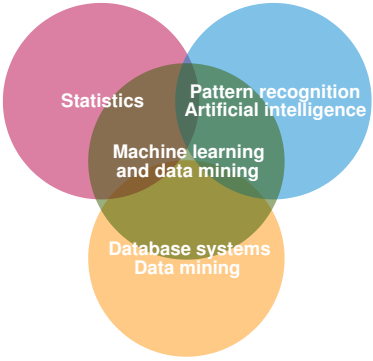
C Diagnosing images of eyes

D Physiological information about cataracts (genetic markers, disease progression, etc.)

E Don't know

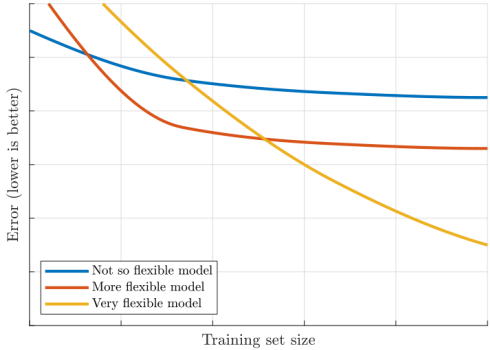
Source: https://www.pexels.com/photo/beautiful-blue-eyes-close-up-dhyamis-kleber-609549/

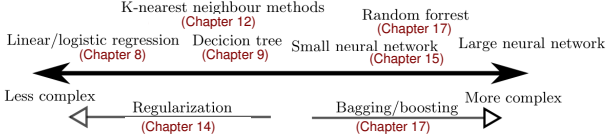
Machine-learning as a separate field



- Focus on a *learning algorithms* (rather than search, pathfinding, etc.)
- De-emphasize explicit knowledge representations, etc.
- Gradual improvements (training time, amount of data)
- *General* algorithms (or algorithmic ideas)

Machine-learning as a separate field





Machine learning advances...

2022-2025, ChatGPT, DeepSeek, Llamma... DeepSeek

2021, alphafold Outperforms all state-of-art expert systems for protein structure prediction.

2020, Breast cancer Outperforms radiologists in breast-cancer detection (Nature)

2019, Lung cancer Outperform six doctors with a 5% reduction in false negatives (DeepMind)

2019, Starcraft 1v1: OpenAI deep reinforcement learning exhibit high-level performance in SC2

2018, BERT: Superhuman performance on the SQuADv1.1 wikipedia question-answer task

2018, alphago: superhuman chess/go learned from scratch

2017, Texas hold'em no limit: Libratus (Carnegie Mellon) beats top professional

2017, Go: Superhuman Go by reinforcement learning + imitation of expert games

2016, libreading : Superhuman libreading from Oxford and Google Deepmind

2016, conversational speech: Microsoft research demonstrate superhuman speech recognition

2016, Geoguessing Google PlaNet win 28 of 50 rounds; median localization error of 1132km vs. 2321km

2015, closed-world image recognition Microsoft report error of 4.94% on ImageNet vs. 5.1% for top-human labeler

2015, Atari Google Deepmind obtain better-than-expert human performance on many Atari video games

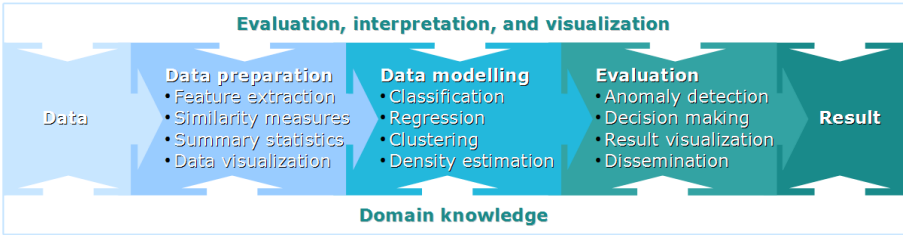
2012, ImageNet and AlexNet

List inspired by: Finn's blog

Why now?

Scientific	Advances in algorithmic ideas
Empirical	Increased availability of large/good datasets
Technological	Faster computers
Social	Libraries which automate routine tasks; increased sharing of code, etc.
Economical	Greatly increased resource allocation

This course (vs other machine learning courses)



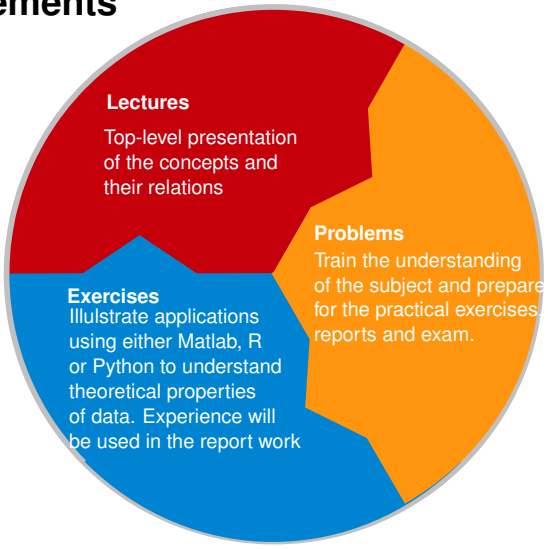
Who should take this course?

Prerequisites

- **Mathematics** (linear algebra, multivariate calculus*, differentiability*, integration*). E.g. 01001/01002/01003/01004/01005
- **Statistics or probability** (**axioms of probability, random variables, expectations, marginalization, product rule, Bayes rule, distributions, densities**, **p-value, paired t-test, confidence intervals**). E.g. 02402/02403 (ideally 02405)
- **Introduction to Programming** in Python / Matlab / R (including basic knowledge of numerical aspects and tools, e.g., NumPy in Python). - E.g. 02002/02101/02102/02525

* Some students will learn about these topics in parallel with this course; extra support has been arranged for those students (e.g. KID students).

Course Elements



Assesment

The assessment consists of two components

- A four-hour written multiple choice exam with negative marking (no aids with two sheets of A4 paper with **handwritten** notes).
  - **Important**: Take notes throughout the course!
- Two mandatory project reports to pass the course (accounts in total for $\approx 10\%$ of the final grade)

  Report 1 March 6 @ 17:00 CET, *Data: Feature extraction, and visualization*

  Report 2 April 10 @ 17:00 CET, *Supervised learning: Classification and regression*
- Final grade based on an **overall assessment** of reports and written exam. The written exam is weighted more than the reports ($\approx 90\%$ vs $\approx 10\%$)

Course format

Lecture session [2 hours per week]

- In-person: main auditorium 116-81 [275 seats] and streamed to 116-83 [100 seats]
- Online: live webcast via Zoom (link on DTU Learn). Exercise rooms can be used during lectures.
- Offline: lectures recorded and uploaded to Panopto (via **Video & Streaming** a few days after the lecture

Exercise session [2 hours per week]

- Physically - find rooms & information on DTU Learn (recommended).
- Online - interaction with TAs using (chat/video/audio) via Microsoft Teams channels.

Detailed structure and activities

- **Workload**: DTU's **nominal** workload for a 5 ECTS course is **9 hours per week** during the 13-week period and 140 hours in total.
- **Structure and study-activities** :
  - Lecture session [2 hours per week]:
    - **High-level lectures** including quizzes and in-class discussions (formative, i.e. not assessed)
  - Supervised **exercises** (formative) [2 hours per week]
    - Focus on the central aspects of the exercises, not the project work.
  - Self-study, preparation & project work [5 hours per week]
    - We assign **readings, homeworks, quizzes** and provide **old exams** to help you study more effectively (not assessed).
    - **Project work** (**assessed**), you apply the taught theory to your own data.
    - Online help/assistance on any aspect of the course via Piazza.
  - **Exam** (**assessed**) [4h + preparation (19h)]

Group Learning

- We encourage group learning during lectures, exercises, and project work.
  - **During lectures** You are encouraged to work together and discuss the online Quiz questions.
  - **During exercises** Each exercise consists of computer exercises relevant to the reported work and a conceptual multiple-choice question from a previous exam. Please only spend about 15 minutes on the multiple-choice question.
  - **Project reports** Submitted as group work (3 people) based on your own dataset.
- For the project reports, you must register your group at DTU Learn > My Course > Groups (target is **three** persons per group).

Online help

- Piazza: https://piazza.com/dtu.dk/spring2025/02450 (Sign up!)
- Use Discussion Forum (i.e. Piazza) for 24/7 help
- Ensures that everyone have access to the same information
  - **Bad: very general questions, i.e., can you explain GMM?**
    Good: Here is what I understand, but I don't get equation ...
  - **Bad: error without context, i.e., I tried to do a PCA, but I got an error that the matrix has the wrong size.**
    Good: What language are you using and what is the error; code which produced the mistake; what did you try to accomplish?
  - **Bad: Can you explain the PCA question in the Fall 2009 exam?**
    Good: Insert a screenshot of fall 2009 exam, explain what part of the solution is unclear

Subsequent machine learning courses

This course (02450) is a **prerequisite** for:

- (02462 Signals and data —currently only for KID students)
- 02456 Deep learning
- 02465 Reinforcement Learning
- 02471 Machine learning for signal processing
- 02477 Bayesian machine learning
- 02460 Advanced Machine Learning

  (other prerequisites 02456 / 02471 / 02477)

+ several domain-specific/applied machine learning courses at DTU

Pretest

- The purpose is to assess your background in order to adjust the presentation and measure your learning.
- Some of the questions will be hard and may seem unfair. Do not Google it. We want to know, if you know.
- **Not part of your assessment. We never look at individual results.**
- Start during the break and finish during the exercise session.

Go to: DTU Learn > 02450 Spring 2025 > Quizzes > Pretest

Break

Next up: *An overview of machine learning*

Machine learning tasks

Supervised - predictive tasks

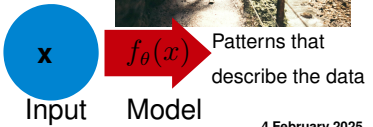
Use some variables to predict unknown or future values of another variable

- **Classification**: *Determine which class a data object belongs to based on the input*
  - Discrete output
- **Regression**: *Determine the output value from the input*
  - Continuous output

Unsupervised - descriptive tasks

Find (human-interpretable) patterns that describe the data

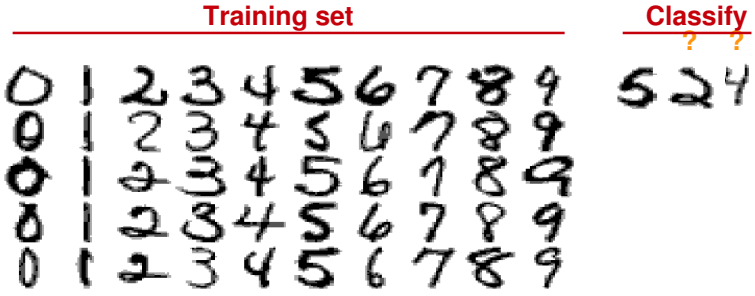
- **Clustering**
  - Discover group structure in data
- **Anomaly Detection** (e.g., via density estimation)
  - Find data objects that are abnormal
- **Association rule discovery**
  - Discover how data objects relate to each other



Classification: Definition

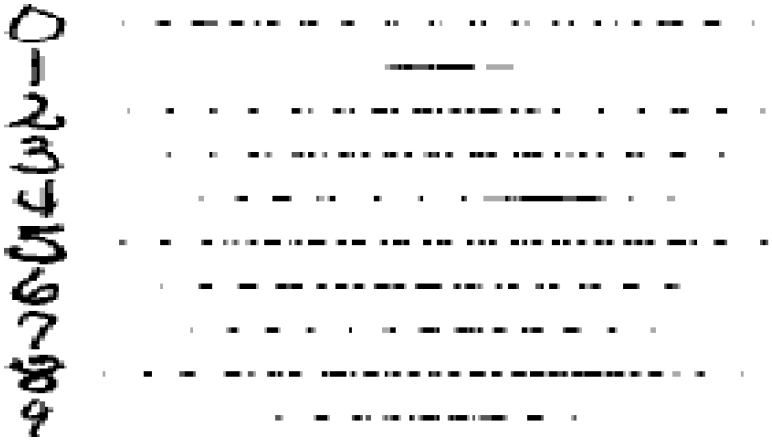
- Given a collection of data objects (**training set**)
  - Each object has associated features (aka variables, attributes)
  - Each object belongs to a certain class
- Define a **model** for the class given the other features
- Goal: Assign a class label to a **previous unseen object**

Example: Image classification



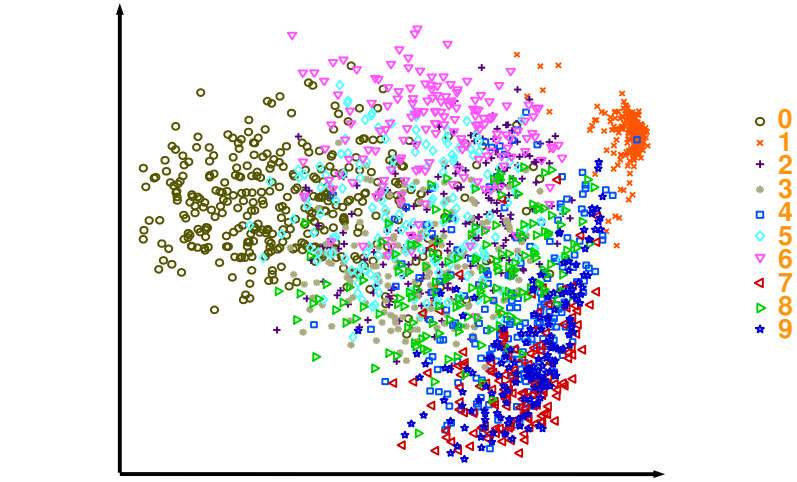
Example: Image classification

Data representation

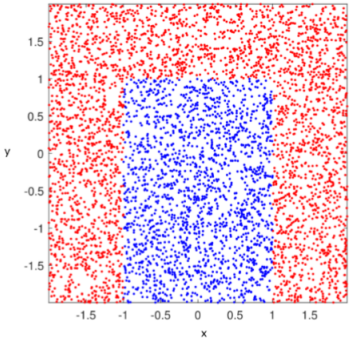


Example: Image classification

Visualization



Quiz 02: Decision rules



The figures shows an example classification problem consisting of a large number of observations (x, y) along with their class (red and blue).

A *decision rule* is just a function which takes an (x, y) coordinate and outputs either the red or the blue class. Suppose we define:

$$z_1 = \max\{0, x - 1\} + \max\{0, -1 - x\}$$
$$z_2 = \max\{0, -1 + y\}$$

Which of the following *decision rules* solve the problem?

A. If $z_1 = z_2 = 0$ classify as blue and otherwise as red

B. If $z_1 = z_2 = 1$ classify as blue and otherwise as red

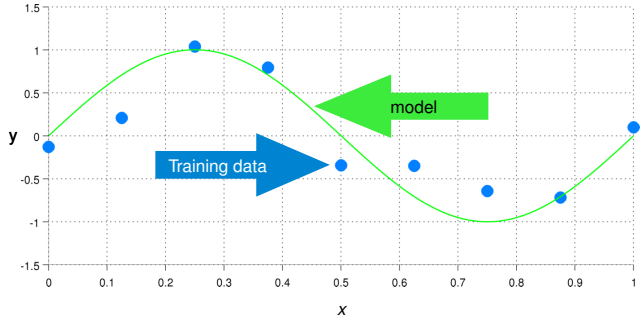
C. If $z_1 = 1$ and $z_2 = 0$ classify as blue and otherwise as red

D. If $z_1 = 0$ and $z_2 = 1$ classify as blue and otherwise as red

E. Don't know

Regression: Definition

- Given a collection of data objects (**training set**)
  - Each object has associated features (aka variables, attributes)
  - Each object is associated with to **countinious valued variable**
- Define a **model** for the variable given the features
- **Goal**: Predict the value of the variable for a **previously unseen input**



Example: Regression

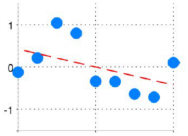
- Predict **sales amounts** of a new product based on **advertising expenditure**.
- Predict **wind velocity** as a function of **temperature**, **humidity**, and **air pressure**
- Predict **stock market index** based on historical **index values** and **market indicators**.

A few possible functions:

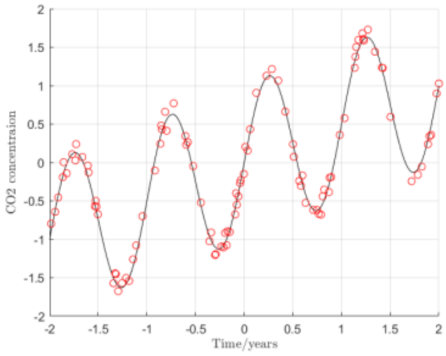
- 1-dimensional input
  $f(x; w) = w_0 + w_1 x$

- 2-dimensional input
  $f(x; w) = w_0 + w_1 x + w_2 x_2$

- 2-dimensional input, polynomial feature transformation/basis expansion
  $f(x; w) = w_0 + w_1 x + w_2 x_2 + w_1 x^2 + w_2 x_2^2$



Quiz 03: Regression



The figures shows an example regression problem where the CO2 concentration is measured as a function of the time of year.

We wish to come up with a prediction rule $y = f(x)$ where y is the relative CO2 concentration and x is the time of year. Which of the following functions would be a good candidate?

A. $y = 0.5x + \cos(x)$

B. $y = -0.5x + \cos(x)$

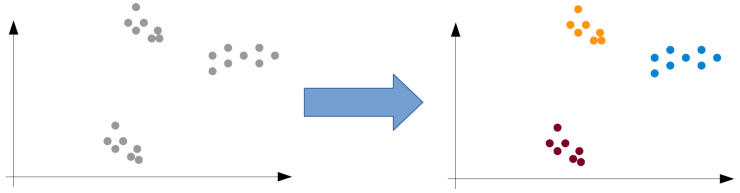
C. $y = 0.5x + \sin(2\pi x)$

D. $y = -0.5x + \sin(2\pi x)$

E. Don't know

Clustering: Definition

- Given a collection of data objects
  - Each object is associated with a number of features
  - A measure of **similarity** between objects is defined
- **Goal**: **Group the objects** into clusters such that
  - Objects within each cluster are similar
  - Objects in separate clusters are less similar



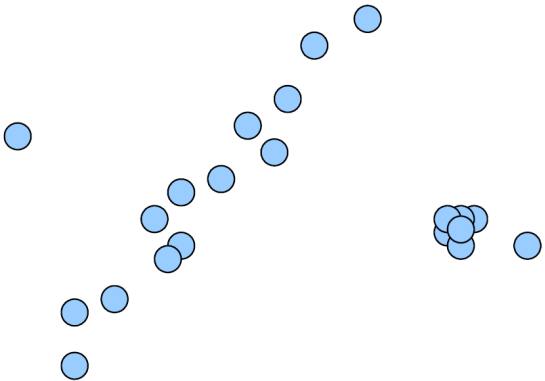
Example: Clustering

Document clustering

- Goal
  - Find groups of similar documents based on the words appearing in them.
- Approach
  - Identify frequently occurring words in each document
  - Define a similarity measure based on the word frequencies
  - Perform clustering to find groups of documents
- Motivation
  - Use the clusters to relate a new document to existing documents
  - Better search algorithms: Return documents that are similar but do not have the exact search keywords

Anomaly detection: Definition

- Given a collection of data objects
  - Each object is associated with a number of features
- **Goal**: Detect which object **deviates** from normal behavior



Example: Anomaly detection

- Credit card **fraud detection**
  - Recognize dubious behavior of credit card transactions based on the transaction history of teh card holder.
- Detect **outliers** in data measurement behavior
  - Remove erroneous measurements due to misreading from an instrument
- **Fault detection** in system health monitoring
  - Detect when a wind turbine performs poorly due to ice coating on blades

Association rule discovery: Definition

- Given a **set of records**
  - Each containing a number of **items from a set**
- Goal: Produce dependency rules
  - Predict the occurrence of an item based on occurrences of other items

Example: Association rule discovery

Market basket analysis

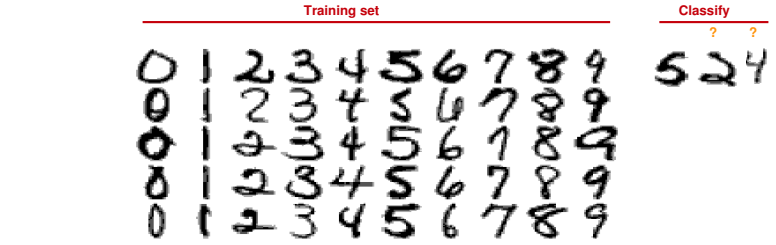
Training set

1.{Bread, Coke, Milk}
2.{Beer, Bread}
3.{Beer, Coke, Diaper, Milk}
4.{Beer, Bread, Diaper, Milk}
5.{Coke, Milk}

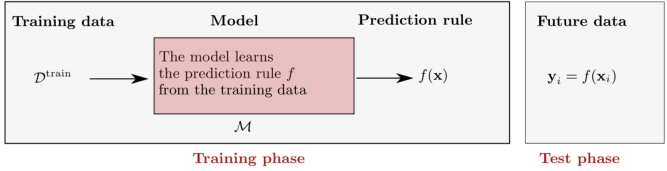
Rules discovered

{Milk} ▸ {Coke}
{Diaper, Milk} ▸ {Beer}

Models in machine learning



Classifying a digit is a function: $f : \mathbb{R}^M \to \{0, 1, \ldots, 9\}$

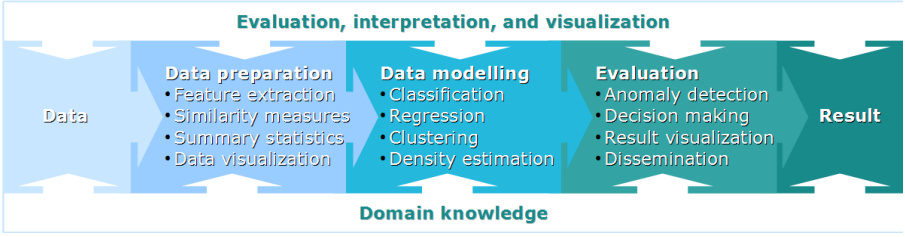


How often the function f makes a mistake on the future/unseen data is called the **generalization error**.

Break

Next up: *What is data?*

What is data?



- Garbage IN, garbage OUT!

What is data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Also known as variable, features, characteristic, or featrues.
- Collection of attributes describe an object
  - Also know as a record, point, case, sample, entity, of instance.

Attributes

Data objects

ID	Age	Gender	Name
1	31	F	Alex
2	24	M	Ben
3	52	F	Cindy
4	35	M	Dan
5	58	M	Eric
6	46	F	Fay
7	42	M	George

Discrete / continuous attributes

- **Discrete**
  - Finite (or countable infinite) set of values
    - Zip codes
    - Counts
    - Set of words in a collection of documents
  - Often represented as integer values
- **Continuous**
  - Has real numbers as attributes values)
  - Examples
    - Temperature
    - Height
    - Weight
  - Often represented as floating point variables

Types of attributes

- **Nominal:** Objects belong to a category (equal / not equal)
  - ID numbers
  - Eye color
  - Zip codes
- **Ordinal:** Object can be ranked (greate than / less than)
  - Taste of potato chips on a scale from 1-10
  - Grades
  - Height in {short, medium, tall}
- **Interval:** Distance between object can be measured (addition / subtraction)
  - Calendar dates
  - Temperature in Fahrenheit and Celcius
- **Ratio:** Zero means absence of what is measured (multiplication / division)
  - Length
  - Time
  - Counts
  - Temperature in Kelvin

Discussion

Classify the following attributes

a) Military rank
b) Angles measured in degrees
c) A person's year of birrth
d) A person's age in years
e) Coat check number
f) Distance from center of campus
g) Number of patiens in a hospital

- **Discrete**
  - Finite (or countably infinite) set of values
- **Continuous**
  - Real number

- **Nominal** (Equal / Not equal)
  - Objects belong to a category
- **Ordinal** (Greater than / Less than)
  - Objects can be ranked
- **Interval** (Addition / Subtraction)
  - Distance between objects can be measured
- **Ratio** (Multiplication / Division)
  - Zero means absence of what is measured

Quiz 04: Attribute types

No.	Attribute description	Abbrev.
x_1	Type (0 = served cold, 1 = served hot)	TYPE
x_2	Calories per serving	CAL
x_3	Grams of protein	PROT
x_4	Grams of fat	FAT
x_5	Milligrams of sodium	SOD
x_6	Grams of dietary fiber	FIB
x_7	Grams of complex carbohydrates	CARB
x_8	Grams of sugars	SUG
x_9	Milligrams of potassium	POT
x_{10}	Vitamins and minerals in 0%, 25%, or 100% of FDA recommendations	VIT
x_{11}	Shelf position (1, 2, or 3, counting from the floor)	SHELF
x_{12}	Weight in ounces of one serving	WEIGHT
x_{13}	Number of cups in one serving	CUPS
x_{14}	Name of cereal brand	NAME
y	Average rating of the cereal (from 0 to 100)	RAT

Table 1: Attributes in a study of cereals (i.e. breakfast products, taken from http://lib.stat.cmu.edu/DASL/Datafiles/Cereals.html).

In a study of healthy breakfast habits 77 cereal brands were investigated. The attributes of the data are given in Table 1. There are a total of 14 attributes denoted x_1–x_{14} and one output variable y which defines the average rating of the cereal products by the consumers.

Which statement about the attributes in the data set is *incorrect*?

A. NAME is discrete and nominal.

B. PROT, FAT and SOD are all continuous and ratio.

C. TYPE and VIT are both discrete and ordinal.

D. An attribute that is ratio will also be interval.

E. Don't know.

Types of data sets

- **Record / tabular**
  - Collection of data objects and their attributes
  - Representation: Table
- **Relational data**
  - Collection of data objects and their relation
  - Representation: Graph
- **Ordered data**
  - Ordered collection of data objects
  - Representation: Sequence

Example: Record example — market basket data

- **Transaction data table**

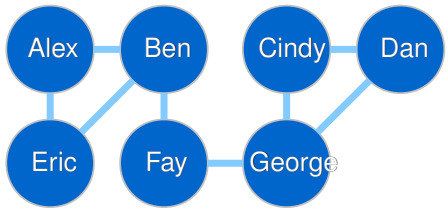
ID	Items
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

- **Matrix**

ID	Bread	Soda	Milk	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	0	1

Example: Relational data — who knows who?

- **Graph**

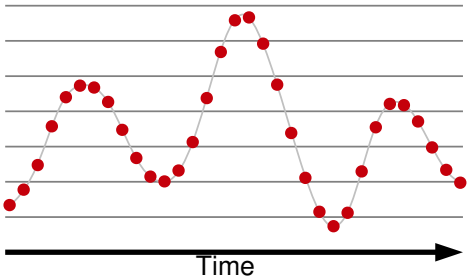


- **Matrix**

	A	B	C	D	E	F	G
A	0	1	0	0	1	0	0
B	1	0	0	0	1	1	0
C	0	0	0	1	0	0	1
D	0	0	1	0	0	0	1
E	1	1	0	0	0	0	0
F	0	1	0	0	0	0	1
G	0	0	1	1	0	1	0

Example: Sequential data / time series

- **Sequence**



- **Matrix**

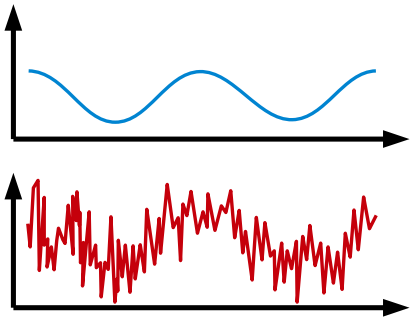
Time	Value
0	1.3
1	1.8
2	2.5
3	3.6
4	4.4
5	4.7
6	4.6
7	4.3
8	2.4
9	2.1
10	2.0
11	2.3
12	3.1

Quality

- **Data is of high quality if they**
  - Are fit for their intended purpose
  - Correctly represent the phenomena they correspond to
- **Examples of quality problems**
  - Noise
  - Outliers
  - Missing values

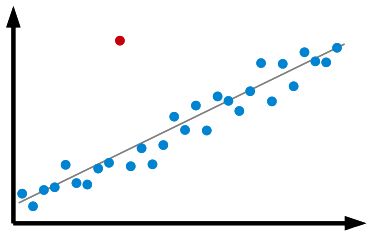
Noise

- **Definition**
  - Unwanted pertubation to a signal
  - Unwanted data
- **Causes of noise**
  - Limits in measurement accuracy
  - Interference from other signals
  - Measurement of attributes note related to the data modellign task
- **Handling noise**
  - Exclude noisy attributes
  - Remoev noise by filtering
  - Model the noise (with explicit assumptions)



Outliers

- **Definition**
  - Data object which are significantly different from most others
  - Unwanted data
- **Causes of outliers**
  - Measurement error
  - Natural property of the data
- **Handling outliers**
  - Identify and explude outliers
  - Model the outliers



Missing values

- **Definition**
  - No value is stored for an attribute in a data object (or a specific value to indicate missing values)
- **Causes of missing values**
  - Information not collected or measures (e.g. people declide to give their age or gender)
  - Attribute is not applicable (e.g. annual income is not relavant for a child)
- **Handling missing values**
  - Remove data objects
  - Eliminate attributes
  - Estimate missing values (e.g. an average)
  - Ignore the missing value in analysis
  - Model the missing value

ID	Age	Gender	Name
1	31	F	Alex
2	(?)	M	Ben
3	52	F	Cindy
4	35	(?)	Dan
5	(?)	M	Eric
6	(?)	F	Fay
7	42	M	(?)

Discussion

- A group of people were asked to write how many children they have - Their response was this

- A reserach assistant typed the results into a table - His table looked like this

Children	3	1	0	2	7	5	15	0	1	3	-2	0	0	0	1

- Are there any data quality issues?
  - Noise?
  - Outliers?
  - Missing values
- Why have these issues occured, and how should they be handled?

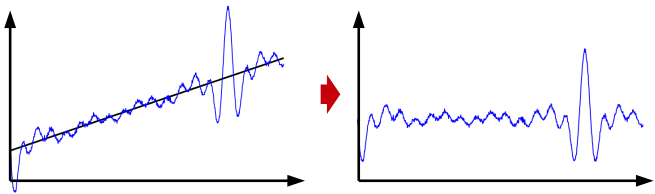
Dataset manipulations

- **Sampling**
  - Selecting a representative subset of the data
- **Feature subset selection**
  - Choose a subset of attributes
- **Feature extraction / transformation**
  - Create new features from existing features
  - Discretization and binarization
  - Apply a fixed transformation to an attribute
  - Aggregation several attributes into a signle attribute
- **Dimensionality reduction**
  - Projet data to a low-dimensional subspace

Feature processing

- Eliminating, suppressing, or attenuating certain aspects of the data
  - Noise removal in audio signals
  - Elimination of common words in text documents
  - Removal of background in images
  - Removal of examples which are corrupted
  - De-trending data (if it is not stationary)

Example of de-trending data



Common feature transformation

ID	MPG	Cylinders	Horsepower	Weight	Year	Safety	Acceleration	Origin
1	18	8	150	3436	70	4	11	France
2	28	4	79	2625	82	4	18.6	USA
3	26	4	79	2255	76	3	17.7	USA
3	29	4	70	1937	76	1	14.2	Germany
4	NaN	8	175	3850	70	2	11	USA
5	24	4	90	2430	70	3	14.5	Germany
6	17.5	6	95	3193	76	4	17.8	USA
7	25	4	87	2672	70	-100	17.5	France
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
142	15	8	198	4341	70	2	10	USA

$$\boldsymbol{X} = \begin{bmatrix} 18 & 8 & 150 & 3436 & 70 & 4 & 11 & 3 \\ 28 & 4 & 79 & 2625 & 82 & 4 & 18.6 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 15 & 8 & 198 & 4341 & 70 & 2 & 10 & 1 \end{bmatrix}$$

Standardize:

$$\boldsymbol{X} = \begin{bmatrix} \cdots & (X_{1j} - \hat{\mu}_j)/\hat{\sigma}_j & \cdots \\ \cdots & (X_{2j} - \hat{\mu}_j)/\hat{\sigma}_j & \cdots \\ & \vdots & \\ \cdots & (X_{Nj} - \hat{\mu}_j)/\hat{\sigma}_j & \cdots \end{bmatrix}$$

$$\hat{\mu}_j = \frac{1}{N}\sum_{i=1}^{N} X_{ij}, \quad \hat{\sigma}_j = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_{ij} - \mu_j)^2}$$

Binarize/threshold:

$$\boldsymbol{X} = \begin{bmatrix} \cdots & 1_{[\theta,\infty[}(x_{1j}) & \cdots \\ \cdots & 1_{[\theta,\infty[}(x_{2j}) & \cdots \\ & \vdots & \\ \cdots & 1_{[\theta,\infty[}(x_{Nj}) & \cdots \end{bmatrix}$$

$$1_{[\theta,\infty[}(x) = 1 \text{ if } x \geq \theta \text{ otherwise } 0$$

Example: One-out-of K encoding

One-out-of-K coding

$X =$

Age	Height	Weight	Nationality
-0.2248	-0.4762	-0.2097	'Sweden'
-0.5890	0.8620	0.6252	'Sweden'
-0.2938	-1.3617	0.1832	'Sweden'
-0.8479	0.4550	-1.0298	'Sweden'
-1.1201	-0.8487	0.9492	'Norway'
2.5260	-0.3349	0.3071	'Norway'
1.6555	0.5528	0.1352	'Norway'
0.3075	1.0391	0.5152	'Norway'
-1.2571	-1.1176	0.2614	'Norway'
-0.8655	1.2607	-0.9415	'Sweden'
-0.1765	0.6601	-0.1623	'Norway'
0.7914	-0.0679	-0.1461	'Denmark'
-1.3320	-0.1952	-0.5320	'Denmark'
-2.3299	-0.2176	1.6821	'Sweden'
-1.4491	-0.3031	-0.8757	'Sweden'
0.3335	0.0230	-0.4838	'Sweden'
0.3914	0.0513	-0.7120	'Denmark'
0.4517	0.8261	-1.1742	'Sweden'
-0.1303	1.5270	-0.1922	'Norway'
0.1837	0.4669	-0.2741	'Denmark'

Age	Height	Weight			
-0.2248	-0.4762	-0.2097	0	0	1
-0.5890	0.8620	0.6252	0	0	1
-0.2938	-1.3617	0.1832	0	0	1
-0.8479	0.4550	-1.0298	0	0	1
-1.1201	-0.8487	0.9492	0	1	0
2.5260	-0.3349	0.3071	0	1	0
1.6555	0.5528	0.1352	0	1	0
0.3075	1.0391	0.5152	0	1	0
-1.2571	-1.1176	0.2614	0	1	0
-0.8655	1.2607	-0.9415	0	0	1
-0.1765	0.6601	-0.1623	0	1	0
0.7914	-0.0679	-0.1461	1	0	0
-1.3320	-0.1952	-0.5320	1	0	0
-2.3299	-0.2176	1.6821	0	0	1
-1.4491	-0.3031	-0.8757	0	0	1
0.3335	0.0230	-0.4838	0	0	1
0.3914	0.0513	-0.7120	1	0	0
0.4517	0.8261	-1.1742	0	0	1
-0.1303	1.5270	-0.1922	0	1	0
0.1837	0.4669	-0.2741	1	0	0

Example: Bag-of-words representation

- First three sentences on wikipeadia.org
  - The bag-of-words model is a simplifying assumption used in natural language processing and information retrieval
  - In this model, a text (such as a sentence or a document) is represented as an unordered collection of words, disregarding grammar and even word order
  - The bag-of-words model is used in some methods of document classification

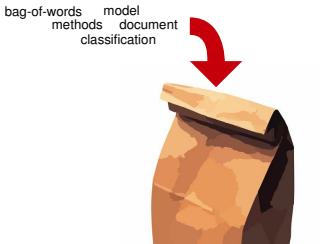
Example: Bag-of-words representation

- The bag-of-words model is a simplifying assumption used in natural language processing and information retrieval
- In this model, a text (such as a sentence or a document) is represented as an unordered collection of words, disregarding grammar and even word order
- The bag-of-words model is used in some methods of document classification

We will treat **this text** as a data set and create a bag-of-words model of it

Example: Bag-of-words representation

- Elimination of common words (so-called stop words)
  - The bag-of-words model is a simplifying assumption used in natural language processing and information retrieval
  - In this model, a text (such as a sentence or a document) is represented as an unordered collection of words, disregarding grammar and even word order
  - The bag-of-words model is used in some methods of document classification



Example: Bag-of-words

- Representation as a matrix

Word	Sentence		
	1	2	3
bag-of-words	1		1
model	1	1	1
simplifying	1		
assumption	1		
natural	1		
language	1		
processing	1		
information	1		
retrieval	1		
text		1	
sentence		1	
document		1	1
represented		1	
unordered		1	
collection		1	
words		1	
disregarding		1	
grammar		1	
word		1	
order		1	
methods			1
classification			1

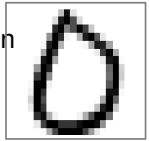
Example: Bag-of-words

- Stemming

Word	Sentence		
	1	2	3
bag-of-word*	1		1
model*	1	1	1
simplif*	1		
assum*	1		
natural*	1		
languag*	1		
process*	1		
information*	1		
retriev*	1		
text*		1	
sentence*		1	
document*		1	1
represent*		1	
unorder*		1	
collect*		1	
word*		2	
disregard*		1	
grammar*		1	
order*		1	
method*			1
classif*			1

Example: Image representation

- **Example: Handwritten digits**
- Preprocessing
  - Digitalization
  - Centering
  - Rotation
  - Scaling



28×28

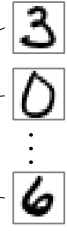
$$M_0 = \begin{bmatrix} 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0.3 & 1 & 0.2 & 0 & \dots & 0 \\ \vdots & & & & & & & \vdots \\ 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

- Vectorization 1×784

$$x_0 = \begin{bmatrix} 0 & \cdots\cdots & 0 & 0.3 & 1 & 0.2 & 0 & \cdots\cdots & 0 \end{bmatrix}^\top$$

- Matrix representation of data set

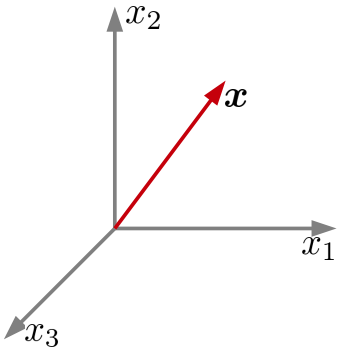
$$\boldsymbol{X} = \begin{bmatrix} \text{---} & \boldsymbol{x}_1 & \text{---} \\ \text{---} & \boldsymbol{x}_2 & \text{---} \\ & \vdots & \\ \text{---} & \boldsymbol{x}_N & \text{---} \end{bmatrix}$$



If each image is 28×28 pixels then $\boldsymbol{X}$ is a $N \times 784$ matrix.

Key concept: Data in a vector space

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix}$$



Summary

Exercises

We support **Matlab**, **Python**, and **R**

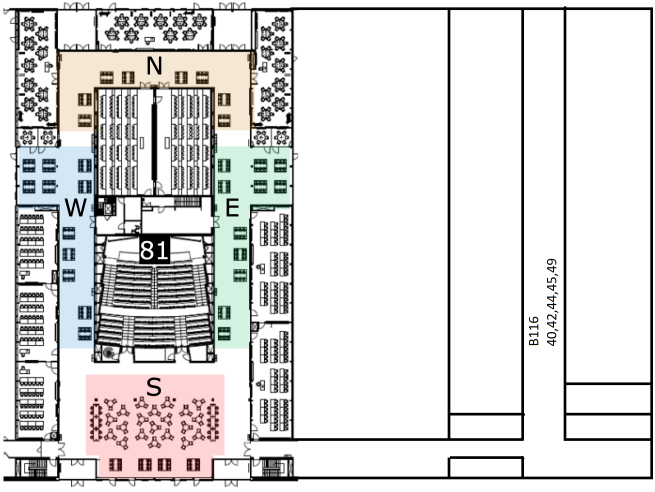
- Exercise 0 guided your through installing your chosen environment
- If you have no experience with any of these programming languages, you should consider if this course is really for you and be willing to learn one (we recommend **Python**)

Exercise rooms

Rooms for exercises:

- Building 116-A081, (Python,Matlab,R)
- Building 116-Lobby South, (Python)
- Building 116-A083, (Python)
- Building 116-H012, (Python)
- Building 116-H013, (Python)
- Building 116-H015, (Python)
- Building 116-Lobby North, (Python)
- Building 116-H016 (reserved for Kunstig Intelligens og Data students), (Python)
- Building 116-H019 (reserved for Kunstig Intelligens og Data students, (Python)

Building 116



Exercises Today

- Follow exercise instructions available on DTU learn (Exercise 1)
- Start forming groups (target is 3 students per group)
  - Find team members via the exercise session, Discussion Forum (i.e. Piazza), or other channels.
  - Unable to find a group (say by end of week 2)? Enter your info in MS Teams $>$ General $>$ Shared files $>$ `02450missing_a_group.xlxs`. We will then attempt to assign you to a group (pelase be proactive an contact other people on the list).
  - Once formed, register your group on DTU Learn.
- Start looking for a dataset and discuss the suitability of a dataset with a TA.
  - Instructions for finding a dataset on DTU Learn $>$ 02450 $>$ Project descriptions $>$ `02450finding_a_dataset_for_reports.pdf` and the project 1 description.