



1

02450: Introduction to Machine Learning and Data Mining

Introduction

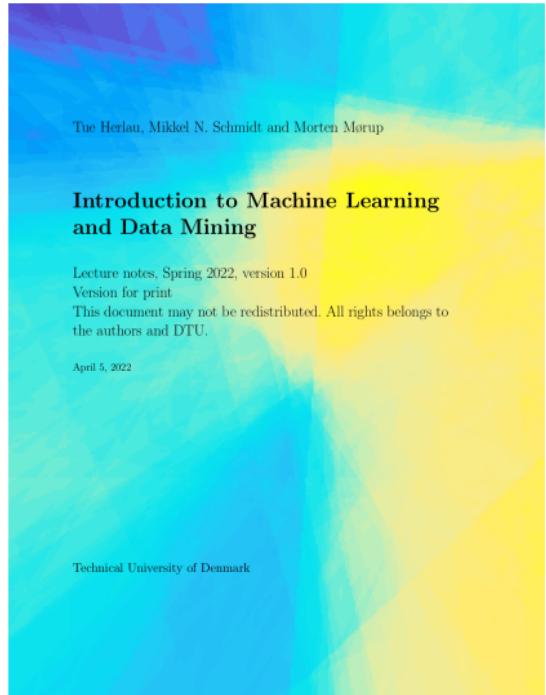
Bjørn Sand Jensen

DTU Compute, Technical University of Denmark (DTU)

DTU Compute

Department of Applied Mathematics and Computer Science

Reading material: Chapter 1



Tue Herlau, Mikkel N. Schmidt and Morten Mørup

Introduction to Machine Learning and Data Mining

Lecture notes, Spring 2022, version 1.0

Version for print

This document may not be redistributed. All rights belongs to
the authors and DTU.

April 5, 2022

Technical University of Denmark

Lecture Schedule

1 Introduction

30 January: C1

Data: Feature extraction, and visualization

2 Data, feature extraction and PCA

6 February: C2, C3

3 Measures of Similarity, summary statistics and probabilities

13 February: C4, C5

4 Probability densities and data visualization

20 February: C6, C7

Supervised learning: Classification and regression

5 Decision trees and linear regression

27 February: C8, C9 (Project 1 due 29 February at 17:00)

6 Overfitting, cross-validation and Nearest Neighbor

5 March: C10, C12

7 Performance evaluation, Bayes, and Naive Bayes

12 March: C11, C13

Online 24/7 help: Discussion Forum/Piazza
Streaming: Zoom (see link on DTU Learn)
Recordings: <https://panopto.dtu.dk/>
Online exercises: MS Teams

8 Artificial Neural Networks and Bias/Variance

19 March: C14, C15

9 AUC and ensemble methods

2 April: C16, C17

Unsupervised learning: Clustering and density estimation

10 K-means and hierarchical clustering

9 April: C18 (Project 2 due 11 April at 17:00)

11 Mixture models and density estimation

16 April: C19, C20

12 Association mining

23 April: C21

Recap

13 Recap and discussion of the exam

30 April: C1-C21

Public homepage

- Syllabus/homework also available on public homepage
<http://compute.dtu.dk/courses/02450>
and on DTU Learn ([02450syllabus_and_practicals.pdf](#))

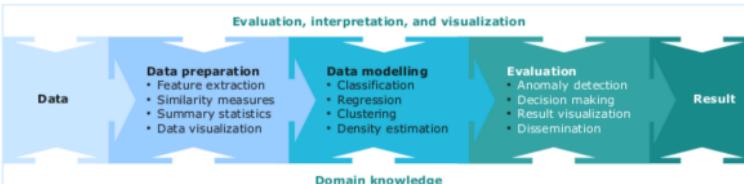
   www2.compute.dtu.dk/courses/02450/

 Section for Cognitive Systems
DTU Compute

02450 Introduction to Machine Learning and Data Mining - Spring 2024

Machine learning and data mining

The course is designed around a data modeling framework shown in the figure. Each lecture/assignment will focus a subset of the data modeling framework.



We emphasize the holistic view of modeling in order to motivate and stress the relevance of individual components and building blocks, disseminate the obtained competence (see the course [learning objectives](#)), and make them applicable for a broad spectrum of engineering problems in e.g. biomedical engineering, chemistry, electrical engineering, and informatics.

Resources

DTU Learn

If you are enrolled in the course you can access material and participate in the course through the [DTU Learn homepage](#).

Lectures

The lectures will take place in Building 116 auditorium 81 on Tuesdays from 13:00-15:00.

Due to restrictions on the auditorium capacity, we will stream the lecture to Building 116 auditorium 83. Seats in Building 116 auditorium 81 and Building 116 auditorium 83 will be allocated on a first come, first served principle. You can use the rooms allocated to exercises (except H013 and H015) to stream the lecture yourself (feel free to use the projectors).



Bjørn Sand Jensen



Georgios Arvanitidis

Plan for today:

- What is machine learning?
- Why do we learn different methods?
- Impact of machine learning
- This course
- Pre-test
- Break
- Lecture 1, basic terminology
- Exercises in your favourite programming language (15:00–17:00)

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

Please answer this quiz on DTU Learn:

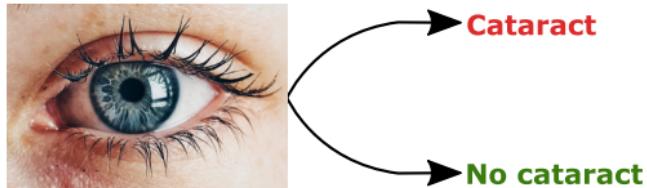
<https://learn.inside.dtu.dk/d2l/home/187740>

-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 tasks in T, as measured by P, improves with experience E."



Tom Michell

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 ?



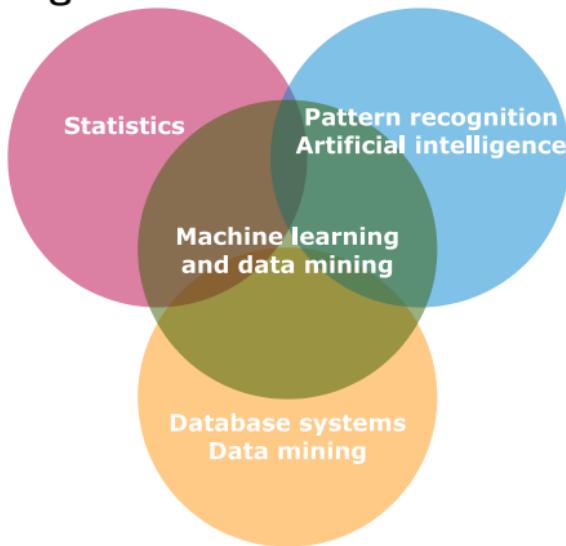
- A. The number of correctly diagnosed patients
- B. A database containing images of eyes with their labels
- C. The errors the program commits when trying to label the images.
- D. Physiological information about cataracts (genetic markers, disease progression, etc.)
- E. Don't know.

Solution:

The correct answer is 2, the images of eyes along their diagnosis. In a normal machine-learning setting, the system would learn a mapping from images to

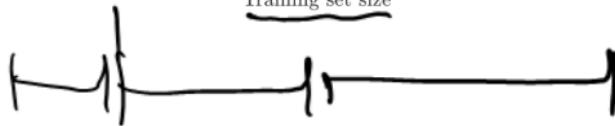
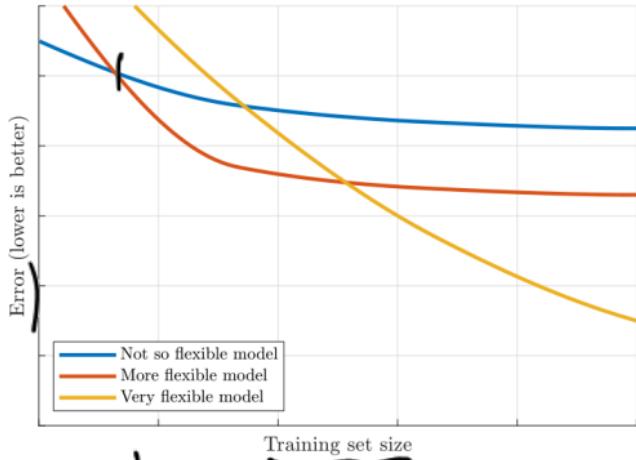
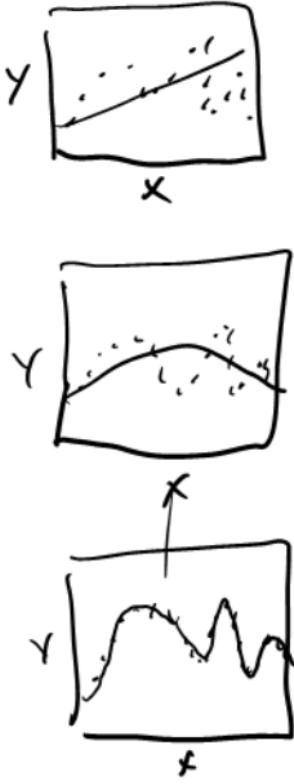
disease labels based on a large set of examples of this mapping.

Machine-learning as a research area

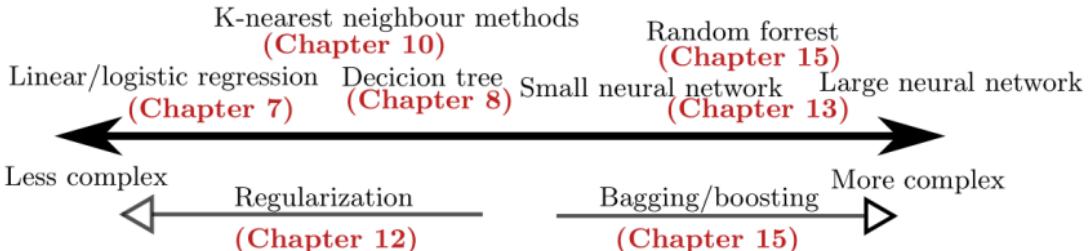
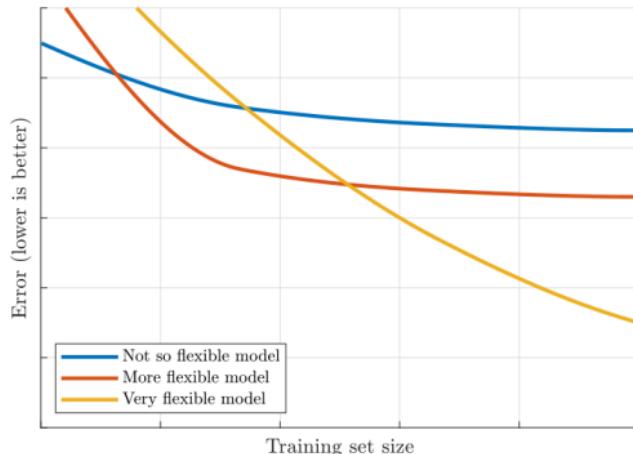


- Focus on a *learning algorithms* (rather than search, pathfinding, etc.)
- De-emphasize explicit knowledge representations, etc.
- Gradual improvements (training time, amount of data)
- *General* algorithms (or algorithmic ideas)

Machine-learning as a research area



Machine-learning as a research area



Man vs. Machine

2022, ChatGPT

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 / Nature Medicine)

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

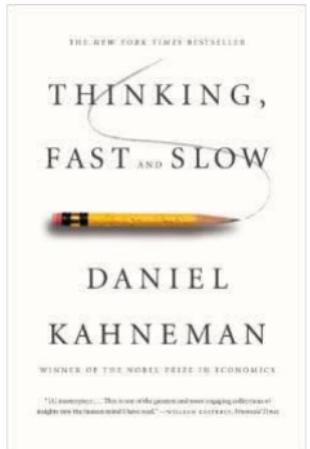
2015, closed-world image recognition Microsoft report error of 4.94% on ImageNet vs. 5.1% for top-human labeler

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

List inspired by:

<https://finnaarupnielsen.wordpress.com/2015/03/15/status-on-human-vs-machines/>

Human abilities examined



The number of studies reporting comparisons of clinical and statistical predictions has increased to roughly two hundred (...) About 60% of the studies have shown significantly better accuracy for the algorithms. The other comparisons scored a draw in accuracy, but a tie is tantamount to a win for the statistical rules, which are normally much less expensive to use than expert judgement. No exception has been convincingly documented.

The range of predicted outcomes has expanded to cover medical variables such as longevity of cancer patients, the length of hospital stays, the diagnosis of cardiac disease, and the susceptibility of babies to sudden infant death syndrome; economic measures such as the prospect of success for new businesses, the evaluation of credit risks by banks, and the future career satisfaction of workers; questions of interest to government agencies, including assessment of the suitability of foster parents, the odds of recidivism among juvenile offenders, and the likelihood of other forms of violent behaviour; and the miscellaneous outcomes such as the evaluation of scientific presentations, the winners of football games, and the future prices of Bordeaux wine. Each of these domains entails a significant degree of uncertainty and unpredictability. We describe them as “low-validity environments.”. In every case, the accuracy of experts was matched or exceeded by a simple algorithm. (Kahneman 2011)

Why now?

Scientific Advances in algorithmic ideas

Empirical Increased availability of large/good datasets

Technological Faster computers

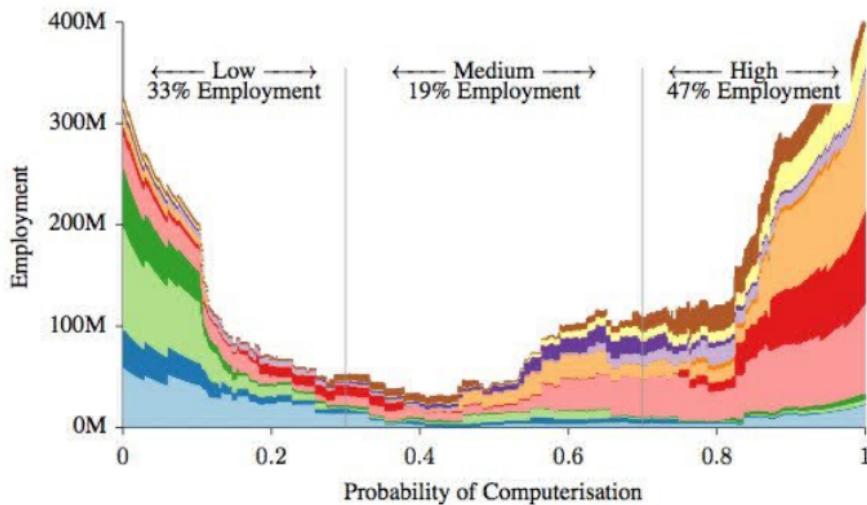
Social Libraries which automate routine tasks; increased sharing of code, etc.

Economical Greatly increased resource allocation

Machine learning as a disruptive technology

- A recent Oxford study suggest about 47% of all US jobs could be automated within two decades (Frey & Osborne, 2013)

| |
|--|
| Management, Business, and Financial |
| Computer, Engineering, and Science |
| Education, Legal, Community Service, Arts, and Media |
| Healthcare Practitioners and Technical |
| Service |
| Sales and Related |
| Office and Administrative Support |
| Farming, Fishing, and Forestry |
| Construction and Extraction |
| Installation, Maintenance, and Repair |
| Production |
| Transportation and Material Moving |



Economical impact I

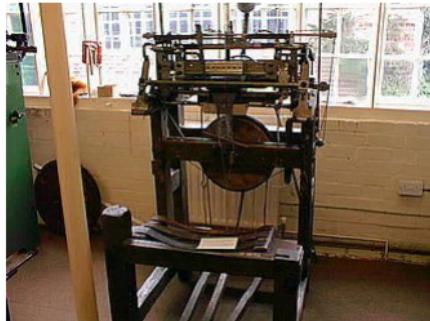
- Basic economics: When things become cheap, we will use it in more places
- Example: Microprocessors perform computations
- Microprocessors did not change the world because people did a lot of computations in 1950, but because **nearly everything can at least partially be turned into a computation problem** (bookkeeping, telephony, photography, entertainment, navigation, design, education, economics, science, etc.)
- We should not ask what situations **are as of now** a machine-learning problem, but which **can be turned into one**

Economical impact II

Many jobs match this description

- ① Recognize what situation you are in
- ② Collect relevant data
- ③ Given data about situation make a **prediction** such as: (i) outcome of performing a given action in the situation or (ii) which action is appropriate
- ④ Perform action
- ⑤ Repeat ...

Machine learning can, in principle, learn 3



[https://en.wikipedia.org/wiki/William_Lee_\(inventor\)](https://en.wikipedia.org/wiki/William_Lee_(inventor))

William Lee (1563–1614) was an English clergyman and inventor who devised the first stocking frame knitting machine in 1589

Elizabeth I: "Thou aimest high, Master Lee. Consider thou what the invention could do to my poor subjects. It would assuredly bring to them ruin by depriving them of employment, thus making them beggars."

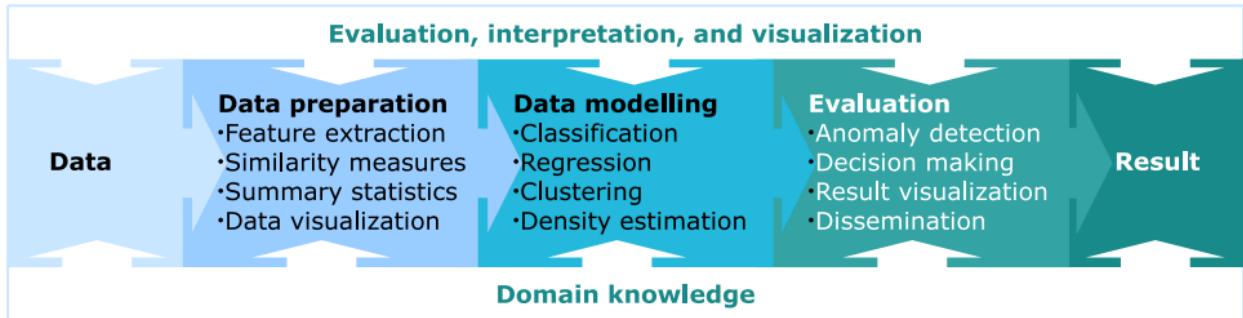
"our discovery of means of economising the use of labour can outrun the pace at which we can find new uses for labour, as Keynes (1933) pointed out.

The reason why human labour has prevailed relates to its ability to adopt and acquire new skills by means of education (Goldin and Katz, 2009). Yet as computerisation enters more cognitive domains this will become increasingly challenging (Brynjolfsson and McAfee, 2011)." (Taken from Frey & Osborne, 2013)

Economical impact III

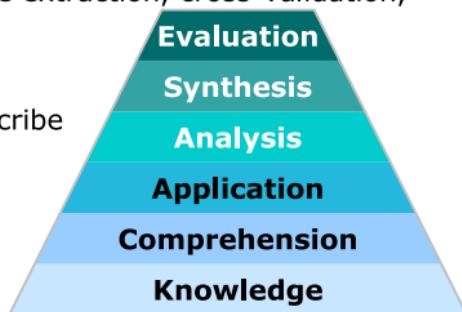
- We don't know what will happen
- Plausibly, many tasks will become so cheap humans will no longer perform them
 - Essential, non-automated tasks will both become more valuable, inhibit progress
 - Humans will do fewer jobs, play a relatively smaller role in the economy (the share of capital will increase relative to labor)
 - The most **destructive** forms of automation is when tasks are only **slightly** better done by machines

Machine learning and data mining pipeline of this course



Learning objectives

1. Describe the major steps involved in data modeling from preparing the data, modeling the data to evaluating and disseminating the results.
(Knowledge)
2. Discuss key machine learning concepts such as feature extraction, cross-validation, generalization and over-fitting, prediction and curse of dimensionality.
(Comprehension)
3. Sketch how the data modeling methods work and describe their assumptions and limitations.
(Knowledge and Comprehension)
4. Match practical problems to standard data modeling problems such as regression, classification, density estimation, clustering and association mining.
(Comprehension and Application)
5. Apply the data modeling framework to a broad range of application domains in medical engineering, bio-informatics, chemistry, electrical engineering and computer science.
(Application)
6. Compute the results of the data modeling framework by use of Matlab, R or Python.
(Application)
7. Use visualization techniques and statistics to evaluate model performance, identify patterns and data issues.
(Analysis)
8. Combine and modify data modeling tools in order to analyze a data set of their own and disseminate the results of the analysis.
(Application, Analysis, Synthesis and Evaluation)



Assesment

The assesment consists of two components

- A four hour written multiple choice exam with negative marking (all aids, closed internet).
- Two mandatory project reports which **must be approved/passed to attend the exam!**

Report 1: *Data, Feature extraction, and visualization*

Report 1 - Main Thursday February 29 at 17:00 ~~March 14~~

Report 1 - Feedback available Thursday ~~April 11~~ at 17:00

Report 1 - Resubmission Thursday March 21 at 17:00 (only if the report was not approved)

Report 2: *Supervised learning: Classification and regression*

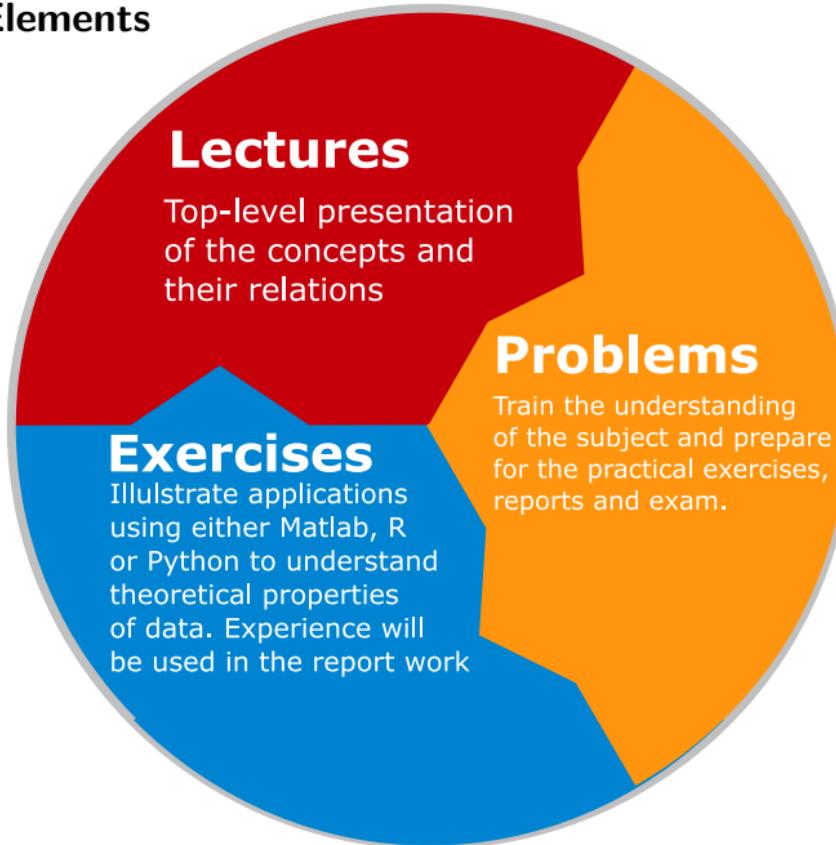
Report 2 - Main Thursday April 11 at 17:00

Report 2 - Feedback available Thursday April ~~11~~²⁵ at 17:00

Report 2 - Resubmission Thursday May 2 at 17:00 (only if the report was not approved)

- Final grade based on overall assessment of reports and written exam. The written exam is weighted more than the reports ($\approx 90\%$ vs $\approx 10\%$)

Course Elements



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 (except 116-13 and 116-15) can be used during lecture.
- 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.
 - TAs will go over some aspect of the exercise/homework around 16:45 (starting in week 2).
- 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 for the report work as well as 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 have to register your group at DTU Learn > My Course > Groups (target is 3 persons per group).

Online help

- <https://piazza.com/dtu.dk/spring2024/02450> (Sign up!)
- Use Discussion Forum (i.e. Piazza) for 24/7 help
- Ensure 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 get an error 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 screenshot of fall 2009 exam, explain what part of the solution is unclear

Subsequent ML courses

This course (02450) is recommended prerequisite for:

- 02456 Deep learning
- 02471 Machine learning for signal processing
- 02477 Bayesian machine learning
- 02460 Advanced Machine Learning (also recommended prerequisites 02456 / 02471 / 02477)

More courses will probably be added soon.

- 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 > Quizzes > Pretest

Data Mining and Machine Learning Tasks

Predictive tasks (Supervised learning)

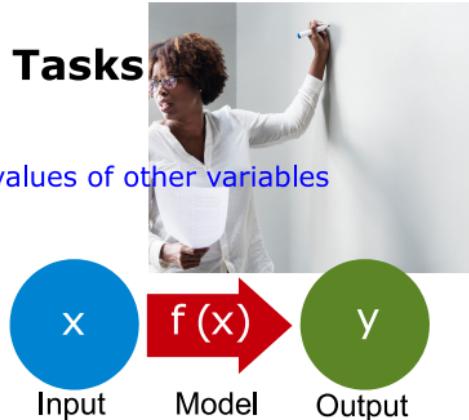
- Use some variables to predict unknown or future values of other variables

- **Classification**

- Discrete output
(Determine which class a new data object belongs to)

- **Regression**

- Continuous output
(Determine the output value from the input variables)



Descriptive tasks (Unsupervised learning)

- Find human-interpretable patterns that describe the data

- **Clustering**

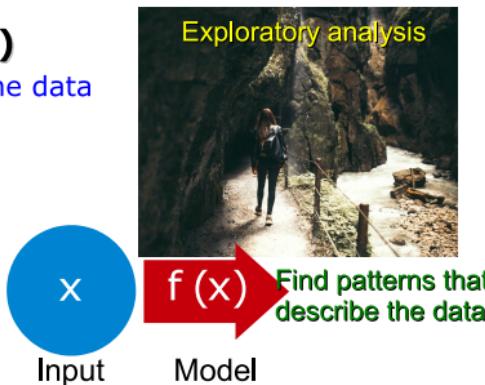
- Discover group structure in data

- **Association rule discovery**

- Discover how data objects relate to each other

- **Anomaly detection**

- Find data objects that are abnormal



Classification: Definition

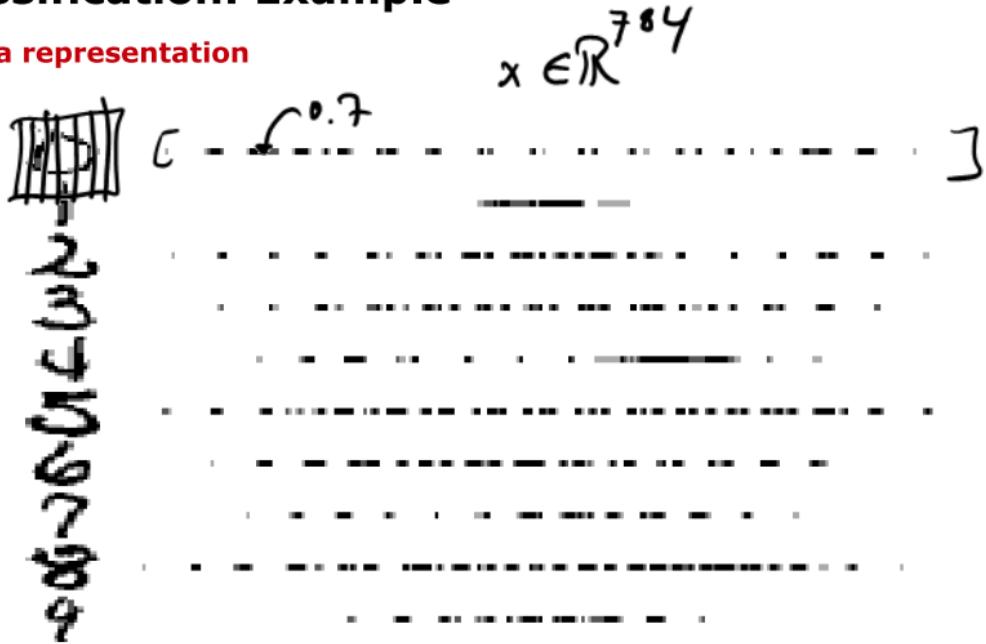
- Given a collection of data objects (**training set**)
 - Each object has associated a number of features, \mathbf{x}
 - Each object belongs to a certain class, y
- Define a **model** for the class given the other features
- Goal: Assign a class label to a **previously unseen object**

Classification: Example

| Training set | | | | | | | | | | Classify | ? | ? | ? |
|--------------|---|---|---|---|---|---|---|---|---|----------|---|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ? | ? | ? | |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 5 | 2 | 4 | |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 5 | 2 | 4 | |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 5 | 2 | 4 | |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 5 | 2 | 4 | |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 5 | 2 | 4 | |

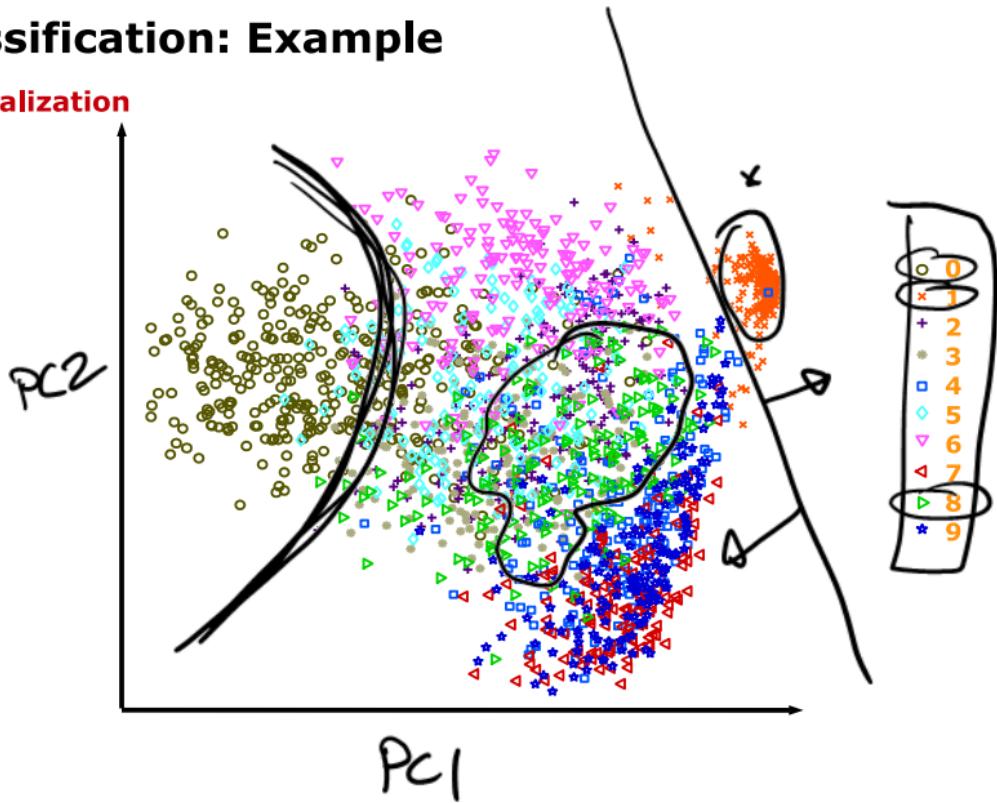
Classification: Example

Data representation



Classification: Example

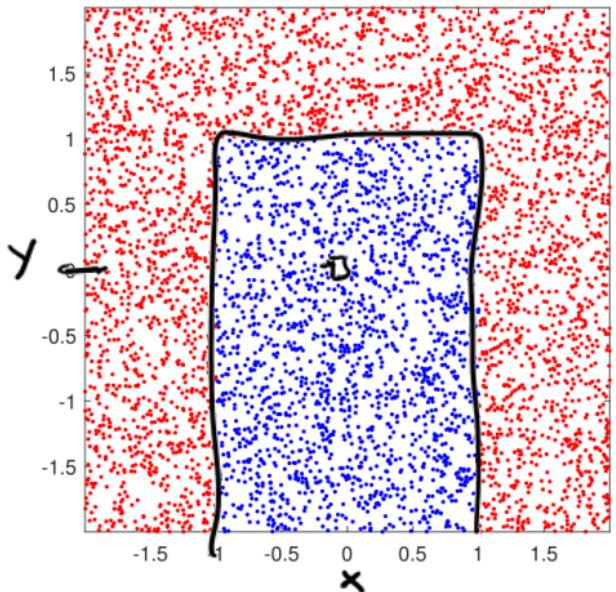
Visualization



$$\max\{-10, 1\} = 1$$

$x: (x, y)$

Quiz 02 (DTU Learn): Decision rules



The figure 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:

$$\begin{aligned} z_1 &= \max\{0, x - 1\} + \max\{0, -1 - x\} \geq 0 \\ z_2 &= \max\{0, -1 + y\} \geq 0 \end{aligned} \quad \left. \right\}$$

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

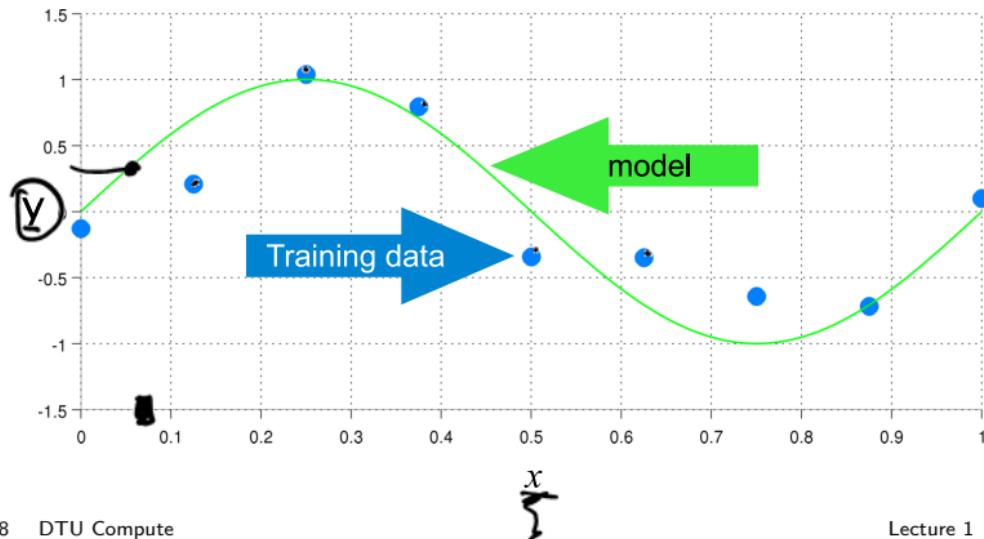
Solution:

The first option will give the correct boundary. To see this, check out $(x, y) = 0$. In this case $z_1 = 0$ and $z_2 = 0$. Therefore, only the first option gives the correct answer.

Actually, this decision rule is a (slightly obscured) example of a neural network with RELU units. We will learn more about neural networks later in this course.

Regression: Definition

- Given a collection of data objects
 - Each object has associated a number of features
 - Each object has associated a **continuous valued variable**
- Define a **model** for the variable given the features
- Goal: Predict the value of the variable for a **previously unseen object**



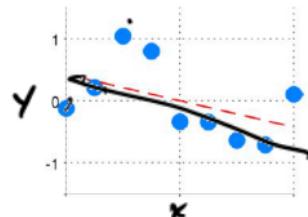
Regression: Example

- Predict **sales amounts** of new product based on
 - advertising expenditure
- Predict **wind velocity** as a function of
 - temperature, humidity, and air pressure
- Predict the value of a **stock market index** based on
 - previous index time series and market indicators

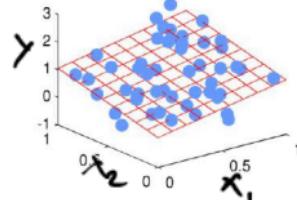
Regression: Example

- Predict **sales amounts** of new product based on
 - advertising expenditure
- Predict **wind velocity** as a function of
 - temperature, humidity, and air pressure
- Predict the value of a **stock market index** based on
 - previous index time series and market indicators

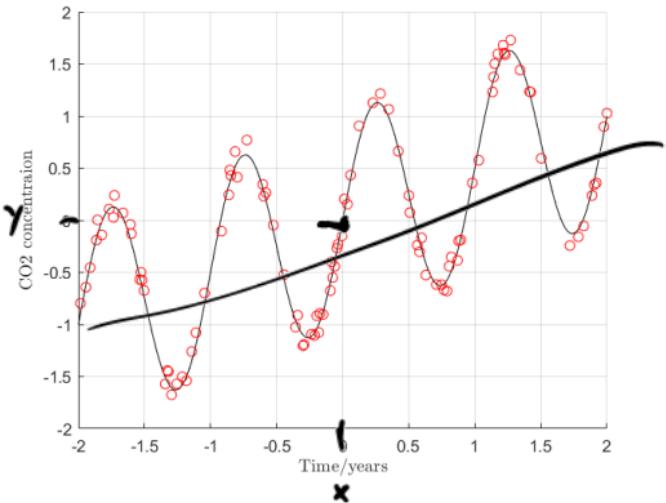
1-dimensional inputs
 $f(x) = w_0 + w_1 x$



2-dimensional inputs
 $f(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2$



Quiz 03 (DTU Learn): 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

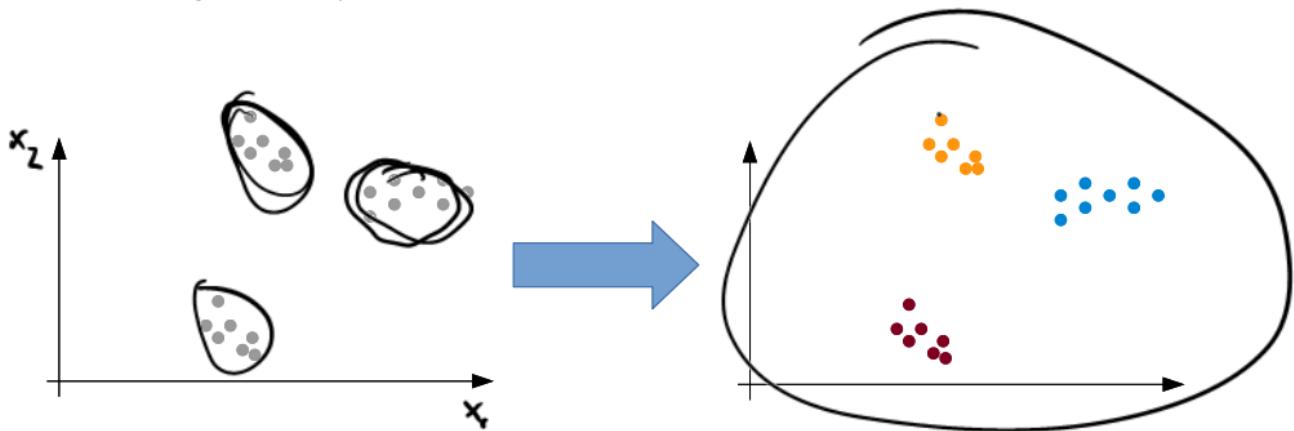
Solution:

The right answer is $y = 0.5x + \sin(2\pi x)$. The functions involving cosines can be ruled out since they have the wrong period; meanwhile the function has an

upwards trend which means it must involve the term $0.5x$.

Clustering: Definition

- Given a collection of data objects
 - Each object has associated 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



Clustering: Example

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

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

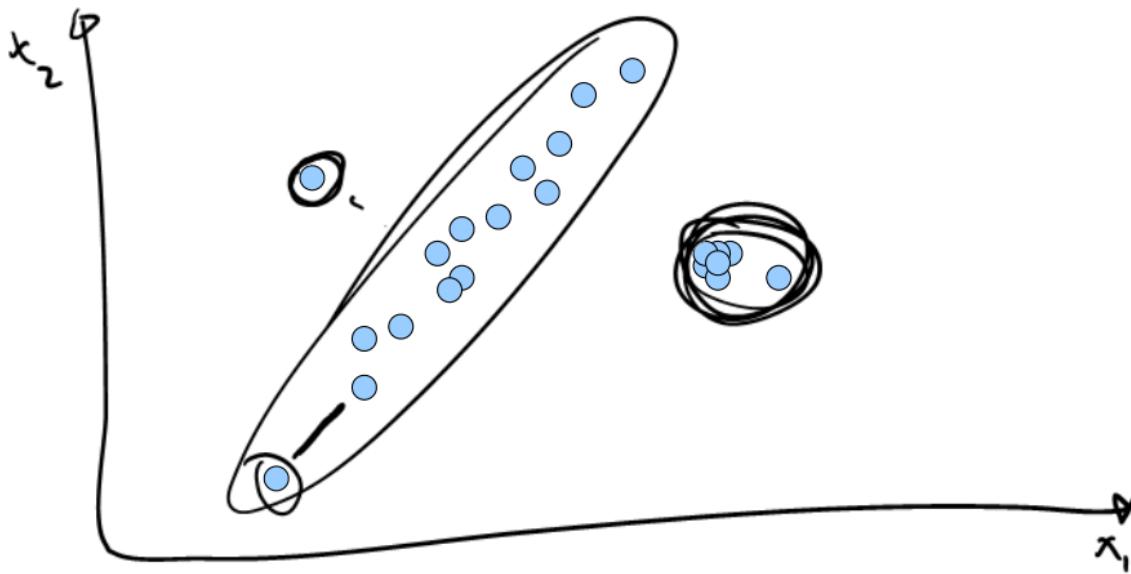
Association rule discovery: Example

Market basket analysis

| Training set | Rules discovered |
|-------------------------------|------------------------|
| 1.{Bread, Coke, Milk} | {Milk} ►{Coke} |
| 2.{Beer, Bread} | {Diaper, Milk} ►{Beer} |
| 3.{Beer, Coke, Diaper, Milk} | |
| 4.{Beer, Bread, Diaper, Milk} | |
| 5.{Coke, Milk} | |

Anomaly detection: Definition

- Given a collection of data objects
 - Each object has associated a number of features
- Detect which objects **deviate from normal** behaviour



Anomaly detection: Example

- Credit card **fraud detection**
 - Recognize dubious credit card transactions based on the transaction history of the card holder
- Detection of **outliers** in data measurements
 - 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

Models in machine learning

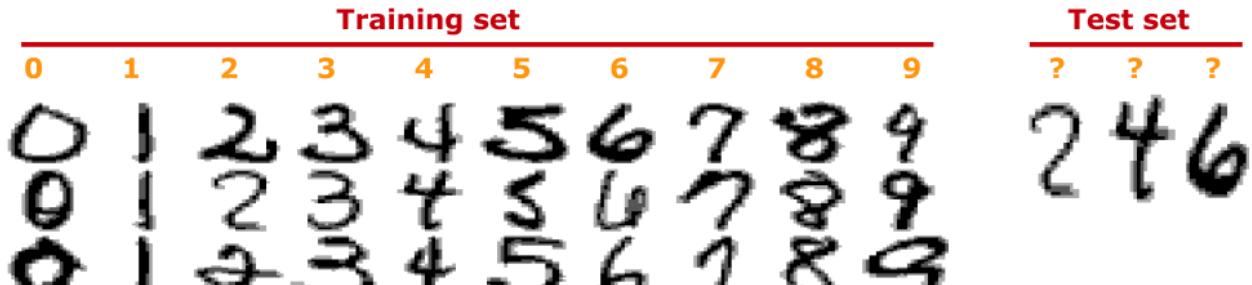
| Training set | | | | | | | | | | Test set | | |
|--------------|---|---|---|---|---|---|---|---|---|----------|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ? | ? | ? |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 2 | 4 | 6 |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | | | |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | | | |

Models in machine learning

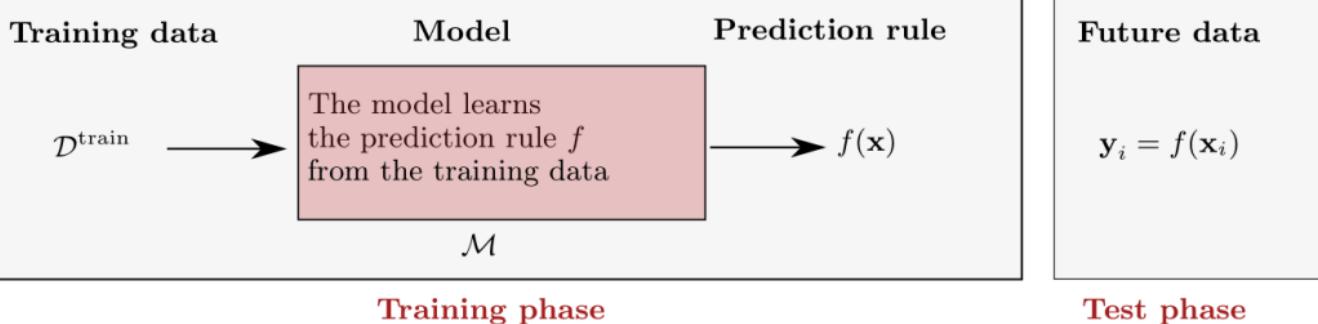
| Training set | | | | | | | | | | Test set | | |
|--------------|---|---|---|---|---|---|---|---|---|----------|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ? | ? | ? |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 2 | 4 | 6 |

Classifying digits is a mapping $f : \mathbb{R}^M \rightarrow \{0, 1, \dots, 9\}$

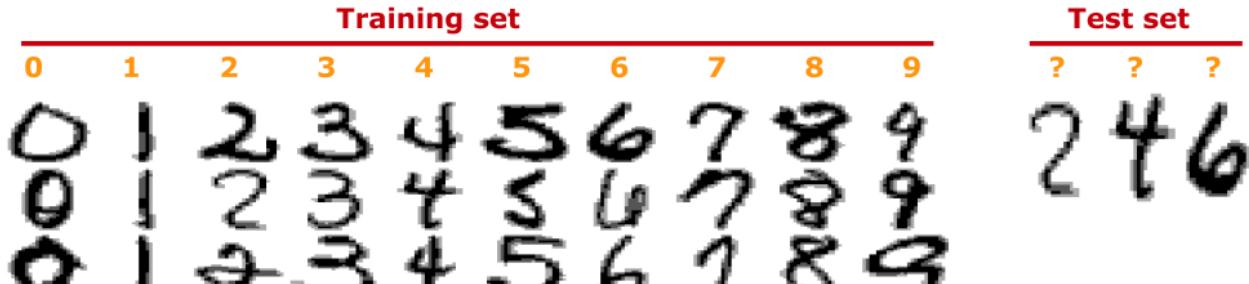
Models in machine learning



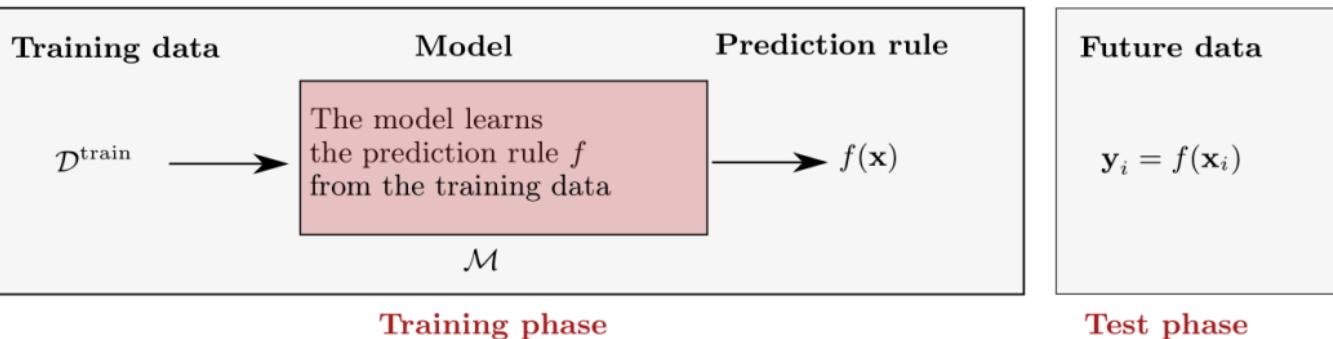
Classifying digits is a mapping $f : \mathbb{R}^M \rightarrow \{0, 1, \dots, 9\}$



Models in machine learning



Classifying digits is a mapping $f : \mathbb{R}^M \rightarrow \{0, 1, \dots, 9\}$



How often the learned function f makes errors in the future is the **generalization error**

Exercises

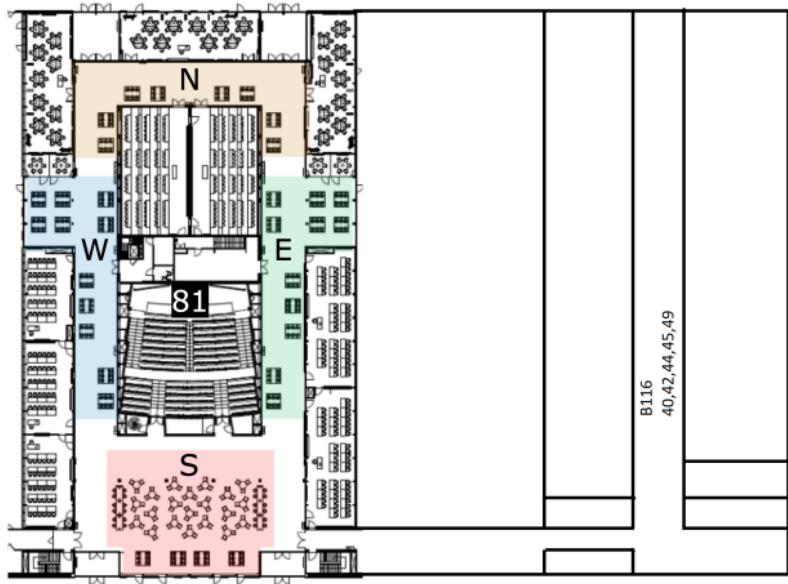
We support **Matlab**, **Python**, and **R**

- Exercise 0 guided you through installing your chosen environment
- If you have no experience with either, we recommend **Python**

Rooms for exercises:

- Building 116-A081, (Python,Matlab)
- Building 116-A083, (Python)
- Building 116-H010, (Python,R)
- Building 116-H011, (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 week 3)? Enter your info in MS Teams > General > Shared files > 02450missing_a_group.xlsx. We will then assign you to a group.
 - Once formed, register your group on DTU Learn.
- Start looking for a dataset and discuss the suitability of specific 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.

Resources

<https://www.mckinsey.com> Impact assessment of automation by McKinsey

(<https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy>)

<https://towardsdatascience.com> Another introduction to machine learning basics (<https://towardsdatascience.com/introduction-to-machine-learning-db7c668822c4>)

<https://www.economicsofai.com> conference on modelling economical impact of AI (<https://www.economicsofai.com/nber-conference-toronto-2017/>)

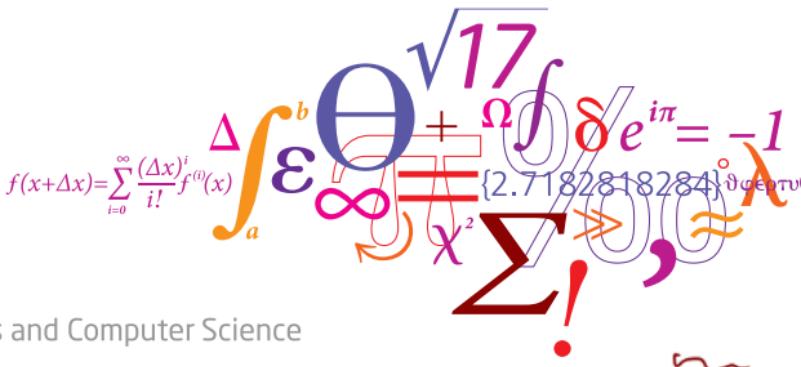
<https://deepmind.com> Obviously google-focused, but otherwise a great resource for what is hot right now (<https://deepmind.com/>)

02450: Introduction to Machine Learning and Data Mining

AUC and ensemble methods

Bjørn Sand Jensen

DTU Compute, Technical University of Denmark (DTU)



Today

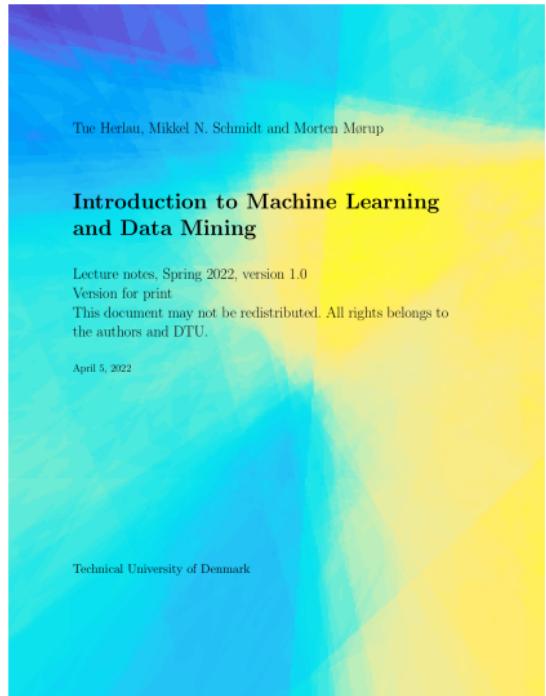
Feedback Groups of the day:

Mathias Schmidt Thomsen, Julius Gregers Gliese Winkel, Swati Tak, Jonathan Tybirk Munk, Lucas Rieneck Gottfried Pedersen, Ona Saulianskaite, Josefine Tvermoes Meineche, Thomas Würtzen, Sami Rana, Hallgrímur Thorsteinsson, Gian Marco Maratta, Lauren Walker, Dhruva Gajanan Sakhare, Karíta Ósk Pálmadóttir, Johanne Margrethe Lund, Sophie Mikkeline Stæhr, Frederik Riis Nielsen, Niels Due Rasmussen, Rodolphe Nakhlé, Christian Damén Schultz-Nielsen, Amalie Drud Nielsen, Alec Ranjithkar, Ditte Stuhr Petersen, Jonas Hvidtfeldt Møller, Lasse Heinrich Rommelfangen, Nicolas Gilles Roger Starc, Rebecca Hjermind Millum, Leonie Meier, Frodo Simon Hviid Mikkelsen, Anna Mitsiou, Cecilie Martine Møller-Jensen, Tazo Mukhadze, Hassen Ali Nasrallah, Radek Nesnídal, Sofie Marie Møller Nielsen, Lasse Hamborg Nielsen, Sifat E Noor, Emilie Nyholm-Christensen, Antoine Francois S Ohn, Mikkel Inti Skærbaek Tabi Espinosa Olsen



Reading material:

Chapter 16, Chapter 17



Lecture Schedule

1 Introduction

30 January: C1

Data: Feature extraction, and visualization

2 Data, feature extraction and PCA

6 February: C2, C3

3 Measures of Similarity, summary statistics and probabilities

13 February: C4, C5

4 Probability densities and data visualization

20 February: C6, C7

Supervised learning: Classification and regression

5 Decision trees and linear regression

27 February: C8, C9 (Project 1 due 29 February at 17:00)

6 Overfitting, cross-validation and Nearest Neighbor

5 March: C10, C12

7 Performance evaluation, Bayes, and Naive Bayes

12 March: C11, C13

Online 24/7 help: Discussion Forum/Piazza
Streaming: Zoom (see link on DTU Learn)
Recordings: <https://panopto.dtu.dk/>
Online exercises: MS Teams

8 Artificial Neural Networks and Bias/Variance

19 March: C14, C15

9 AUC and ensemble methods

2 April: C16, C17

Unsupervised learning: Clustering and density estimation

10 K-means and hierarchical clustering

9 April: C18 (Project 2 due 11 April at 17:00)

11 Mixture models and density estimation

16 April: C19, C20

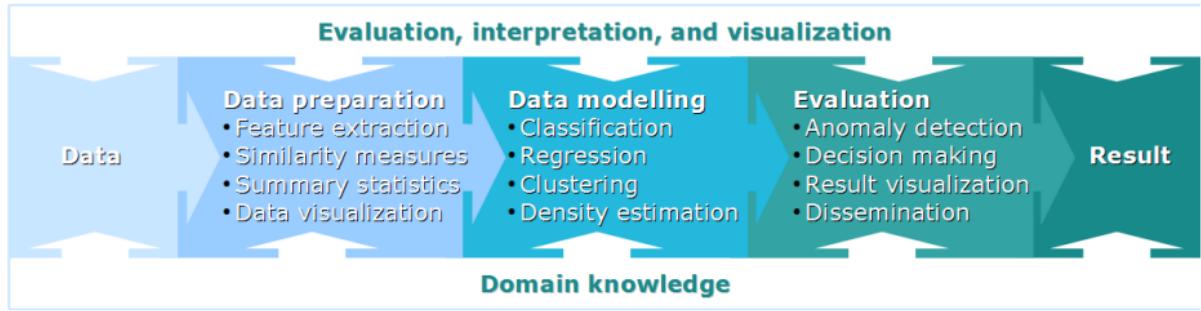
12 Association mining

23 April: C21

Recap

13 Recap and discussion of the exam

30 April: C1-C21

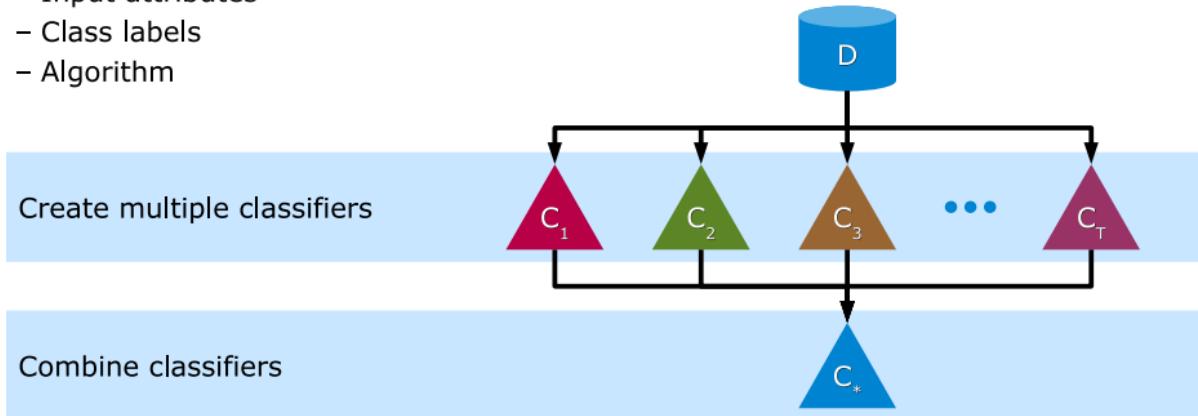


Learning Objectives

- Explain the principle behind boosting and bagging and apply it to improve classifiers
- Be able to address issues of class-imbalance and resampling
- Understand the definition of Precision, Recall, ROC, and AUC

Ensemble methods

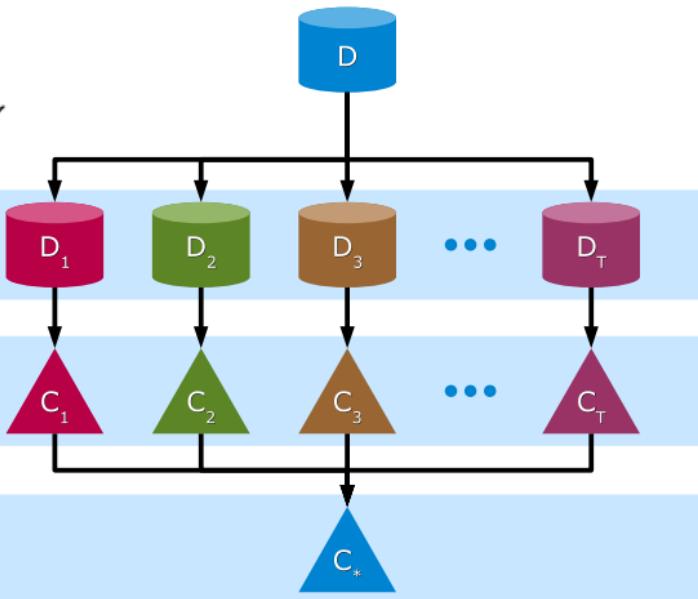
- Combine multiple (weak) classifiers into one (strong) classifier
- Each classifier trained using different variations of
 - Data set
 - Input attributes
 - Class labels
 - Algorithm



Ensemble methods

$$\delta(f_t(x), y) = \begin{cases} 1 & \text{if } f_t(x) = y \\ 0 & \text{if } f_t(x) \neq y \end{cases}$$

Create multiple data sets



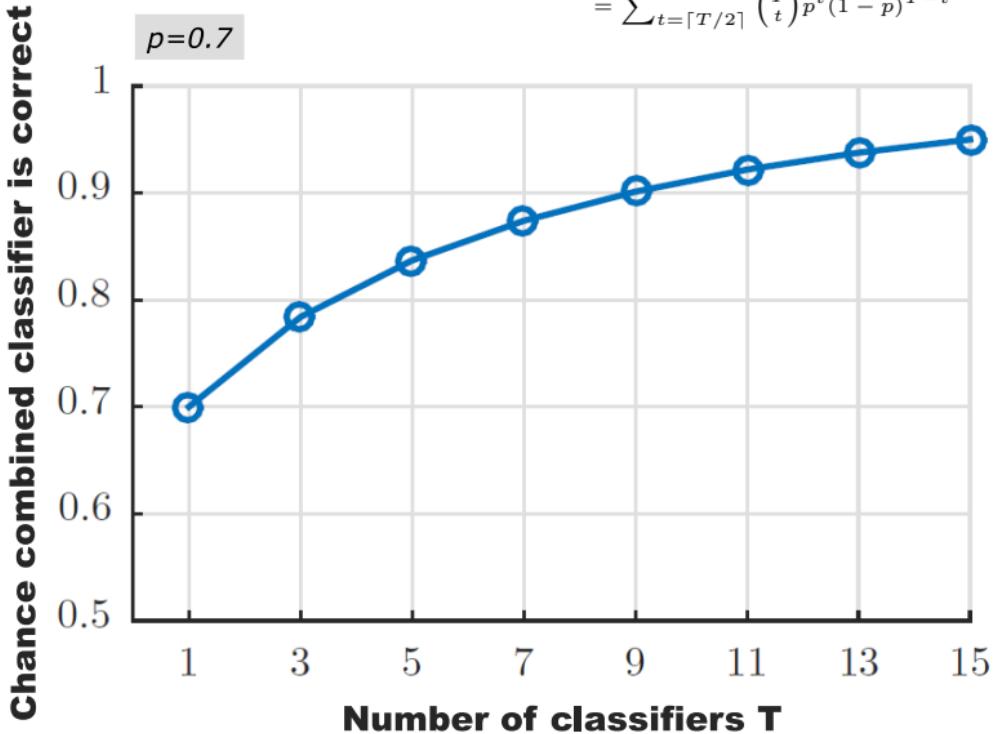
$$f(x) = \operatorname{argmax}_{y \in \{0, 1, 3\}} \sum_{t=1}^T \delta(f_t(x), y)$$

Why ensemble methods?

- Can improve classification algorithms in terms of
 - Better classification accuracy
 - Increased stability
 - Reduced variance
 - Less overfitting
- Consider T independent classifiers for binary classification, each with accuracy p .
- The probability a classifier which use majority voting is correct is then given by:

$$\begin{aligned} P(\text{Majority voting is correct}) &= \sum_{t=\lceil T/2 \rceil}^T P(\{t \text{ classifiers are correct}\}) \\ &= \sum_{t=\lceil T/2 \rceil}^T \binom{T}{t} p^t (1-p)^{T-t} \end{aligned}$$

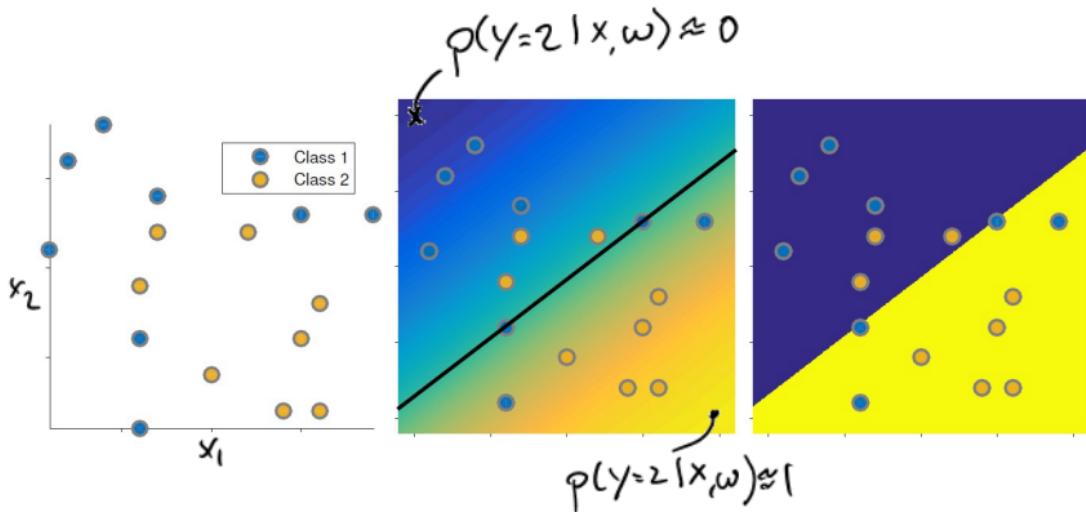
$$\begin{aligned} P(\text{Majority voting is correct}) &= \sum_{t=\lceil T/2 \rceil}^T P(\{t \text{ classifiers are correct}\}) \\ &= \sum_{t=\lceil T/2 \rceil}^T \binom{T}{t} p^t (1-p)^{T-t} \end{aligned}$$



$$p(y=2 | x_1, w) = \frac{1}{1 + e^{-(w_0 + w_1 x_1 + w_2 x_2)}} \in [0, 1]$$

Data example

- Classification using logistic regression



Bagging

- New training data sets drawn randomly from pool with replacement

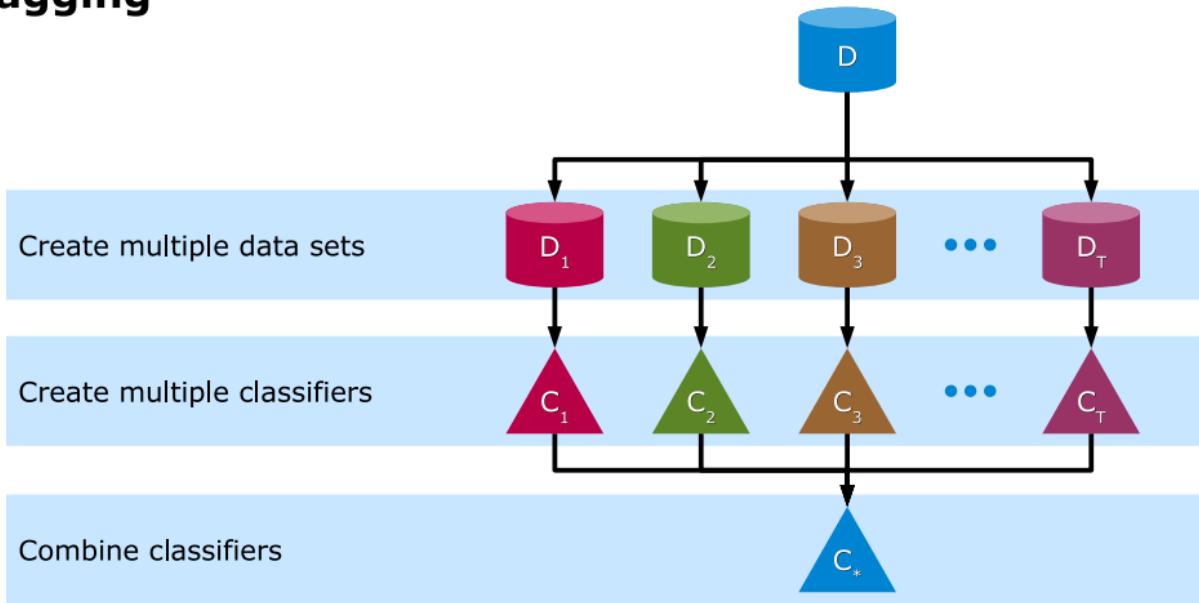
Pool of training data

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|----|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|----|

New training data sets

| | | | | | | | | | |
|---|---|---|---|----|---|----|----|----|---|
| 3 | 5 | 4 | 3 | 9 | 7 | 9 | 5 | 1 | 1 |
| 5 | 8 | 2 | 6 | 2 | 3 | 8 | 3 | 5 | 1 |
| 1 | 7 | 4 | 1 | 10 | 6 | 10 | 8 | 8 | 7 |
| ⋮ | | | | | | | | | |
| 4 | 3 | 8 | 5 | 2 | 4 | 7 | 10 | 10 | 8 |

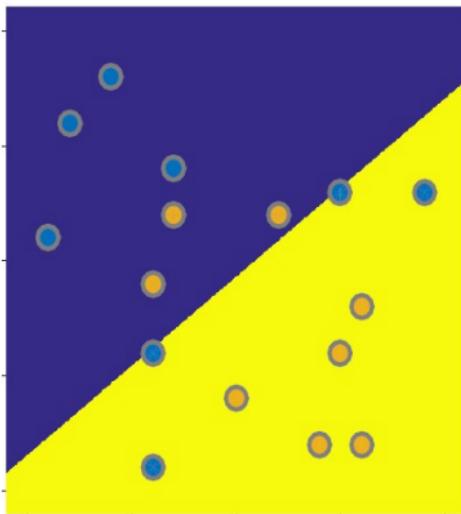
Bagging



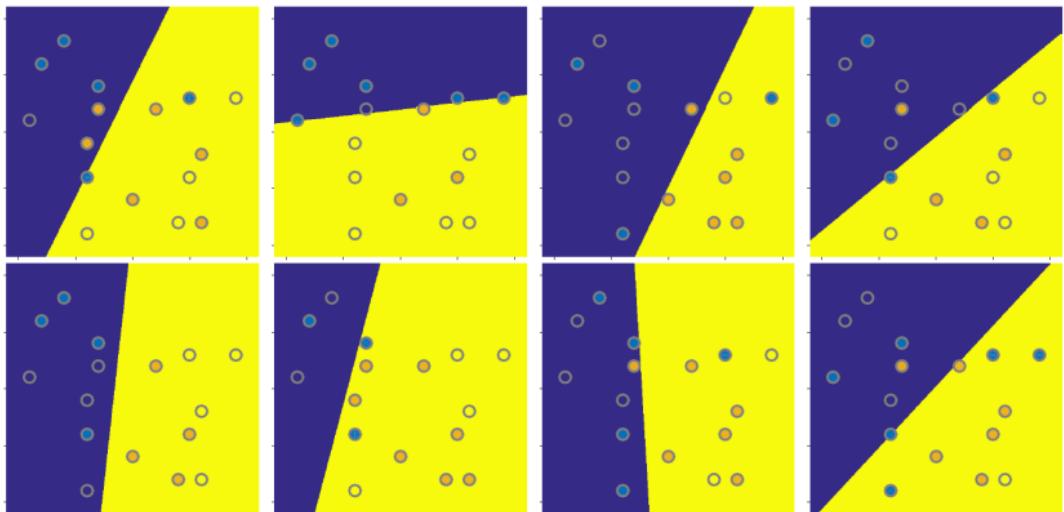
Bagging

- **Single classifier**

- Logistic regression
- Two features, (x, y)



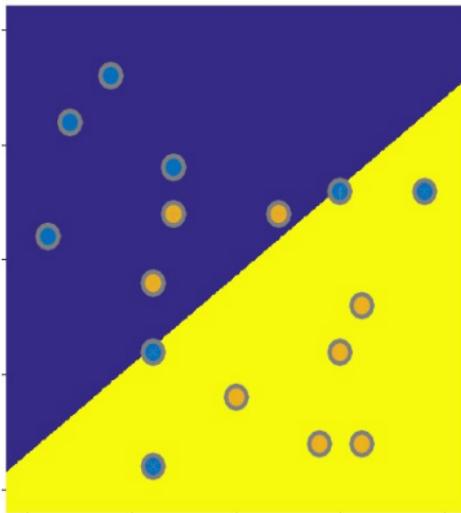
Bagging



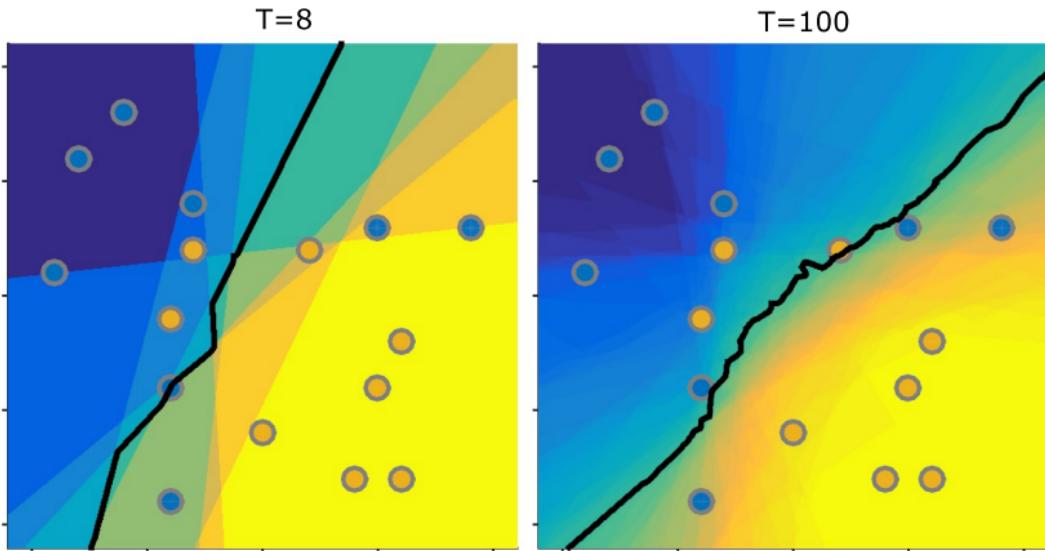
Notice, hollow dots are observations not included in bagging round

Bagging

- Single classifier



Bagging



Boosting

| | | | | | | | | | | |
|-----------------------|----|----|----|----|----|----|----|----|----|----|
| Pool of training data | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Weights | .1 | .1 | .1 | .1 | .1 | .1 | .1 | .1 | .1 | .1 |

| | | | | | | | | | | |
|-----------------------|---|---|---|---|---|---|---|---|---|---|
| New training data set | 3 | 5 | 4 | 3 | 9 | 7 | 9 | 5 | 1 | 1 |
|-----------------------|---|---|---|---|---|---|---|---|---|---|

Train classifier



Boosting

| | | | | | | | | | | |
|---------------------------|---|----|----|----|----|----|----|----|----|-----|
| Pool of training data | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Weights | .1 | .1 | .1 | .1 | .1 | .1 | .1 | .1 | .1 | .1 |
| New training data set | 3 | 5 | 4 | 3 | 9 | 7 | 9 | 5 | 1 | 1 |
| Train classifier |  | | | | | | | | | |
| Classify all data objects | 1✓ | 2✗ | 3✓ | 4✗ | 5✓ | 6✗ | 7✓ | 8✓ | 9✓ | 10✓ |

Boosting

| | | | | | | | | | | |
|---------------------------|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Pool of training data | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Weights | .1 | .1 | .1 | .1 | .1 | .1 | .1 | .1 | .1 | .1 |
| New training data set | 3 | 5 | 4 | 3 | 9 | 7 | 9 | 5 | 1 | 1 |
| Train classifier |  | | | | | | | | | |
| Classify all data objects | 1✓ | 2✗ | 3✓ | 4✗ | 5✓ | 6✗ | 7✓ | 8✓ | 9✓ | 10✓ |
| Update weights | .07 | .17 | .07 | .17 | .07 | .17 | .07 | .07 | .07 | .07 |

Boosting

Pool of training data

| | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| .1 | .1 | .1 | .1 | .1 | .1 | .1 | .1 | .1 | .1 |

Weights

New training data set

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 3 | 5 | 4 | 3 | 9 | 7 | 9 | 5 | 1 | 1 |
|---|---|---|---|---|---|---|---|---|---|

Train classifier



Classify all data objects

| | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1✓ | 2✗ | 3✓ | 4✗ | 5✓ | 6✗ | 7✓ | 8✓ | 9✓ | 10✓ |
| .07 | .17 | .07 | .17 | .07 | .17 | .07 | .07 | .07 | .07 |

Update weights

New training data set

| | | | | | | | | | |
|---|---|---|---|---|---|----|---|---|---|
| 6 | 4 | 7 | 3 | 2 | 4 | 10 | 2 | 5 | 6 |
|---|---|---|---|---|---|----|---|---|---|

Train classifier



AdaBoost

Algorithm 6: AdaBoost algorithm

- 1: Initialize $w_i(1) = \frac{1}{N}$ for $i = 1, \dots, N$
- 2: **for** $t = 1, \dots, T$ **do**
- 3: Create \mathcal{D}_t by sampling (with replacement) from \mathcal{D} according to $\mathbf{w}(t)$
- 4: Let f_t be the classifier *trained* on \mathcal{D}_t
- 5: $\epsilon_t = \sum_{i=1}^N w_i(t) (1 - \delta_{f_t(\mathbf{x}_i), y_i})$ (*weighted error of f_t on all data*).
- 6: $\alpha_t = \frac{1}{2} \log \frac{1-\epsilon_t}{\epsilon_t}$
- 7: For each i update weights using eq. (15.7):

$$w_i(t+1) = \frac{\tilde{w}_i(t+1)}{\sum_{j=1}^N \tilde{w}_j(t+1)}, \quad \tilde{w}_i(t+1) = \begin{cases} w_i(t)e^{-\alpha_t} & \text{if } f_t(\mathbf{x}_i) = y_i \\ w_i(t)e^{\alpha_t} & \text{if } f_t(\mathbf{x}_i) \neq y_i. \end{cases}$$

- 8: **end for**
 - 9: $f^*(\mathbf{x}) = \arg \max_{y=1,2} \sum_{t=1}^T \alpha_t \delta_{f_t(\mathbf{x}), y}$ (*Majority voting classifier*)
-

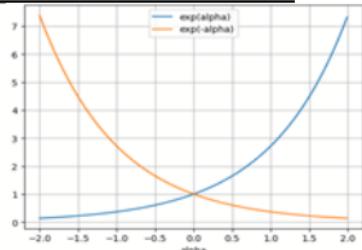
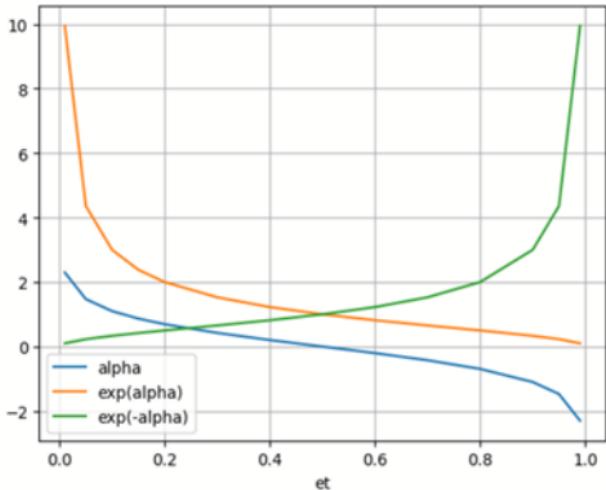
AdaBoost

Algorithm 6: AdaBoost algorithm

- 1: Initialize $w_i(1) = \frac{1}{N}$ for $i = 1, \dots, N$
- 2: **for** $t = 1, \dots, T$ **do**
- 3: Create \mathcal{D}_t by sampling (with replacement) f
- 4: Let f_t be the classifier *trained* on \mathcal{D}_t
- 5: $\epsilon_t = \sum_{i=1}^N w_i(t) (1 - \delta_{f_t(\mathbf{x}_i), y_i})$ (*weighted error*)
- 6: $\alpha_t = \frac{1}{2} \log \frac{1-\epsilon_t}{\epsilon_t}$
- 7: For each i update weights using eq. (15.7):

$$w_i(t+1) = \frac{\tilde{w}_i(t+1)}{\sum_{j=1}^N \tilde{w}_j(t+1)}, \quad \tilde{w}_i(t+1) = \begin{cases} w_i(t)e^{-\alpha_t} & \text{if } f_t(\mathbf{x}_i) = y_i \\ w_i(t)e^{\alpha_t} & \text{if } f_t(\mathbf{x}_i) \neq y_i. \end{cases}$$

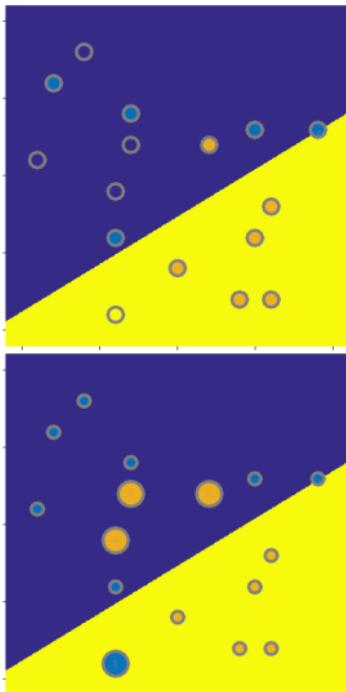
- 8: **end for**
 - 9: $f^*(\mathbf{x}) = \arg \max_{y=1,2} \sum_{t=1}^T \alpha_t \delta_{f_t(\mathbf{x}), y}$ (*Majority voting classifier*)
-



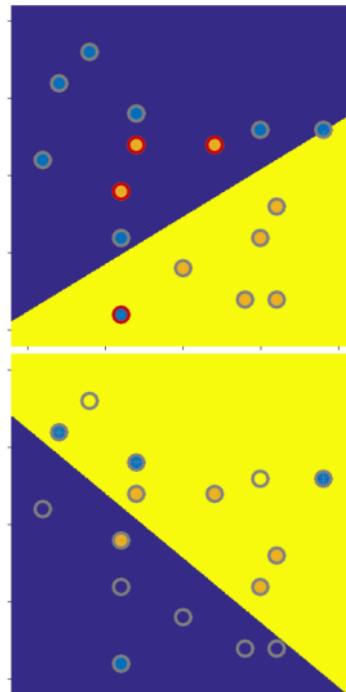
Boosting

A:

A dataset is sampled with replacement and a classifier trained.

**C:**

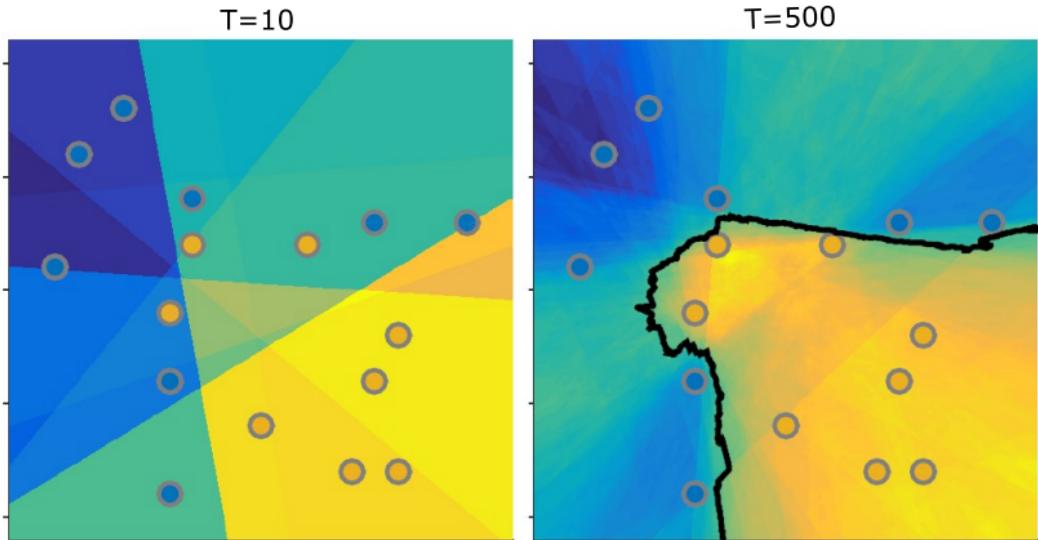
Weights are updated such that more emphasis is given to these mis-classified observations.

**B:**

Mis-classified observations are identified.

New round:
Based on the updated weights a new dataset is sampled and a classifier trained (shown), mis-classified observations identified and given more emphasis...

Boosting

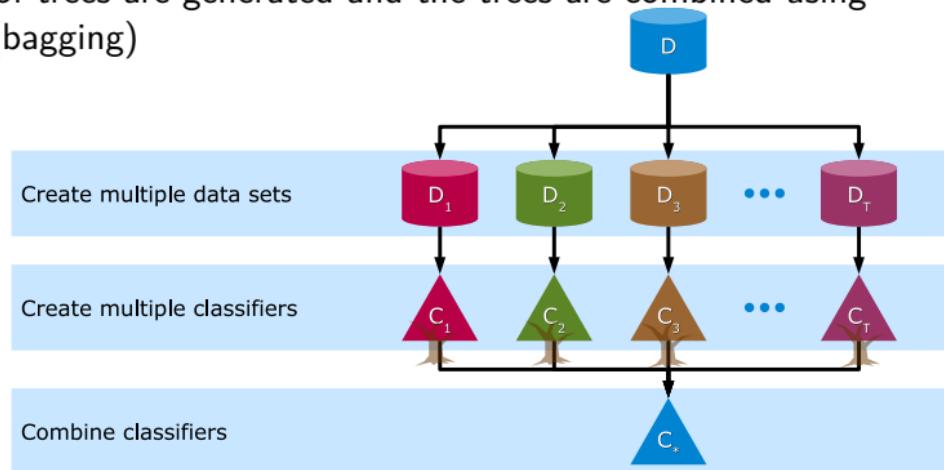


Bagging example: Random forest

Each tree is generated as follows:

- Sample dataset with replacement
- When generating each node in the tree, randomly select a subset of the features and only consider splits using these features

A large number of trees are generated and the trees are combined using majority voting (bagging)



Quiz 1 (please answer on ~~Piazza~~): Adaboost (Spring 2016)

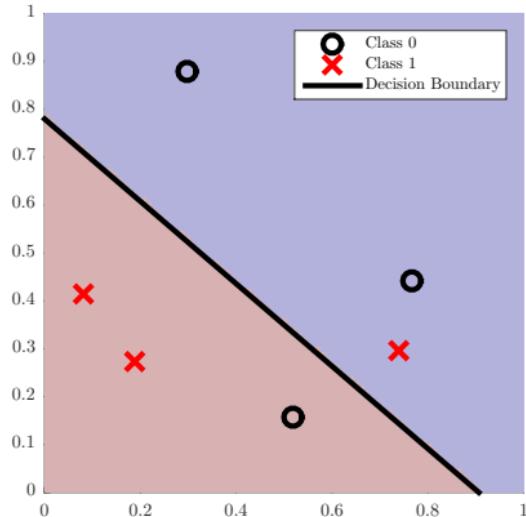


Figure 1: A binary classification problem and the decision boundary obtained by logistic regression. Observations left of the boundary are classified as belonging to the positive class 1 (red crosses) and observations right of the boundary to the negative class 0 (black circles)

$$\epsilon_{t=1} = \frac{1}{6} \times 2 = \frac{1}{3}$$

$$\alpha_{t=1} = \frac{1}{2} \log \frac{1 - \epsilon_t}{\epsilon_t} = \frac{1}{2} \log 2$$

We wish to apply a logistic regression model to the binary classification problem shown in Figure 1. We attempt to improve the performance by applying AdaBoost. AdaBoost works by first sampling a new dataset with replacement, then training a classifier on the dataset and then proceeding with the subsequent steps of the AdaBoost algorithm.

Suppose in the first iteration of the AdaBoost algorithm the classification boundary of the trained classifier is as indicated by the black line (i.e. observations left of the black line are classified as in the positive class). What is the resulting value of the weights w ?

- A. $w = [0.125 \quad 0.250 \quad 0.125 \quad 0.125 \quad 0.125 \quad \underline{0.250}]$
- B. $w = [0.026 \quad 0.447 \quad 0.026 \quad 0.026 \quad 0.026 \quad 0.447]$
- C. $w = [0.235 \quad 0.029 \quad 0.235 \quad 0.235 \quad 0.235 \quad 0.029]$
- D. $w = [0.1 \quad 0.3 \quad 0.1 \quad 0.1 \quad 0.1 \quad 0.3]$
- E. Don't know.

(Hint: First compute ϵ_1 , then α_1 , then the weights)

$$\tilde{W} = \frac{1}{6} [e^{-\alpha_6} e^{\alpha_6} e^{-\alpha_6} e^{-\alpha_6} e^{-\alpha_6} e^{\alpha_6}]$$

Solution:

$$w_i = \frac{\tilde{w}_i}{\sum_{i \sim i} \tilde{w}_i} \approx$$

The classifier classifies two out of $N = 6$ observations incorrectly. We therefore have:

$$\varepsilon_i = \left[\sum_{j=1}^N w_j I(\hat{y}_j \neq y_j) \right]$$

$$\alpha_i = \frac{1}{2} \log \left(\frac{1 - \varepsilon_i}{\varepsilon_i} \right)$$

and accordingly $\varepsilon_1 = \frac{1}{N} \times 2 = \frac{1}{3}$. This gives

$$\alpha_1 = \frac{1}{2} \log \frac{1 - \frac{1}{3}}{\frac{1}{3}} = \frac{1}{2} \log 2$$

and so for w we get

$$w \propto [e^{-\alpha_1} \quad e^{\alpha_1} \quad e^{-\alpha_1} \quad e^{-\alpha_1} \quad e^{-\alpha_1} \quad e^{\alpha_1}]$$

Simplifying by moving $\frac{1}{\sqrt{2}}$ outside the vector:

$$w \propto [1 \quad 2 \quad 1 \quad 1 \quad 1 \quad 2]$$

and normalizing:

$$w = \frac{1}{8} [1 \quad 2 \quad 1 \quad 1 \quad 1 \quad 2]$$

accordingly option A is correct.

Class imbalance problem

- Many data sets have **imbalanced class distributions**
 - Example: Detection of defects that only occur rarely (e.g. 1/1,000,000)
 - Danger: Algorithm that says nothing is defect will be 99.999% correct
- **Solution approaches**
 - Resample to balance data sets
 - Modify existing classification algorithms
 - Measure performance in a way that takes balance into account

Resampling balanced data

- New sample has equal number of data objects from each class

- **Approaches**

- **Undersampling** majority class: Throws out potentially useful data
- **Oversampling** minority class: Increase data size and computational burden
- **Somewhere in between...**

| | | | | | | | | | | |
|--------------------------|---|---|---|---|---|---|---|----|---|----|
| Imbalanced training data | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Oversampling | 1 | 2 | 3 | 4 | 5 | 7 | 9 | 10 | 6 | 6 |
| | 6 | 6 | 8 | 8 | 8 | 8 | | | | |
| Undersampling | 3 | 5 | 6 | 8 | | | | | | |
| Somewhere in between | 3 | 5 | 4 | 3 | 9 | 6 | 6 | 8 | 8 | 8 |

$$Acc = \frac{TP + TN}{TP + TN + FN + F}$$

Confusion matrix

| | | <i>Predicted</i> | |
|---------------|-----------------|-----------------------------|-----------------------------|
| | | <i>positive</i> | <i>negative</i> |
| <i>Actual</i> | <i>positive</i> | TP True Positive | FN False Negative |
| | <i>negative</i> | FP False Positive | TN True Negative |

Precision and recall

- **Precision**

- Fraction of true positive among objects predicted to be positive

$$p = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- **Recall**

- Fraction of objects predicted to be positive among all positive objects

$$r = \frac{\text{TP}}{\text{TP} + \text{FN}}$$



| | | <i>Predicted</i> | |
|---------------|----------|----------------------|----------------------|
| | | positive | negative |
| <i>Actual</i> | positive | TP True Positive | FN False Negative |
| | negative | FP False Positive | TN True Negative |

The table illustrates the four outcomes of a classification test:

- True Positive (TP):** Objects correctly predicted to be positive.
- False Positive (FP):** Objects incorrectly predicted to be positive (among negative objects).
- True Negative (TN):** Objects correctly predicted to be negative.
- False Negative (FN):** Objects incorrectly predicted to be negative (among positive objects).

Two specific cells in the matrix are highlighted with colored circles: TP (True Positive) is circled in blue, and FN (False Negative) is circled in red. A green circle highlights the entire row for "Actual positive".

Quiz 2 (please answer on ~~Piazza~~): Precision/Recall

Consider two different classifiers, and suppose on a test set with 20 positive observations:

- Classifier 1 detects 54 positive of which 18 are actually positive
- Classifier 2 detects 16 positive of which 14 are actually positive

| | | Predicted | |
|--------|----------|----------------|----------------|
| | | positive | negative |
| Actual | positive | TP 14 | FN 2 |
| | negative | FP 2 | TN 16 |
| | | True Positive | False Negative |
| | | False Positive | True Negative |

• Precision

- Fraction of true positive among objects predicted to be positive

$$p = \frac{TP}{TP + FP}$$

What is the precision and recall of the two classifiers?

- A. Classifier 1: $p_1 = \frac{2}{3}, r_1 = \frac{7}{10}$
Classifier 2: $p_2 = \frac{1}{3}, r_2 = \frac{1}{3}$
- B. Classifier 1: $p_1 = \frac{1}{3}, r_1 = \frac{9}{10}$
Classifier 2: $p_2 = \frac{2}{3}, r_2 = \frac{9}{10}$
- C. Classifier 1: $p_1 = \frac{2}{3}, r_1 = \frac{7}{10}$
Classifier 2: $p_2 = \frac{1}{3}, r_2 = \frac{9}{10}$
- D. Classifier 1: $p_1 = \frac{1}{2}, r_1 = \frac{9}{10}$
Classifier 2: $p_2 = \frac{7}{8}, r_2 = \frac{7}{10}$



Which classifier would you use if the objective was to detect credit-card fraud (the positive class corresponds to fraud)

• Recall

- Fraction of objects predicted to be positive among all positive objects

$$r = \frac{TP}{TP + FN}$$

Solution:

$$P_1 = \frac{TP}{TP + FP} = \frac{14}{54} = \frac{1}{3}$$

$$r_1 = \frac{TP}{TP + FN} = \frac{14}{20} = \frac{7}{10}$$

The precision is the number of true positives divided by the number *predicted* to be positive. I.e. for the two classifiers

$$p_1 = \frac{18}{54} = \frac{1}{3}, \quad p_2 = \frac{14}{16} = \frac{7}{8}.$$

The recall is the number of true positives divided by the number of *actual* positives:

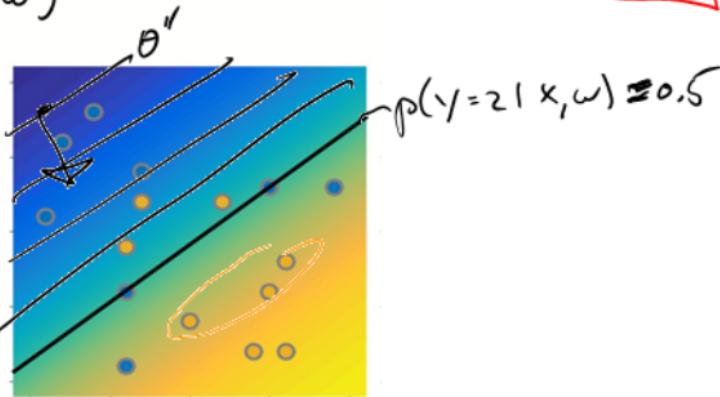
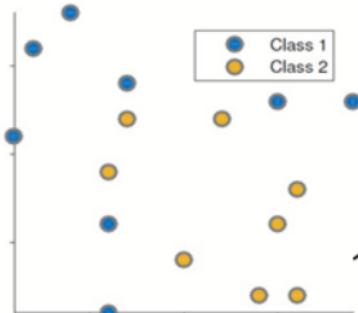
$$r_1 = \frac{18}{20} = \frac{9}{10}, \quad r_2 = \frac{14}{20} = \frac{7}{10}.$$

If the classifiers are supposed to detect credit-card fraud, we probably want a classifier which does not incorrectly *miss* many fraudulent transactions, i.e. it should have a low value of FN. In that case a high recall is desirable, and we should go with classifier 1.

Data example

- Classification using logistic regression

$$p(y=2|x, \omega)$$



| | | 1 | 0 |
|------|---|----|----|
| act. | 1 | TP | FN |
| | 0 | FP | TN |

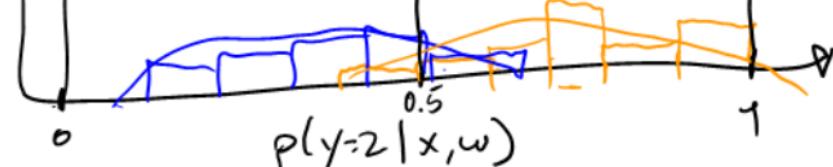
| θ | 1 | 0 |
|----------|---|---|
| 1 | 8 | 6 |
| 0 | 6 | 2 |

$\theta > 0.5$

| | |
|---|---|
| 5 | 3 |
| 4 | 4 |

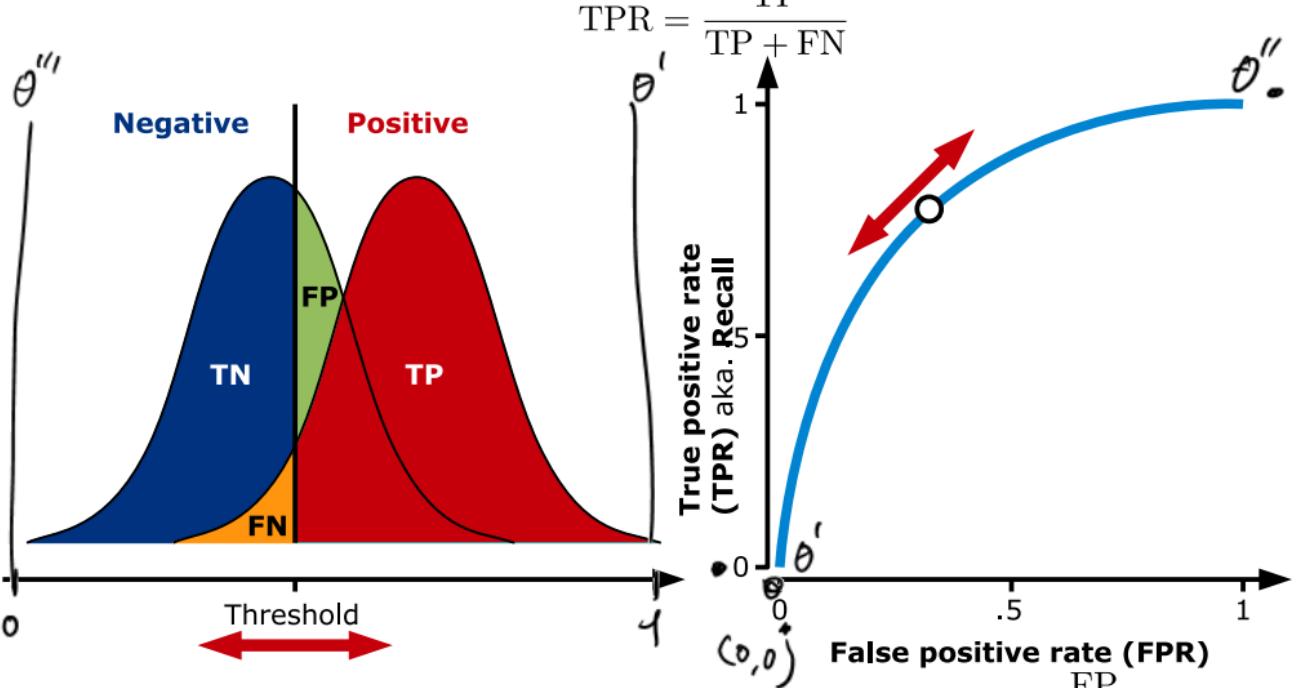
$P = \frac{1}{1 + e^{-\theta}}$

| θ' | 1 | 0 |
|-----------|---|---|
| 1 | 0 | 8 |
| 0 | 0 | 8 |

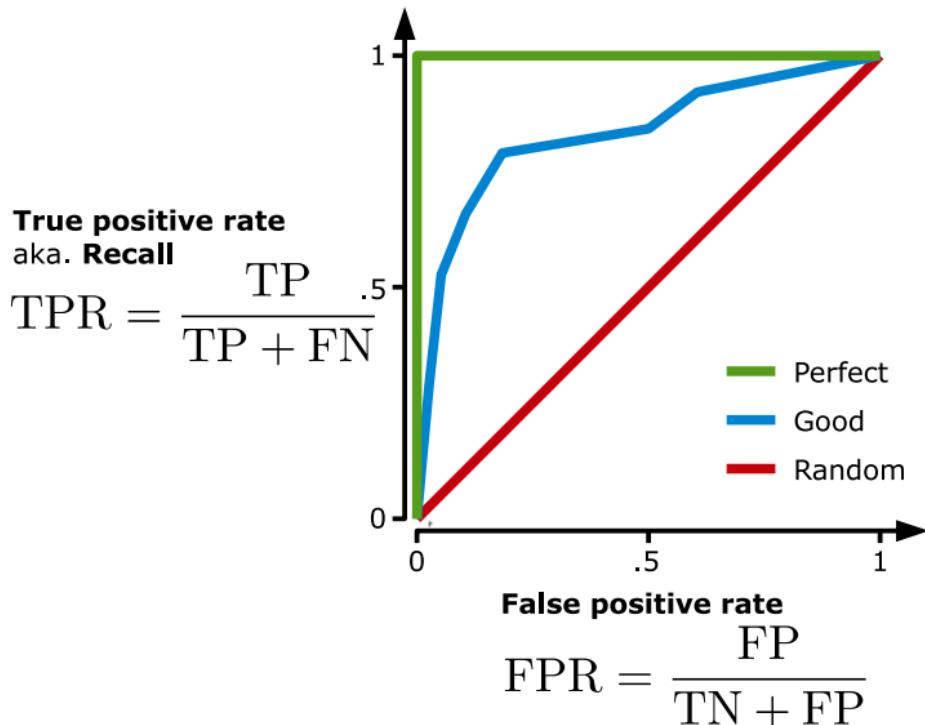


| Predicted | |
|----------------|----------------|
| Actual | |
| positive | negative |
| TP | FN |
| True Positive | False Negative |
| FP | TN |
| False Positive | True Negative |

Receiver operating characteristic



Receiver operating characteristic



Quiz 3 (please answer on ~~Pizza~~): AUC (Spring 2017)

| | 3 gears ($x_5 = 3$) | 4 gears ($x_5 = 4$) | 5 gears ($x_5 = 5$) |
|----------------------|--------------------------|--------------------------|--------------------------|
| Low mpg ($y = 0$) | 13 | 2 | 2 |
| High mpg ($y = 1$) | 2 | 10 | 3 |

Table 1: Number of low mpg and high mpg cars (i.e. $y = 0$ and $y = 1$) according to the number of gears, i.e. $x_5 = 3$, $x_5 = 4$, or $x_5 = 5$.

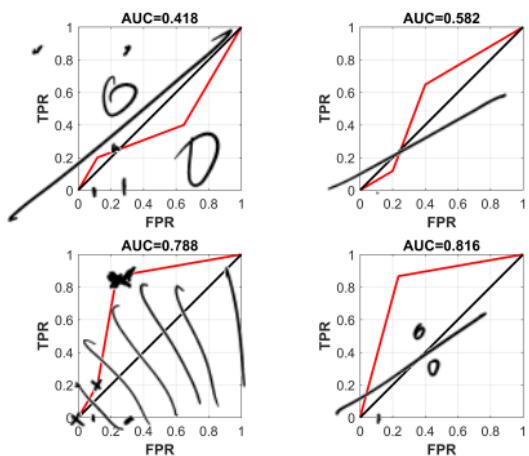


Figure 1: Four different receiver operator characteristic (ROC) curves and their area under curve (AUC) value.

A dataset representing cars contain an attribute x_5 corresponding to the number of gears. We wish to evaluate how well the number of gears predict low mpg, ($y = 0$, considered the negative class) from high mpg, ($y = 1$, considered the positive class) based on the data given in Table 1. For this purpose, we will evaluate the area under curve (AUC) of the receiver operator characteristic (ROC) using the feature x_5 . Which one of the ROC curves given in Figure 1 corresponds to using x_5 to discriminate between low mpg ($y = 0$) and high mpg ($y = 1$)?

- A. The curve having AUC=0.418
- B. The curve having AUC=0.582
- C. The curve having AUC=0.788
- D. The curve having AUC=0.816
- E. Don't know.

(Hint: Select a value e.g. $x_5 = 4.5$. We then predict cars with 5 or more gears as being in the positive class and otherwise negative. Compute the FPR and TPR using this prediction and use the (FPR, TPR) values to discriminate between the curves)

| pred. | | $\theta = 5.5$ | $\theta = 4.5$ | $\theta = 3.5$ | 2.5 |
|-------|---|----------------|----------------|----------------|-----|
| | | 1 | 0 | | |
| act | 1 | TP 15 | FN 17 | | |
| | 0 | FP 17 | TN 15 | | |

$$FPR = \frac{FP}{FP + TN}$$

(0, 0)

$(\frac{2}{17}, \frac{3}{15})$

$$TPR = \frac{TP}{TP + FN}$$

$(0.12, 0.2)$

$(0.24, 0.89)$ (1, 1)

Solution:

The ROC curve can be calculated by lowering the threshold, as no cars have more than 5 forward gears a threshold above 5 will result in the point (0,0). Lowering the threshold we find at the value 5 that $2/17$ of the low mpg cars (FPR) are at 5 and $3/15$ of the high mpg cars (TPR) are at 5 corresponding to

the point $(2/17, 3/15)$. When lowering to a threshold of 4 gears or more we are at the point $(4/17, 13/15)$ and at a threshold at 3 gears or more we have $(1,1)$. Thus, this curve corresponds to the curve having AUC=0.788.

Resources

<https://www.youtube.com> Video tutorial on ROC curve and AUC

(<https://www.youtube.com/watch?v=0Al6eAyP-yo>)

<https://towardsdatascience.com> More in-depth discussion of the Random Forrest algorithm and parameter choices

(<https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd>)

<https://www.datacamp.com> Practical use of the random forest algorithm in python

(<https://www.datacamp.com/community/tutorials/random-forests-classifier-python>)

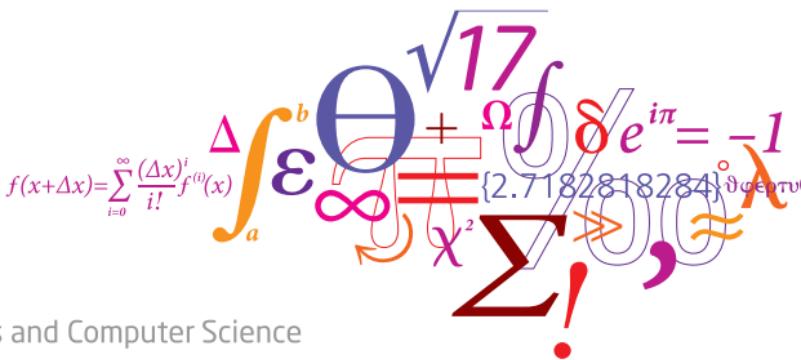
<https://citeseerx.ist.psu.edu> Justification for the AdaBoost algorithm (technical) (<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.51.9525>)

02450: Introduction to Machine Learning and Data Mining

Artificial Neural Networks and Bias/Variance

Bjørn Sand Jensen

DTU Compute, Technical University of Denmark (DTU)



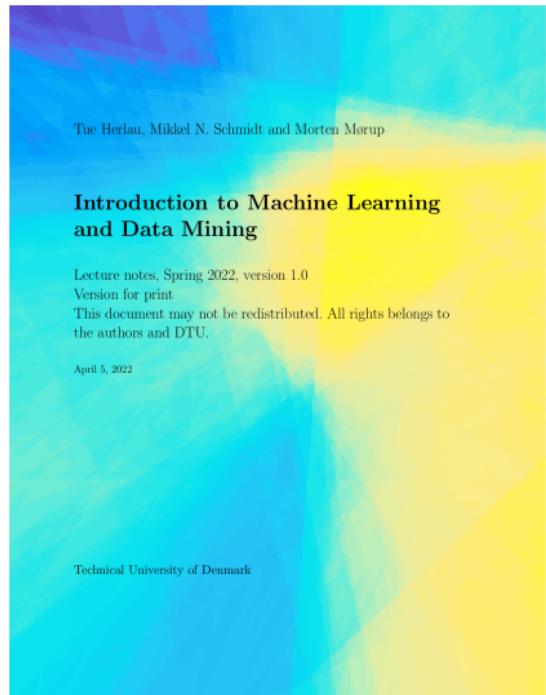
Today

Feedback Groups of the day:

Antonios Otapasidis, Konstantinos Lyroudis, Maxim Zavidei, Oliver Rasmussen, Gabriela Anna Penarska, Sarah Hecht Petersen, Jonas Rose Lund Pedersen, Reinis Muiznieks, Teodor August Obelitz, Jasmin Lundager Aasbjerg Petersen, Viset Raksa, Ninna Juul Ligaard, Niyaz Mahmud Sayem, Peter Christian Svensson, Philippe Flemming Lind Rasmussen, Casper Wayne Sahl, Rasmus Thielsen, William Vuong, Christian Amtoft Nickelsen, Viet Hoang Nguyen, Joseph An Duy Nguyen, Xavier Viñas Margalef, Valeriu Seremet, Alexandra Pouliasi, Rita Martí Torra, Dagfinnur Ari Normann, Tomas Vasconcelos Estacio, Nicolò Pandolfo, Carl Lundsøl Wernersson, Emil Noohr Nielsen, Jakob Lauge Reeh, Christian Ludvig Meinert Sørensen, Alexander Baatz Lorentzen, Christoffer Roland Olesen, Aniq Kashaf Shamim, Ying Li, Aoling Li, Ziwei Li, Fedor Lipskerov, Yiyang Liu, Benjamin Noah Lumbye, Lucas Christopher Dybendal Maack, Aksel Mads Madsen, Karrar Adam Mahdi, Yehudah Marcus, Neokleia Margariti, Adam Matuska, Lila Ines Fatima Meflah

Reading material:

Chapter 14, Chapter 15



Lecture Schedule

1 Introduction

30 January: C1

Data: Feature extraction, and visualization

2 Data, feature extraction and PCA

6 February: C2, C3

3 Measures of Similarity, summary statistics and probabilities

13 February: C4, C5

4 Probability densities and data visualization

20 February: C6, C7

Supervised learning: Classification and regression

5 Decision trees and linear regression

27 February: C8, C9 (Project 1 due 29 February at 17:00)

6 Overfitting, cross-validation and Nearest Neighbor

5 March: C10, C12

7 Performance evaluation, Bayes, and Naive Bayes

12 March: C11, C13

Online 24/7 help: Discussion Forum/Piazza
Streaming: Zoom (see link on DTU Learn)
Recordings: <https://panopto.dtu.dk/>
Online exercises: MS Teams

8 Artificial Neural Networks and Bias/Variance

19 March: C14, C15

9 AUC and ensemble methods

2 April: C16, C17

Unsupervised learning: Clustering and density estimation

10 K-means and hierarchical clustering

9 April: C18 (Project 2 due 11 April at 17:00)

11 Mixture models and density estimation

16 April: C19, C20

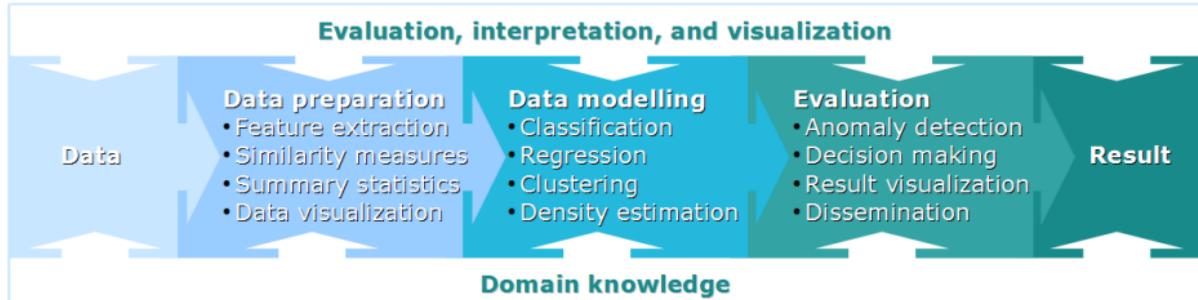
12 Association mining

23 April: C21

Recap

13 Recap and discussion of the exam

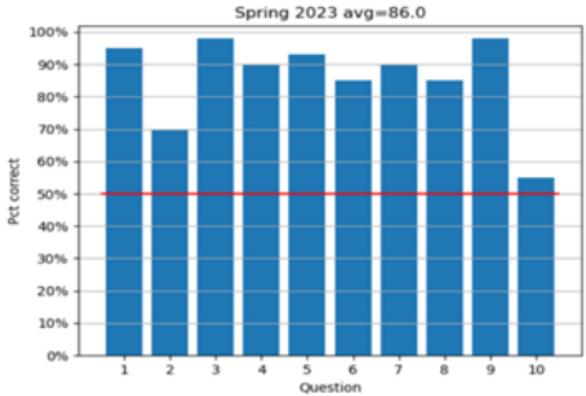
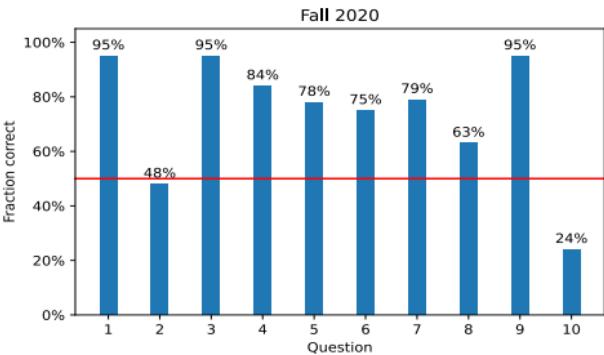
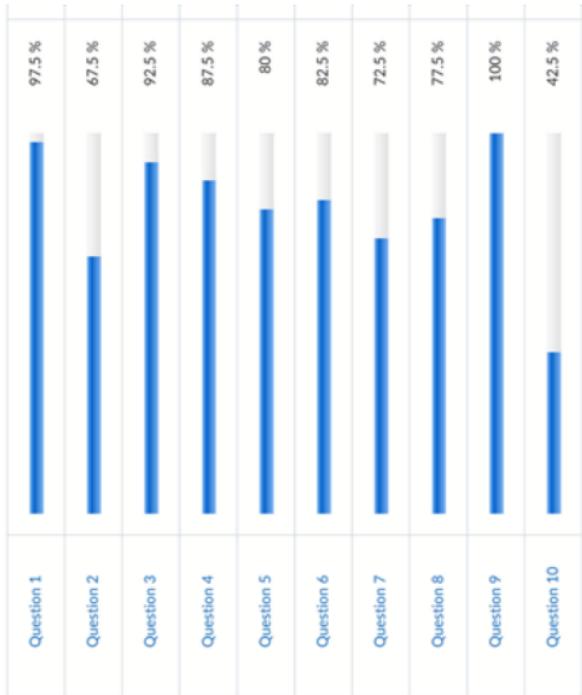
30 April: C1-C21



Learning Objectives

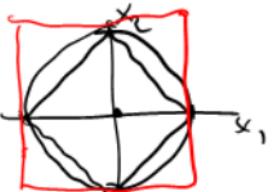
- Understand the Bias-Variance decomposition
- Understand and apply regularized least squares regression (i.e. ridge regression)
- Understand the principles behind artificial neural networks (ANNs) and how ANNs can be used for classification and regression
- Understand how logistic regression and ANNs can be extended to multi-class classification

Midterm practice test results



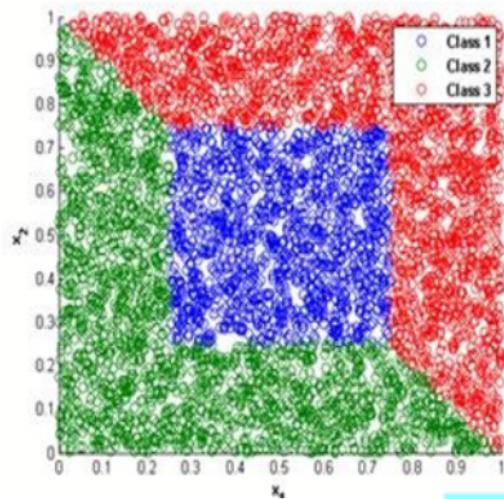
Solutions are at the end of this presentation

$$\|\mathbf{x} - \boldsymbol{\mu}\|_p$$

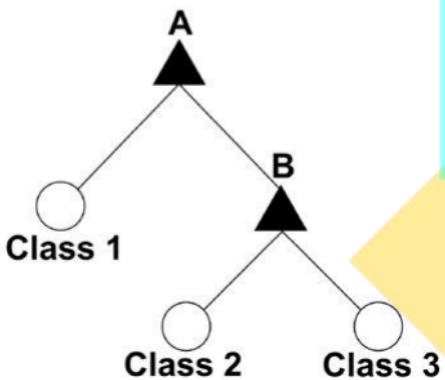
**Question 2:**

Consider the classification problem given in figure 1 and the Decision Tree in figure 2 with two decisions denoted A and B. We will let \mathbf{x}_n define the x_1 and x_2 coordinates of a given observation whereas $\mathbf{x}_n - 0.5 \cdot \mathbf{1}$ denotes the subtraction of 0.5 from x_1 and x_2 .

Which one of the following classification rules would lead to a correct classification of the data?



- A: A: $\|\mathbf{x}_n - 0.5 \cdot \mathbf{1}\|_1 \leq 0.25$, B: $\|\mathbf{x}_n\|_\infty \leq 1$
- B: A: $\|\mathbf{x}_n\|_1 \leq 1$, B: $\|\mathbf{x}_n - 0.5 \cdot \mathbf{1}\|_2 \leq \infty$
- C: A: $\|\mathbf{x}_n - 0.5 \cdot \mathbf{1}\|_2 \leq 0.25$, B: $\|\mathbf{x}_n\|_\infty \leq 1$
- D: A: $\|\mathbf{x}_n - 0.5 \cdot \mathbf{1}\|_\infty \leq 0.25$, B: $\|\mathbf{x}_n\|_1 \leq 1$
- E: Don't know



$$\tilde{X} = V \Sigma V^T$$
$$\Sigma = \begin{bmatrix} 16 & 6 & 6 & 0 \\ 6 & 16 & -2 & 0 \\ 6 & -2 & 16 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Question 8:

When carrying out a principal component analysis of a dataset with four attributes we obtain the following singular values $s_1=4$, $s_2=2$, $s_3=1$, and $s_4=0$. Which one of the following statements is wrong?

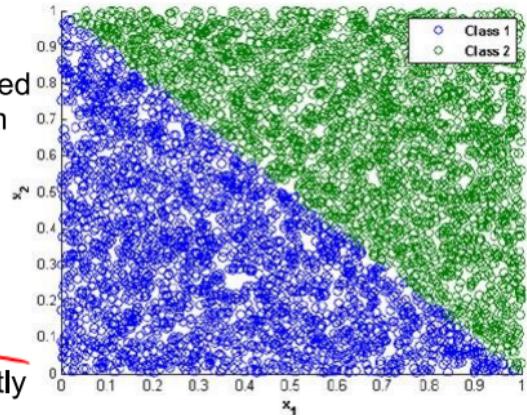
- A: The first principal component accounts for more than 60 % of the variation in the data.
- B: The third principal component accounts for less than 5 % of the variation in the data.
- C: The second principal component accounts for more than 20 % of the variation in the data.
- D: The data can be perfectly represented in a three dimensional sub-space.
- E: Don't know.

$$\frac{\sigma_2^2}{\sum_{i=1}^4 \sigma_i^2} = \frac{2^2}{4^2 + 2^2 + 1^2 + 0^2} = \frac{4}{16 + 4 + 1} = \frac{4}{21}$$

$$\frac{\sigma_2^2 + \sigma_3^2}{\sum_{i=1}^4 \sigma_i^2} = \frac{2^2 + 1^2}{4^2 + 2^2 + 1^2 + 0^2} = \frac{5}{21}$$

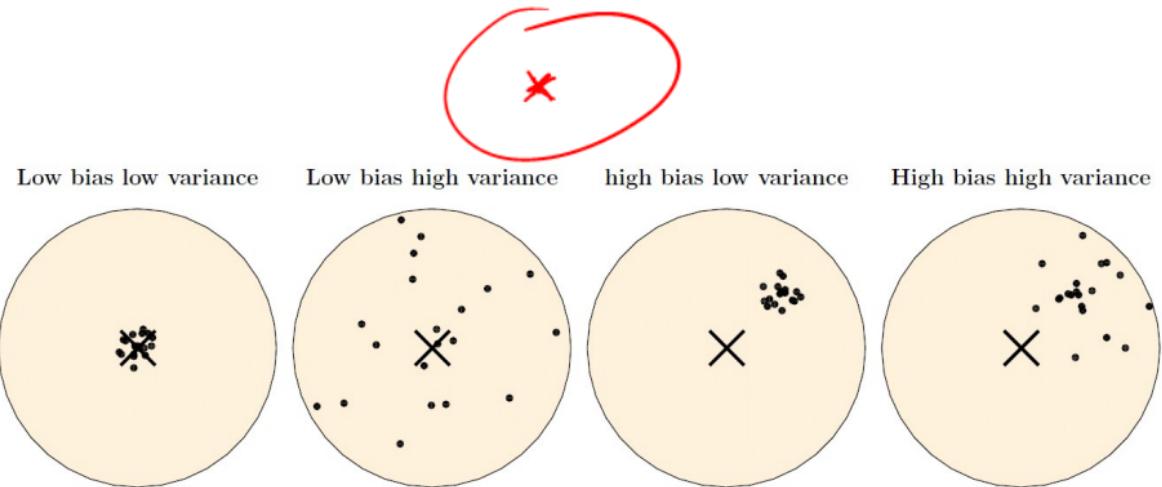
Question 10:

Consider the classification problem given in Figure 5 where x_1 and x_2 are used as features for a logistic regression classifier and a decision tree. The considered logistic regression models all include the constant term w_0 . Which one of the following statements is wrong?

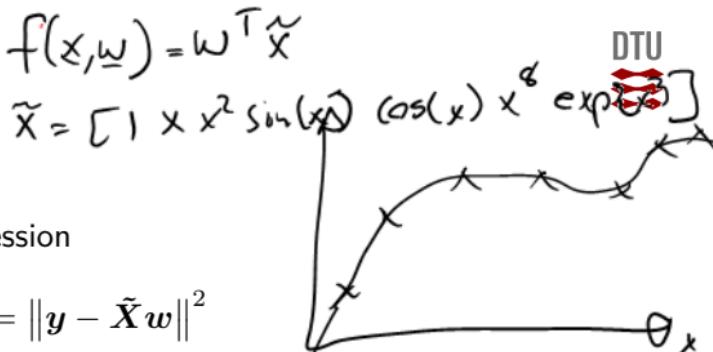


- A: The two classes can be perfectly separated by a logistic regression model using x_1 and x_2 as features.
- B: A decision tree with less than five nodes can perfectly separate the classes using only x_1 and x_2 as features. *(This statement is crossed out with red ink.)*
- C: A logistic regression model can perfectly separate the two classes using only the feature t given by $t = x_1 + x_2$.
- D: In logistic regression the probability that each observation belongs to the two classes can be derived from the logistic function.
- E: Don't know.

What is bias and what is variance?



Regularized least squares



- Recall cost function from linear regression

$$E(w) = \|y - \tilde{X}w\|^2$$

- A parsimonious model can be obtained by **forcing** parameters towards zero.
- Problem: Columns of X have very different scale (i.e. require large/small values of w)
- Therefore, standardize X :

$$\hat{X}_{ij} = \frac{X_{ij} - \mu_j}{\hat{s}_j}, \quad \mu_j = \frac{1}{N} \sum_{i=1}^N X_{kj}, \quad \hat{s}_j = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (X_{ij} - \mu_j)^2}$$

- Note \hat{X} contains no constant term.

$$E(\omega) = \|\mathbf{y} - \tilde{\mathbf{x}}\omega\|_2^2 = \sum_{i=1}^N (y_i - f(\tilde{x}_i, \omega))^2$$

- Introduce regularization term $\lambda\|\mathbf{w}\|^2$ to penalize large weights:

$$E_\lambda(\mathbf{w}, w_0) = \sum_{i=1}^N (y_i - w_0 - \hat{\mathbf{x}}^\top \mathbf{w})^2 + \lambda\|\mathbf{w}\|^2 = \left\| \mathbf{y} - w_0 \mathbf{1} - \hat{\mathbf{X}}\mathbf{w} \right\|^2 + \lambda\|\mathbf{w}\|^2$$
$$\|\mathbf{w}\|^2 = w_1^2 + w_2^2 + \dots + w_M^2$$

- We can solve for w_0 and \mathbf{w} :

$$\frac{dE_\lambda}{dw_0} = \sum_{i=1}^N -2(y_i - w_0 - \hat{\mathbf{x}}_i^\top \mathbf{w}) = -2N\mathbb{E}[y] - 2Nw_0 - N \left(\frac{1}{N} \sum_{i=1}^N \hat{\mathbf{x}}_i^\top \right) \mathbf{w}$$
$$\Rightarrow w_0 = \mathbb{E}[y]$$

- With $\hat{y}_i = y_i - \mathbb{E}[y]$

$$E_\lambda = \left\| \hat{\mathbf{y}} - \hat{\mathbf{X}}\mathbf{w} \right\|^2 + \lambda\|\mathbf{w}\|^2$$

- Setting the derivative wrt. \mathbf{w} equal to zero and solving for \mathbf{w} yields

$$\mathbf{w}^* = (\hat{\mathbf{X}}^\top \hat{\mathbf{X}} + \lambda \mathbf{I}) \backslash (\hat{\mathbf{X}}^\top \hat{\mathbf{y}})$$

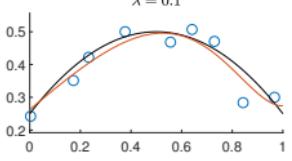
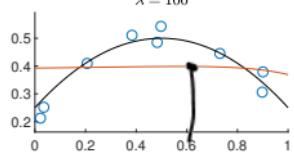
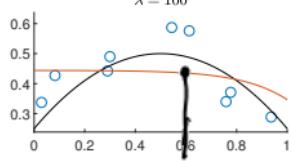
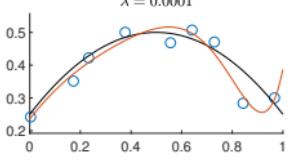
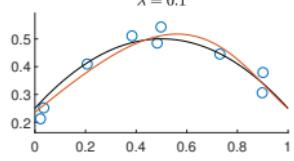
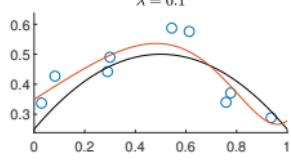
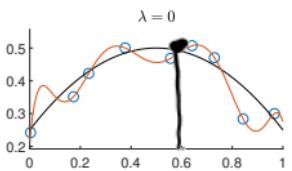
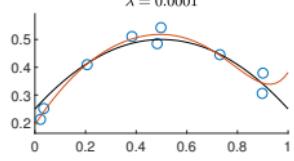
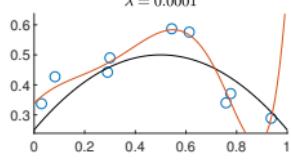
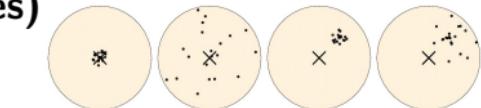
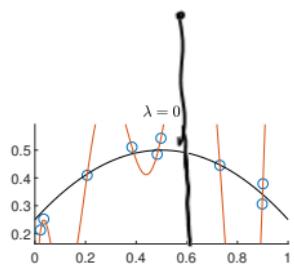
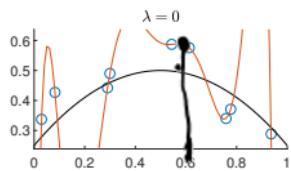
Selecting λ

- Suppose

$$\boldsymbol{w}^* = (\hat{\mathbf{X}}^\top \hat{\mathbf{X}} + \lambda \mathbf{I}) \backslash (\hat{\mathbf{X}}^\top \hat{\mathbf{y}}) \propto \frac{Xy}{X^2 + \lambda}$$

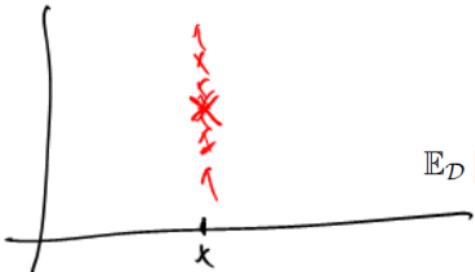
- So if $\lambda = 0$ then no effect, else if $\lambda \rightarrow \infty$ then $\boldsymbol{w}^* \rightarrow 0$
- λ controls complexity of model. Select λ using cross-validation

How does different values of λ (vertical) affect the bias/variance of learned function (red lines)



The Bias-Variance decomposition

$$(a+b)^2 = a^2 + b^2 + 2ab$$

$$\mathbb{E}_{\mathcal{D}} [E^{\text{gen}}] = \mathbb{E}_{\mathcal{D}, (\mathbf{x}, y)} [(y - f_{\mathcal{D}}(\mathbf{x}))^2] = \mathbb{E}_{\mathcal{D}} \left\{ \mathbb{E}_{(\mathbf{x}, y)} \left\{ L(y, f_{\mathcal{D}}(\mathbf{x})) \right\} \right\}$$

We first consider \mathbf{x} fixed

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}, y|\mathbf{x}} [(y - f_{\mathcal{D}}(\mathbf{x}))^2] \\ &= \mathbb{E}_{\mathcal{D}, y|\mathbf{x}} \left[(y - \bar{y}(\mathbf{x}) + \bar{y}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right] \\ &= \mathbb{E}_{y|\mathbf{x}} [(y - \bar{y}(\mathbf{x}))^2] + \mathbb{E}_{\mathcal{D}} [(\bar{y}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2] + 2\mathbb{E}_{\mathcal{D}, y|\mathbf{x}} [(y - \bar{y}(\mathbf{x})) (\bar{y}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))] \end{aligned}$$

$$\bar{y}(\mathbf{x}) = \underline{\mathbb{E}_{y|\mathbf{x}} [y]}$$

$$\mathbb{E}_{\mathcal{D}} \left\{ \mathbb{E}_{y|\mathbf{x}} \left\{ (y - \bar{y}(\mathbf{x})) (\bar{y}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x})) \right\} \right\}$$



The Bias-Variance decomposition

$$\mathbb{E}_{\mathcal{D}} [E^{\text{gen}}] = \mathbb{E}_{\mathcal{D},(\mathbf{x},y)} \left[(y - f_{\mathcal{D}}(\mathbf{x}))^2 \right]$$

We first consider \mathbf{x} fixed

$$\begin{aligned} & \mathbb{E}_{\mathcal{D},y|\mathbf{x}} \left[(y - f_{\mathcal{D}}(\mathbf{x}))^2 \right] & \bar{y}(\mathbf{x}) = \mathbb{E}_{y|\mathbf{x}} [y] \\ &= \mathbb{E}_{\mathcal{D},y|\mathbf{x}} \left[(y - \bar{y}(\mathbf{x}) + \bar{y}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right] \\ &= \mathbb{E}_{y|\mathbf{x}} \left[(y - \bar{y}(\mathbf{x}))^2 \right] + \mathbb{E}_{\mathcal{D}} \left[(\bar{y}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right] + \cancel{2\mathbb{E}_{\mathcal{D},y|\mathbf{x}} [(y - \bar{y}(\mathbf{x})) (\bar{y}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))]} \end{aligned}$$

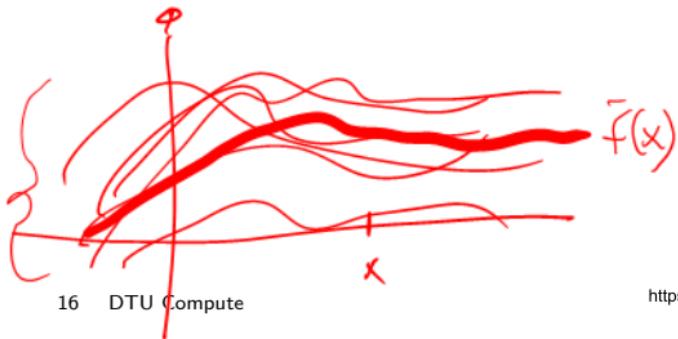


The Bias-Variance decomposition

$$\mathbb{E}_{\mathcal{D}, y|\mathbf{x}} \left[(y - f_{\mathcal{D}}(\mathbf{x}))^2 \right] = \mathbb{E}_{y|\mathbf{x}} \left[(y - \bar{y}(\mathbf{x}))^2 \right] + \mathbb{E}_{\mathcal{D}} \left[(\bar{y}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right]$$

$$\bar{f}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}} [f_{\mathcal{D}}(\mathbf{x})]$$

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}} \left[(\bar{y}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right] \\ &= \mathbb{E}_{\mathcal{D}} \left[(\bar{y}(\mathbf{x}) - \bar{f}(\mathbf{x}) + \bar{f}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right] \\ &= \mathbb{E}_{\mathcal{D}} \left[(\bar{y}(\mathbf{x}) - \bar{f}(\mathbf{x}))^2 \right] + \mathbb{E}_{\mathcal{D}} \left[(\bar{f}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right] + 2\mathbb{E}_{\mathcal{D}} \left[(\bar{y}(\mathbf{x}) - \bar{f}(\mathbf{x})) (\bar{f}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x})) \right] \end{aligned}$$



The Bias-Variance decomposition



$$\mathbb{E}_{\mathcal{D}, y|\mathbf{x}} \left[(y - f_{\mathcal{D}}(\mathbf{x}))^2 \right] = \mathbb{E}_{y|\mathbf{x}} \left[(y - \bar{y}(\mathbf{x}))^2 \right] + \mathbb{E}_{\mathcal{D}} \left[(\bar{y}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right]$$

$$\bar{f}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}} [f_{\mathcal{D}}(\mathbf{x})]$$

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}} \left[(\bar{y}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right] \\ &= \mathbb{E}_{\mathcal{D}} \left[(\bar{y}(\mathbf{x}) - \bar{f}(\mathbf{x}) + \bar{f}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right] \\ &= \mathbb{E}_{\mathcal{D}} \left[(\bar{y}(\mathbf{x}) - \bar{f}(\mathbf{x}))^2 \right] + \mathbb{E}_{\mathcal{D}} \left[(\bar{f}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right] + \cancel{2\mathbb{E}_{\mathcal{D}} \left[(\bar{y}(\mathbf{x}) - \bar{f}(\mathbf{x})) (\bar{f}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x})) \right]} \end{aligned}$$

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}, y|\mathbf{x}} \left[(y - f_{\mathcal{D}}(\mathbf{x}))^2 \right] \\ &= \mathbb{E}_{y|\mathbf{x}} \left[(y - \bar{y}(\mathbf{x}))^2 \right] + (\bar{y}(\mathbf{x}) - \bar{f}(\mathbf{x}))^2 + \mathbb{E}_{\mathcal{D}} \left[(\bar{f}(\mathbf{x}) - f_{\mathcal{D}}(\mathbf{x}))^2 \right] \\ &= \text{Var}_{y|\mathbf{x}} [y] + (\bar{y}(\mathbf{x}) - \bar{f}(\mathbf{x}))^2 + \text{Var}_{\mathcal{D}} [f_{\mathcal{D}}(\mathbf{x})] \end{aligned}$$

The Bias-Variance decomposition

$$\mathbb{E}_{\mathcal{D}} [E^{\text{gen}}] = \mathbb{E}_{\mathbf{x}} \left[\mathbb{E}_{\mathcal{D}, y|\mathbf{x}} \left[(y - f_{\mathcal{D}}(\mathbf{x}))^2 \right] \right]$$
$$\mathbb{E}_{\mathcal{D}} [E^{\text{gen}}] = \mathbb{E}_{\mathbf{x}} \left[\text{Var}_{y|\mathbf{x}} [y] + (\bar{y}(\mathbf{x}) - \bar{f}(\mathbf{x}))^2 + \text{Var}_{\mathcal{D}} [f_{\mathcal{D}}(\mathbf{x})] \right]$$



[https://commons.wikimedia.org/wiki/File:Frustrated_man_at_a_desk_\(cropped\).jpg](https://commons.wikimedia.org/wiki/File:Frustrated_man_at_a_desk_(cropped).jpg)

The Bias-Variance decomposition



$$\mathbb{E}_{\mathcal{D}} [E^{\text{gen}}] = \mathbb{E}_{\mathbf{x}} \left[\text{Var}_{y|\mathbf{x}} [y] + (\bar{y}(\mathbf{x}) - \bar{f}(\mathbf{x}))^2 + \text{Var}_{\mathcal{D}} [f_{\mathcal{D}}(\mathbf{x})] \right]$$

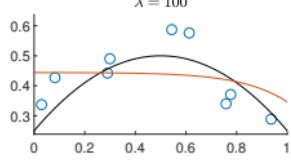
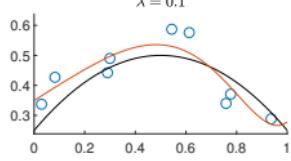
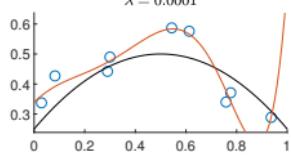
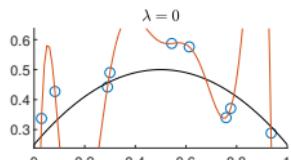
The first term does not depend at all upon our choice of model but simply represents the intrinsic difficulty of the problem. We cannot make this term any larger or smaller by selecting one model over another.

The second term is the **bias** term. It tells us how much the average values of models trained on different training datasets differ compared to the true mean of the data.

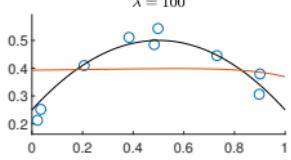
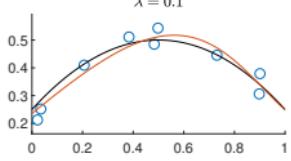
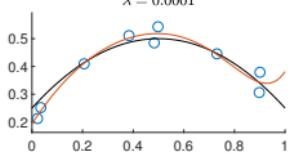
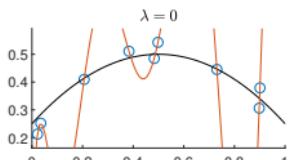
The third term is the **variance** term. It tells us how much the model wiggles when trained on different sets of training data. That is, when you train the models on N different (random) sets of training data and the models (the prediction curves) are nearly the same this term is small.

The bias variance decomposition

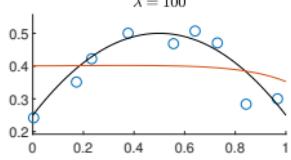
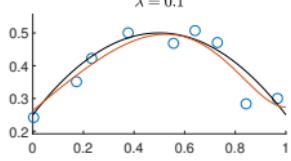
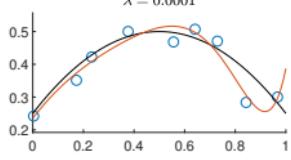
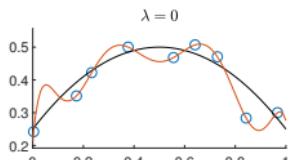
Dataset 1



Dataset 2



Dataset 3



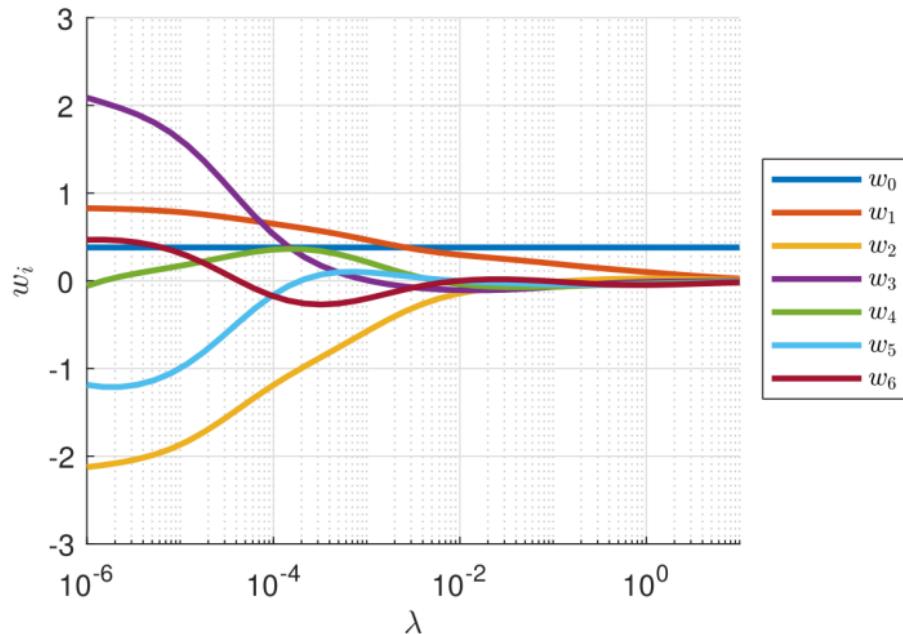
Low bias
High var

Low var
High bias

By regularization we can tradeoff bias and variance, in particular, we can hope to substantially reduce variance without introducing too much bias!

Parameters w^* as function of λ

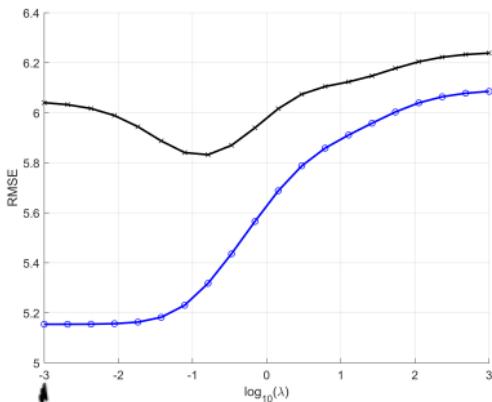
$$E_\lambda(\mathbf{w}) = \sum_{i=1}^N (\hat{y}_i - w_0 - \hat{\mathbf{x}}_i^\top \mathbf{w})^2 + \lambda \|\mathbf{w}\|^2$$



$$\underline{w}^\top \underline{w} = \|\underline{w}\|^2$$

14:13

Quiz 1, Bias-variance (Fall 2017)



Using 54 observations of a dataset about Basketball, we would like to predict the average points scored per game (y) based on the four features. For this purpose we consider regularized least squares regression which minimizes with respect to \mathbf{w} the following cost function:

$$E(\mathbf{w}) = \sum_n (y_n - [1 \ x_{n1} \ x_{n2} \ x_{n3} \ x_{n4}] \mathbf{w})^2 + \lambda \mathbf{w}^\top \mathbf{w},$$

We consider 20 different values of λ and use leave-

one-out cross-validation to estimate the performance (measured by mean-squared error) of each of these different values of λ and plot the result in the figure. For the value of $\lambda = 0.6952$ the following model is identified:

$$f(\mathbf{x}) = 2.76 - 0.37x_1 + 0.01x_2 + 7.67x_3 + 7.67x_4.$$

Which one of the following statements is correct?

- A. In the figure the blue curve with circles corresponds to the training error whereas the black curve with crosses corresponds to the test error.
- B. According to the model defined for $\lambda = 0.6952$ increasing a players height x_1 will increase his average points scored per game.
- C. There is no optimal way of choosing λ since increasing λ reduces the variance but increases the bias.
- D. As we increase λ the 2-norm of the weight vector \mathbf{w} will also increase.
- E. Don't know.

The correct answer is A: The blue curve monotonically increases with λ reflecting a worse fit to the training set as we increase λ using regularization we can reduce the variance by introducing bias and the black curve indicates that an optimal tradeoff at around $10^{-0.8}$ as reflected by the test error indicated in the black curve being minimal. As we increase λ we will

penalize the weights according to the squared 2-norm more and more and thus the 2-norm will be reduced. Finally, according to the fitted model we observe that the coefficient in front of x_1 (Height) is negative thus indicating that an increase in height will reduce the models prediction of average points scored per game.

General linear model

- Remember the generalized linear model?

- Data

$$\{\mathbf{x}_n, y_n\}_{n=1}^N$$

- Model

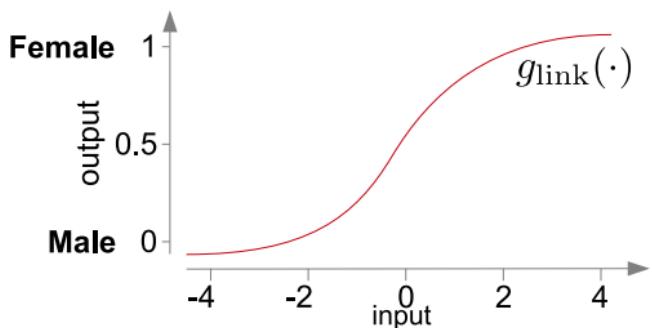
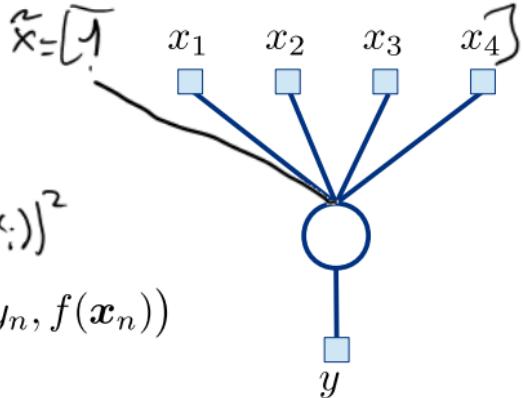
$$f(\mathbf{x}) = \underbrace{g_{\text{link}}(\mathbf{x}^\top \mathbf{w})}$$

- Cost function

$$d(y, f(\mathbf{x})) = \frac{1}{N} \sum_i (y_i - f(\mathbf{x}_i))^2$$

- Parameters

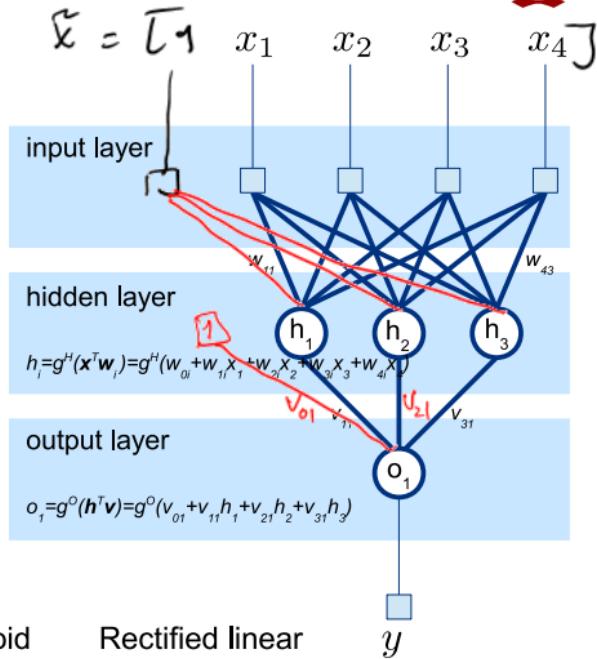
$$\mathbf{w} = \arg \min_{\mathbf{w}} \sum_{n=1}^N d(y_n, f(\mathbf{x}_n))$$



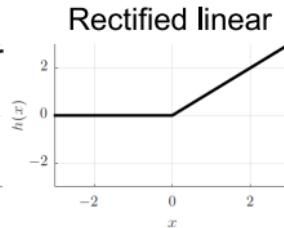
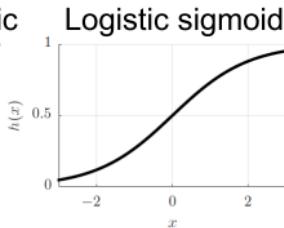
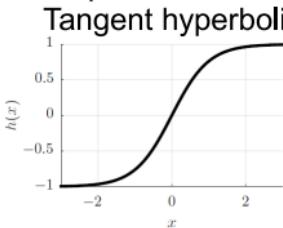
Artificial neural networks

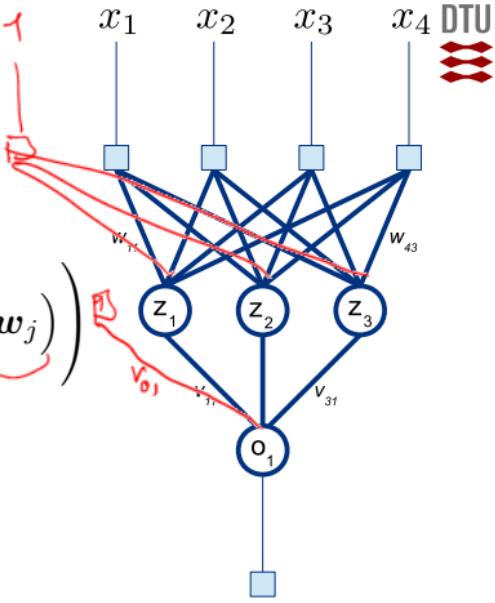
Feed forward network

- Each “neuron”
 - Computes a non-linear function of the sum of its inputs
 - Is just like a generalized linear model
 - Has its own set of parameters
- Modeling choices
 - Cost function
 - Non-linearities
 - Number of neurons and hidden layers
 - Selection of inputs
- Parameter estimation using numerical optimization methods
- Very flexible model: Can easily overfit



Example of non-linearities:





Data: $\{\mathbf{x}_i, y_i\}$

Model: $f(\mathbf{x}) = h^{(2)} \left(v_{01} + \sum_{j=1}^H v_{1j} h^{(1)} \left(\tilde{\mathbf{x}}^\top \mathbf{w}_j \right) \right)$

Distance: $d(y, f(\mathbf{x}))$

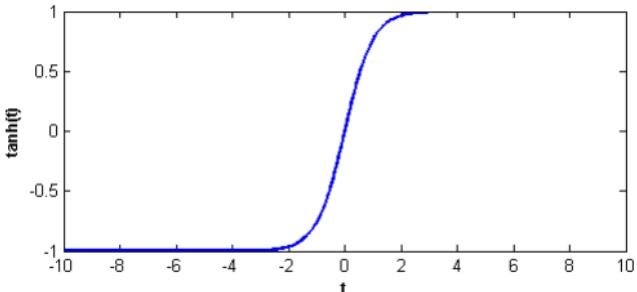
Cost: $E = \sum_{i=1}^N d(y_i, f(\mathbf{x}_i))$

Common choices

$$h^{(1)}(x) = \tanh(x)$$

$$h^{(2)}(x) = x$$

$$d(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$$



$$f(\mathbf{x}, \mathbf{w}) = h^{(2)}\left(\mathbf{\tilde{w}}^{(2)}[1, h^{(1)}(\mathbf{\tilde{w}}^{(1)}\mathbf{\tilde{x}})]\right)$$

Neurons and layers

Recall:

$$f(\mathbf{x}) = h^{(2)} \left(v_{01} + \sum_{j=1}^H v_{1j} h^{(1)} \left(\tilde{\mathbf{x}}^\top \mathbf{w}_j \right) \right)$$

- Let $z_j^{(1)}$ be output of j 'th hidden unit

$$\mathbf{\tilde{w}}^{(1)} = \begin{bmatrix} -w_1 \\ -w_2 \\ \vdots \end{bmatrix}$$

$$z_j^{(1)} = h^{(1)} \left(\mathbf{w}_j^{(1) \top} \tilde{\mathbf{x}} \right)$$

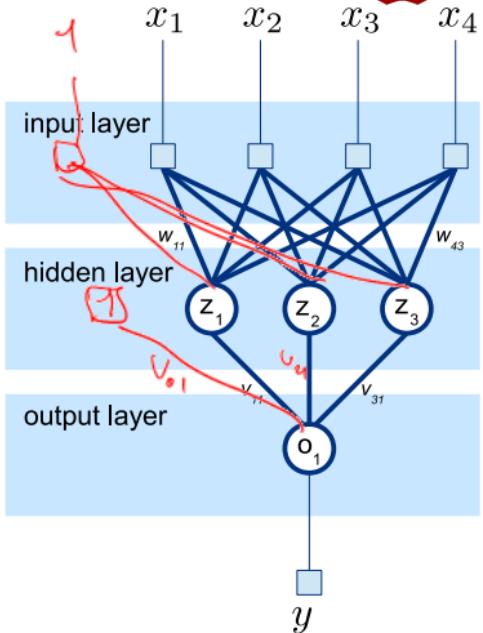
Abbreviated $\mathbf{z}^{(1)} = h^{(1)} \left(\mathbf{W}^{(1)} \tilde{\mathbf{x}} \right)$

- Output

$$f(\mathbf{x}) = h^{(2)} \left(v_{01} + \sum_{j=1}^H v_{1j} z_j^{(1)} \right) = h^{(2)} \left(\mathbf{W}^{(2)} \mathbf{\tilde{z}}^{(1)} \right)$$

$$\mathbf{\tilde{w}}^{(2)} = \begin{bmatrix} -w_1 \\ -w_2 \\ \vdots \end{bmatrix}$$

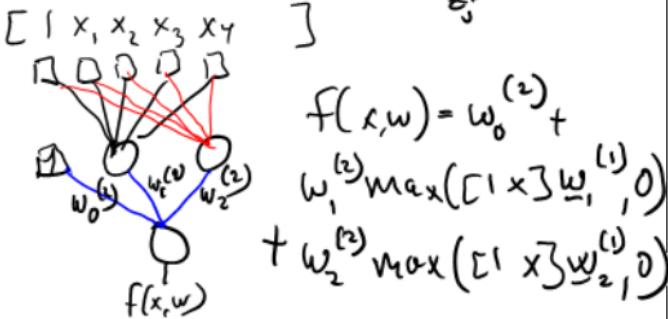
We consider each $z_j^{(1)}$ a neuron and $\mathbf{z}^{(1)}$ a (hidden) layer



Quiz 2, Artificial Neural Network (Fall 2017)

We will consider an artificial neural network (ANN) trained to predict the average score of a player (i.e., y). The ANN is based on the model:

$$f(\mathbf{x}, \mathbf{w}) = w_0^{(2)} + \sum_{j=1}^2 w_j^{(2)} h^{(1)}([1 \ \mathbf{x}] \mathbf{w}_j^{(1)}).$$



where $h^{(1)}(x) = \max(x, 0)$ is the rectified linear function used as activation function in the hidden layer (i.e., positive values are returned and negative values are set to zero). We will consider an ANN with two hidden units in the hidden layer defined by:

$$\mathbf{w}_1^{(1)} = \begin{bmatrix} 21.78 \\ -1.65 \\ 0 \\ -13.26 \\ -8.46 \end{bmatrix}, \quad \mathbf{w}_2^{(1)} = \begin{bmatrix} -9.60 \\ -0.44 \\ 0.01 \\ 14.54 \\ 9.50 \end{bmatrix},$$

and $w_0^{(2)} = 2.84$, $w_1^{(2)} = 3.25$, and $w_2^{(2)} = 3.46$.

What is the predicted average score of a basketball player with observation vector $\mathbf{x}^* = [6.8 \ 225 \ 0.44 \ 0.68]$?

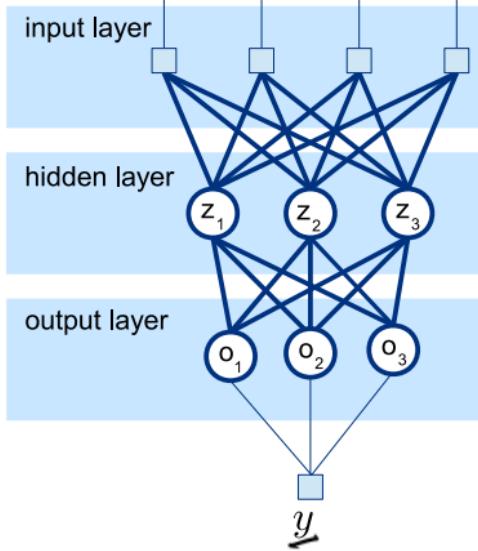
- A. 1.00
- B. 3.74
- C. 8.21
- D. 11.54
- E. Don't know.

The output is given by:

$$f(\mathbf{x}, \mathbf{w}) = 2.84$$

$$\begin{aligned} & + 3.25 \cdot \max([1 \ 6.8 \ 225 \ 0.44 \ 0.68] \cdot \begin{bmatrix} 21.78 \\ -1.65 \\ 0 \\ -13.26 \\ -8.46 \end{bmatrix}, 0) \\ & + 3.46 \max([1 \ 6.8 \ 225 \ 0.44 \ 0.68] \cdot \begin{bmatrix} -9.60 \\ -0.44 \\ 0.01 \\ 14.54 \\ 9.50 \end{bmatrix}, 0) \\ & = 2.84 + 3.25 \cdot \max(-1.027, 0) + 3.46 \max(2.516, 0) \\ & = 11.54 \end{aligned}$$

Generalization 1: Multiple outputs



- As before define: $\underline{z}^{(1)} = h^{(1)} \left(\mathbf{W}^{(1)} \tilde{\underline{x}} \right)$
- Now let $\mathbf{W}^{(2)}$ be a $C \times H$ matrix then:

$$\underline{y} = \mathbf{f}(\underline{x}) = h^{(2)} \left(\mathbf{W}^{(2)} \tilde{\underline{z}}^{(1)} \right)$$

will be \underline{C} -dimensional

- Re-define error function

$$E = \underbrace{\sum_{i=1}^N \|\underline{y}_i - \mathbf{f}(\underline{x}_i)\|_2^2}_{\text{ }}$$

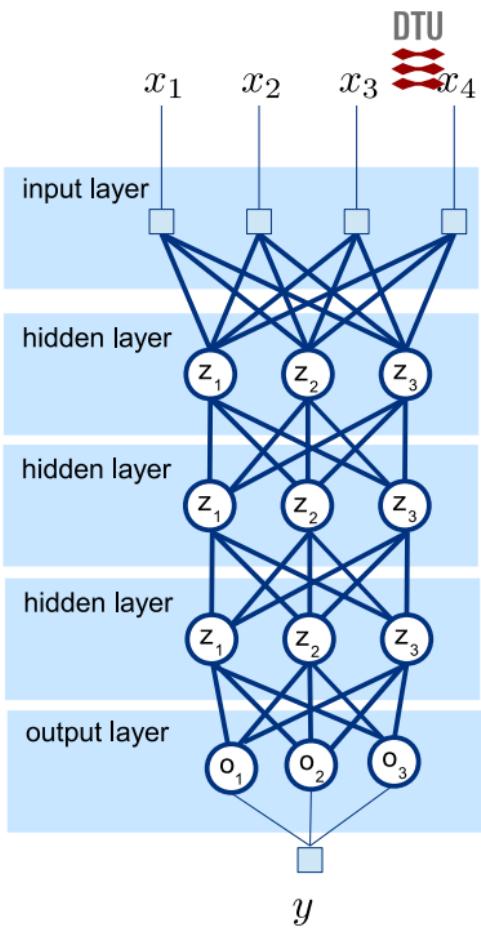
Generalization 2: Multiple layers

- Define $z^{(0)} = x$
- For each layer $l = 1, \dots, L$ compute

$$z_j^{(l)} = h^{(l)} \left(\mathbf{W}^{(l)} \tilde{z}^{(l-1)} \right)$$

- Output is simply

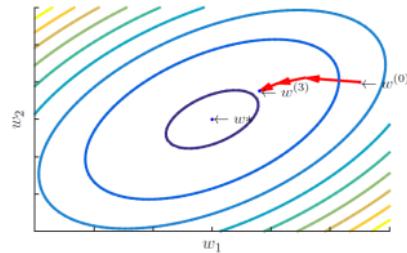
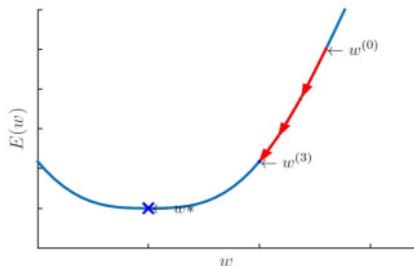
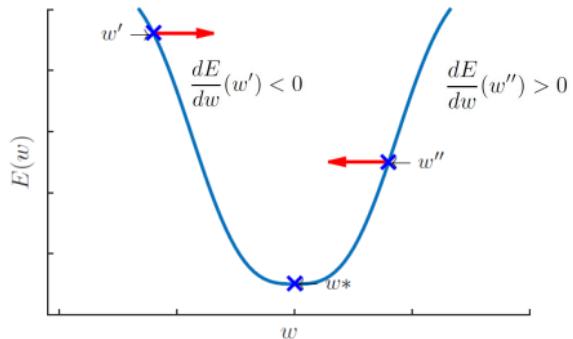
$$\mathbf{f}(x) = z^{(L)}$$



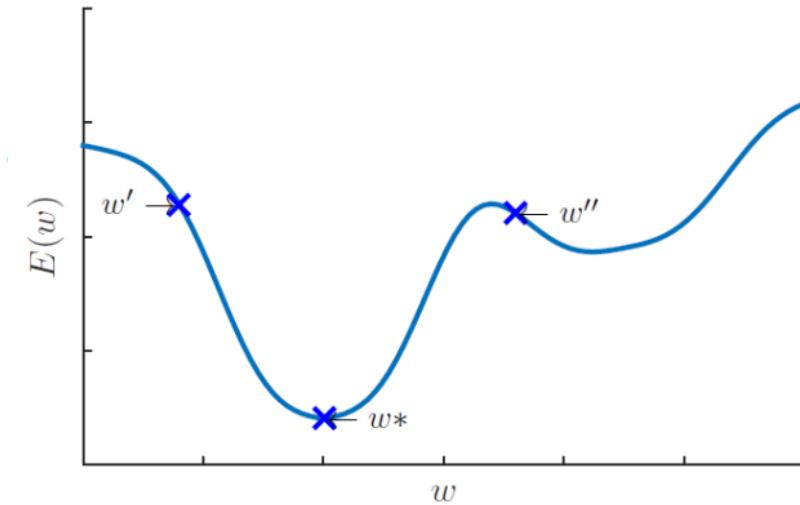
Gradient descent

- Start from an initial guess at \mathbf{w}^* , $\mathbf{w}^{(0)}$
- At step t of the algorithm, modify $\mathbf{w}^{(t-1)}$ to produce a better guess $\mathbf{w}^{(t)}$:

$$\mathbf{w}^{(t)} = \mathbf{w}^{(t-1)} - \epsilon \frac{dE}{d\mathbf{w}}(\mathbf{w}^{(t-1)})$$



Contrary to least-squares linear regression and logistic regression ANNs have issues of local minima



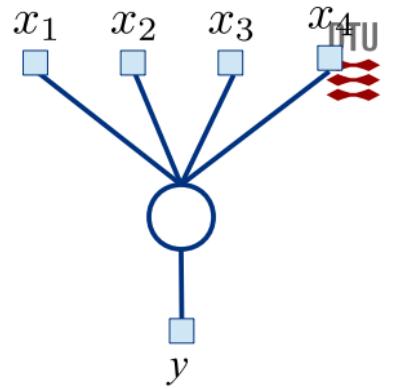
Single and multi-class: One out of K coding

Nationality

TXT=

| | | Denmark | Norway | Sweden |
|-----------|--------|---------|--------|--------|
| 'Sweden' | X_tmp= | 0 | 0 | 1 |
| 'Sweden' | | 0 | 0 | 1 |
| 'Sweden' | | 0 | 0 | 1 |
| 'Sweden' | | 0 | 0 | 1 |
| 'Norway' | | 0 | 0 | 1 |
| 'Norway' | | 0 | 1 | 0 |
| 'Norway' | | 0 | 1 | 0 |
| 'Norway' | | 0 | 1 | 0 |
| 'Norway' | | 0 | 1 | 0 |
| 'Norway' | | 0 | 1 | 0 |
| 'Norway' | | 0 | 1 | 0 |
| 'Denmark' | | 1 | 0 | 0 |
| 'Denmark' | | 1 | 0 | 0 |
| 'Sweden' | | 0 | 0 | 1 |
| 'Sweden' | | 0 | 0 | 1 |
| 'Sweden' | | 0 | 0 | 1 |
| 'Denmark' | | 1 | 0 | 0 |
| 'Sweden' | | 0 | 0 | 1 |
| 'Norway' | | 0 | 1 | 0 |
| 'Denmark' | | 1 | 0 | 0 |

One-out-of-K coding



Multi-class classification

- Logistic regression, $y = 0, 1$:

$$p(y|\theta) = \theta^y (1 - \theta)^{1-y}$$

$$\theta = \sigma(\mathbf{x}^\top \mathbf{w})$$

- Multinomial regression, $y = 1, 2, \dots, K$

z_k : one-of- K encoding of y ,

$$p(y|\theta) = \prod_{i=1}^K \theta_k^{z_k}$$

$$\theta = \text{softmax}([\mathbf{x}^\top \mathbf{w}_1 \quad \dots \quad \mathbf{x}^\top \mathbf{w}_K])$$

$$= \begin{bmatrix} \frac{e^{\mathbf{x}^\top \mathbf{w}_1}}{\sum_{c=1}^K e^{\mathbf{x}^\top \mathbf{w}_c}} & \dots & \frac{e^{\mathbf{x}^\top \mathbf{w}_{K-1}}}{\sum_{c=1}^K e^{\mathbf{x}^\top \mathbf{w}_c}} & \frac{e^{\mathbf{x}^\top \mathbf{w}_K}}{\sum_{c=1}^K e^{\mathbf{x}^\top \mathbf{w}_c}} \end{bmatrix}$$

or: $\theta = \begin{bmatrix} \frac{e^{\mathbf{x}^\top \mathbf{w}_1}}{1 + \sum_{c=1}^{K-1} e^{\mathbf{x}^\top \mathbf{w}_c}} & \dots & \frac{e^{\mathbf{x}^\top \mathbf{w}_{K-1}}}{1 + \sum_{c=1}^{K-1} e^{\mathbf{x}^\top \mathbf{w}_c}} & \frac{1}{1 + \sum_{c=1}^{K-1} e^{\mathbf{x}^\top \mathbf{w}_c}} \end{bmatrix} \mathbf{y}_1 \quad \mathbf{y}_2 \quad \mathbf{y}_3$

Connection to neural networks

Multinomial regression:

- Define:

$$\boldsymbol{\theta} = \begin{bmatrix} \frac{e^{\mathbf{x}^\top \mathbf{w}_1}}{1 + \sum_{c=1}^{K-1} e^{\mathbf{x}^\top \mathbf{w}_c}} & \cdots & \frac{1}{1 + \sum_{c=1}^{K-1} e^{\mathbf{x}^\top \mathbf{w}_c}} \end{bmatrix}$$

- Cost function is ($z_{i\cdot}$ is one-of- K encoding of y_i)

$$E = - \sum_{i=1}^N \log p(y_i | \mathbf{x}_i) = - \sum_{i=1}^N \sum_{c=1}^K z_{ic} \log \theta_{ic}$$

Multi-class neural network:

- Suppose $\tilde{y}_1, \dots, \tilde{y}_K$ are outputs of a neural network
- Define

$$\boldsymbol{\theta} = \begin{bmatrix} \frac{e^{\tilde{\mathbf{y}}^\top \mathbf{w}_1}}{\sum_{c=1}^K e^{\tilde{\mathbf{y}}^\top \mathbf{w}_c}} & \cdots & \frac{e^{\tilde{\mathbf{y}}^\top \mathbf{w}_K}}{\sum_{c=1}^K e^{\tilde{\mathbf{y}}^\top \mathbf{w}_c}} \end{bmatrix}$$

- Cost function is:

$$E = - \sum_{i=1}^N \log p(y_i | \tilde{\mathbf{y}}_i) = - \sum_{i=1}^N \sum_{c=1}^K z_{ic} \log \theta_{ic}$$

Quiz 3, Multinomial Regression (Spring 2016)

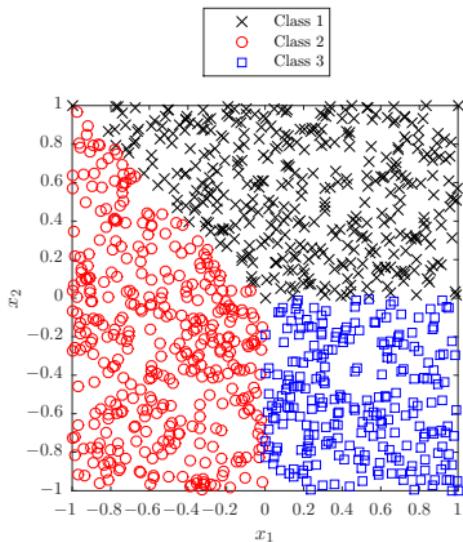


Figure 1: Observations labelled with the most probable class

Consider a multinomial regression classifier for

a three-class problem where for each point $\mathbf{x} = [x_1 \ x_2]^\top$ we compute the class-probability using the softmax function

$$P(\hat{y} = k) = \frac{e^{\mathbf{w}_k^\top \mathbf{x}}}{e^{\mathbf{w}_1^\top \mathbf{x}} + e^{\mathbf{w}_2^\top \mathbf{x}} + e^{\mathbf{w}_3^\top \mathbf{x}}}.$$

A dataset of $N = 1000$ points where each point is labeled according to the maximum class-probability is shown in Figure 1. Which setting of the weights was used?

- A. $w_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $w_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $w_3 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$
- B. $w_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$, $w_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $w_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$
- C. $w_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $w_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $w_3 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$
- D. $w_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $w_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $w_3 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$
- E. Don't know.

Consider for instance the point \mathbf{x} where $x_1 = 0$ and $x_2 = 1$. Then, letting $y_k = \mathbf{w}_k^T \mathbf{x}$, we obtain:

$$A : [y_1 \quad y_2 \quad y_3] = [-1 \quad 1 \quad -1]$$

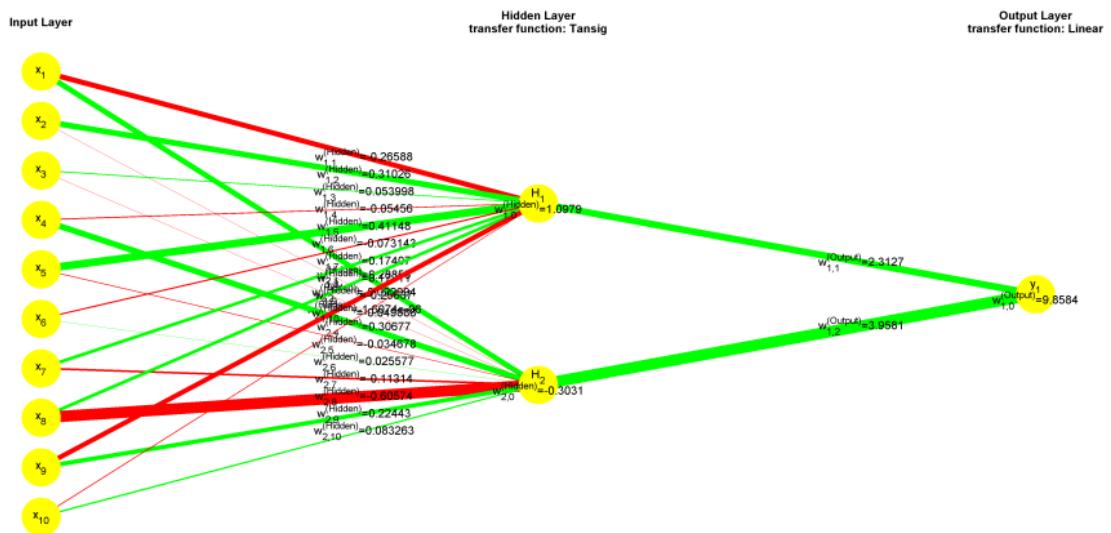
$$B : [y_1 \quad y_2 \quad y_3] = [-1 \quad -1 \quad 1]$$

$$C : [y_1 \quad y_2 \quad y_3] = [1 \quad -1 \quad -1]$$

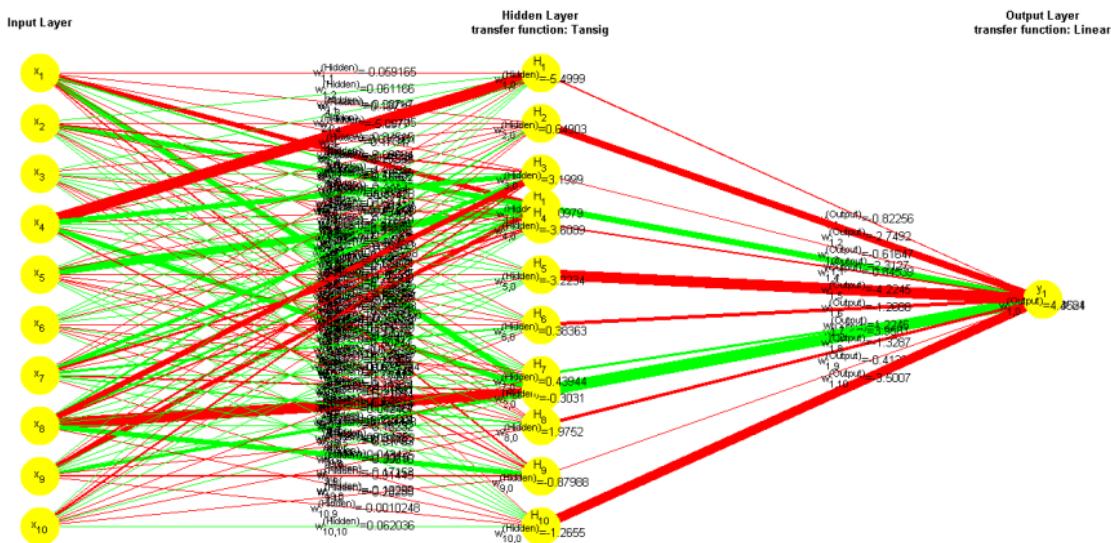
$$D : [y_1 \quad y_2 \quad y_3] = [-1 \quad 1 \quad 1]$$

Next, since the multinomial regression function preserves order we need only consider the maximal value. Accordingly the point \mathbf{x} is only classified to the correct class 1 for option C .

Interpreting neural networks can be difficult



Interpreting neural networks can be difficult



Resources

<https://www.youtube.com> Excellent video resource explaining the concepts behind neural networks

(https://www.youtube.com/watch?v=aircAruvnKk&list=PLZHQQb0WTQDNU6R1_67000Dx_ZCJB-3pi)

<http://playground.tensorflow.org> Sleek interactive neural network example where you can examine the effect of different number of hidden neurons, activation functions, and many other things on training (<http://playground.tensorflow.org/>)

<https://www.tensorflow.org> Most popular and well-documented deep learning framework. While well documented, notice it requires some python knowledge (<https://www.tensorflow.org/>)

<https://pytorch.org> Upcoming (and in some ways slightly simpler) framework for deep learning; alternative to tensorflow
(<https://pytorch.org/>)

Mid-term quiz 1

In the analysis of house prices the following attributes were collected for a house: The year the house was built (denoted YEAR), the size of the house given in square meters (denoted SIZE) the county in which the house is located (denoted LOCATION). Which statement about the three attributes is correct?

- A. YEAR is ratio, SIZE is interval and LOCATION is nominal
- B. YEAR is interval, SIZE is ratio and LOCATION is nominal
- C. YEAR is interval, SIZE is ratio and LOCATION is ordinal
- D. YEAR is interval, SIZE is ratio and LOCATION is interval
- E. Don't know.

Year data types do not have a zero with a physical meaning and are therefore interval. Size has a physically relevant zero and is therefore ratio. Mean-

while, location is just an identifier which only support similarity-comparison and is therefore nominal.

Mid-term quiz 2

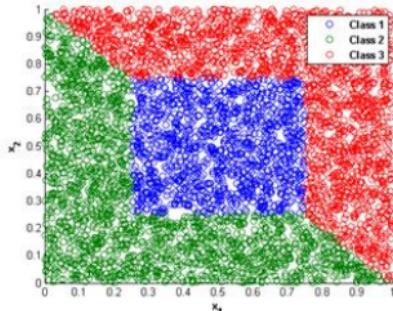
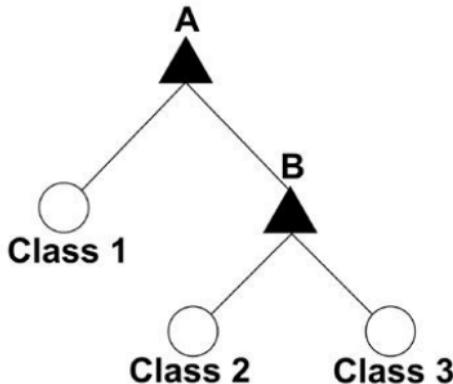


Figure 1



Consider the classification problem given in figure 1 and the Decision Tree shown below it with two decision nodes denoted A and B. We will let $\mathbf{x}_n = (x, y)$ denote a 2-dimensional observation such that $\mathbf{x}_n - 0.5 \cdot \mathbf{1}$ denotes the subtraction of 0.5 from each of the two coordinates of \mathbf{x}_n . Which one of the following classification rules would lead to a correct classification of the data?

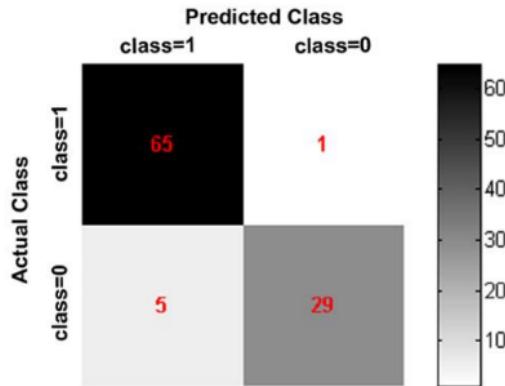
- A. A: $\|\mathbf{x}_n - 0.5 \cdot \mathbf{1}\|_1 \leq 0.25$, B : $\|\mathbf{x}_n\|_\infty \leq 1$
- B. A: $\|\mathbf{x}_n\|_1 \leq 1$, B: $\|\mathbf{x}_n - 0.5 \cdot \mathbf{1}\|_2 \leq \infty$
- C. A: $\|\mathbf{x}_n - 0.5 \cdot \mathbf{1}\|_2 \leq 0.25$, B : $\|\mathbf{x}_n\|_\infty \leq 1$
- D. A: $\|\mathbf{x}_n - 0.5 \cdot \mathbf{1}\|_\infty \leq 0.25$, B : $\|\mathbf{x}_n\|_1 \leq 1$
- E. Don't know.

The right answer is *D*. Recall the shape associated with the different L_p -norms: $p = 2$ is a circle (Euclidean distance), $p = \infty$ a square, and $p = 1$ a square rotated 45 degrees. If we therefore first ask at *A* if the observation is within the square (if yes, classify as the

blue class, otherwise go to next split) and then at the next split ask if it within the L_1 -norm of origo (distance 1), we get the diagonal decision boundary. This can be implemented by option *D*.

Mid-term quiz 3

A classifier has the confusion matrix given in the figure below. Which statement about the classifier is correct?



- A. The Accuracy is 94% and the Error rate is 6%
- B. The Accuracy is 6% and the Error rate is 94%
- C. The Accuracy is 65% and the Error rate is 35%
- D. There is insufficient information in the confusion matrix to determine the Accuracy and Error rate.
- E. Don't know.

Accuracy is total number of correct choices divided by total number of observations. Therefore, the ac-

curacy is $\frac{65+29}{6+65+29} = \frac{94}{100}$ or 95%. The right answer is therefore A.

Mid-term quiz 4

Which statement about crossvalidation is wrong?

- A. Cross-validation can be used to estimate the generalization error.
- B. Leave one out cross-validation is more computationally expensive than 10 fold crossvalidation.
- C. Holding out one third of the data for validation is faster but less accurate than performing 10 fold cross-validation.
- D. The same test set can be used for model selection as well as evaluation of the generalization performance of the model.
- E. Don't know.

The last option D is wrong because if we both select a model on a test set and then later use it for estimating the generalization error we will not obtain

an unbiased estimate of the generalization error since we have already tuned the model on the test set. For this task, one should use two-layer CV.

Mid-term quiz 5

Consider a data set of four features: A , B , C , and D that are applied in a classification algorithm. The table below shows the cross-validated Error rate when using different combinations of the features.

| Feature(s) | Error rate |
|---------------------|------------|
| A | 0.40 |
| B | 0.45 |
| C | 0.33 |
| D | 0.42 |
| A and B | 0.20 |
| A and C | 0.25 |
| A and D | 0.34 |
| B and C | 0.29 |
| B and D | 0.42 |
| C and D | 0.40 |
| A and B and C | 0.13 |
| A and B and D | 0.17 |
| B and C and D | 0.10 |
| A and C and D | 0.15 |
| A and B and C and D | 0.28 |

We will apply a forward feature selection algorithm. Which feature set will the selection algorithm choose?

- A. C
- B. B and C and D
- C. A and B
- D. A and B and C
- E. Don't know.

Forward selection will attempt to minimize the error rate. It will first select C , then select lowest

of the next options containing C , i.e. A, C , and then A, B, C . Therefore, option D is correct.

Mid-term quiz 6

When training a decision tree we will use the classification error as impurity measure $I(t)$ given by $I(t) = 1 - \max_i [p(i|t)]$ where $p(i|t)$ denotes the fraction of data objects belonging to class i at a given node t . We will use Hunt's algorithm to grow the tree and recall that the purity gain is given by:

$$\Delta = I(\text{Parent}) - \sum_{j=1}^k \frac{N(v_j)}{N} I(v_j)$$

where N is the total number of data objects at the parent node, k is the number of child nodes and $N(v_j)$ is the number of data objects associated with the child node, v_j . We will consider classification of Iris flowers into Iris-Setosa, Iris-Virginica and Iris-Versicolor. At a potential split we have:

- Before the split: 5 Iris-Setosa, 10 Iris-Virginica and 10 Iris Versicolor.

After the split

- 0 Iris-Setosa, 8 Iris-Virginica and 2 Iris-Versicolor in the left node.
- 5 Iris-Setosa, 2 Iris-Virginica and 8 Iris-Versicolor in the right node.

Which statement is correct?

- A. The purity gain is $\Delta = \frac{3}{5}$
- B. The purity gain is $\Delta = \frac{3}{15}$
- C. The purity gain is $\Delta = \frac{6}{25}$
- D. The purity gain is $\Delta = \frac{7}{15}$
- E. Don't know.

There are a total of $N = 25$ observations and the number in the two branches are $N_1 = 10$ and $N_2 = 15$. In the base branch, the maximum class-probability is $\frac{10}{25}$ and so $I_0 = 1 - \frac{10}{25} = \frac{15}{25} = \frac{3}{5}$. Similarly, we compute

$I_1 = \frac{1}{5}$ and $I_2 = 1 - \frac{8}{15} = \frac{7}{15}$. We now have

$$\Delta = I_0 - \frac{N_1}{N} I_1 - \frac{N_2}{N} I_2 = \frac{3}{5} - \frac{10}{25} \frac{1}{5} - \frac{15}{25} \frac{7}{15} \quad (1)$$

$$= \frac{3}{5} - \frac{225}{25} - \frac{7}{25} = \frac{15 - 2 - 7}{25} = \frac{6}{25} \quad (2)$$

or C .

Mid-term quiz 7

When people are well rested and take an exam their chance of passing the exam is 90%, however, when people are not well rested there chance of passing the exam is only 40%. On any given day 80% of people are well-rested. What is the chance that a person passing

the test is well rested?

- A. $\frac{4}{10}$
- B. $\frac{8}{10}$
- C. $\frac{9}{10}$
- D. $\frac{10}{11}$
- E. Don't know.

Let R be rested and P be passing. Then the answer is and so C is correct.

$$P(R|P) = \frac{P(P|R)P(R)}{P(P|\bar{R})P(\bar{R}) + P(P|R)P(R)} \quad (1)$$

$$= \frac{0.9 \times 0.8}{0.4 \times 0.2 + 0.9 \times 0.8} \quad (2)$$

$$= \frac{0.9}{0.1 + 0.9} = \frac{9}{1 + 9} = \frac{9}{10} \quad (3)$$

Mid-term quiz 8

When carrying out a principal component analysis of a dataset with four attributes we obtain the following singular values $\sigma_1 = 4$, $\sigma_2 = 2$, $\sigma_3 = 1$, and $\sigma_4 = 0$.

Which one of the following statements is wrong?

- A. The first principal component accounts for more than 60% of the variation in the data.
- B. The third principal component accounts for less than 5% of the variation in the data.
- C. The second principal component accounts for more than 20% of the variation in the data.
- D. The data can be perfectly represented in a three dimensional sub-space.
- E. Don't know.

The variance explained of a given coordinate is $\frac{\sigma_i^2}{\sum_{i=1}^4 \sigma_i^2}$. Therefore, the variance explained by the

second coordinate is $\frac{4}{21} < \frac{1}{5}$ and so C is the right answer.

Mid-term quiz 9

Consider the following sequence of numbers

$$x = [0 \ 1 \ 1 \ 1 \ 2 \ 3 \ 4 \ 4 \ 5 \ 14].$$

What is the sum of the mean, the median and the mode of these numbers, i.e. what is the value: $y =$

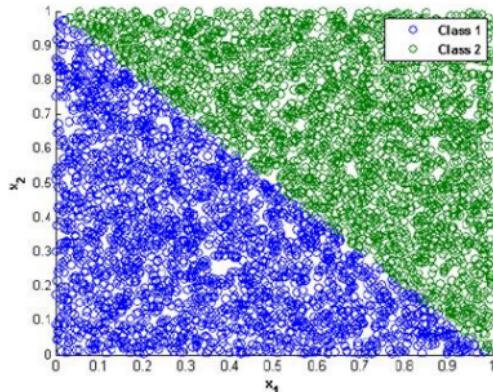
$\text{mean}(x) + \text{median}(x) + \text{mode}(x)$?

- A. $y = 1$
- B. $y = 6$
- C. $y = 7$
- D. $y = 11$
- E. Don't know.

The mode is 1 (most common number). The median is 2.5 (since the list is ordered and contains an even number of elements it is the average of 2 and 3)

and the mean is sum divided by 10 or 3.5. Therefore, the answer is 7.

Mid-term quiz 10



Consider the classification problem given in the figure below where x_1 and x_2 are used as features for a logistic regression classifier and a decision tree. The considered logistic regression models all include the constant term w_0 . Which one of the following

statements is **wrong**?

- A. The two classes can be perfectly separated by a logistic regression model using x_1 and x_2 as features.
- B. A decision tree with less than five nodes, all of the usual axis-aligned form $x_1 > a$ or $x_2 > b$ for different values of a, b , can perfectly separate the classes using only x_1 and x_2 as features.
- C. A logistic regression model can perfectly separate the two classes using only the feature z given by $z = x_1 + x_2$.
- D. In logistic regression the probability that each observation belong to the two classes can be derived from the logistic function.
- E. Don't know.

B: To see why *B* is wrong, note the decision boundary will of such a tree will consist of rectangles with axis-oriented sides. The other options are easily

seen to be correct and for *C*, note that the boundary shown in the plot corresponds to $z > 1$.

02450: Introduction to Machine Learning and Data Mining

Performance evaluation, Bayes, and Naive Bayes

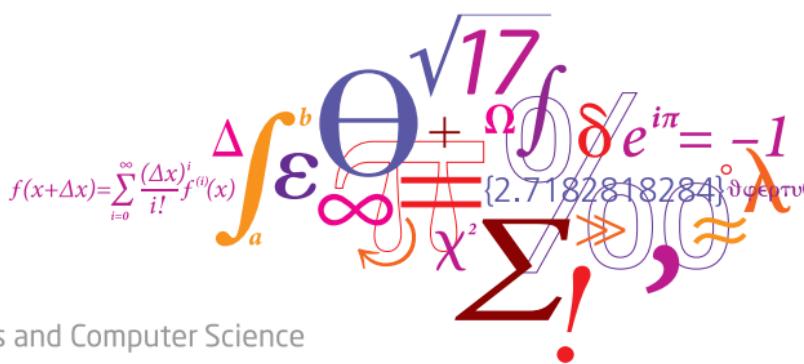
Bjørn Sand Jensen

DTU Compute, Technical University of Denmark (DTU)



DTU Compute

Department of Applied Mathematics and Computer Science

$$f(x+\Delta x) = \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} f^{(i)}(x)$$


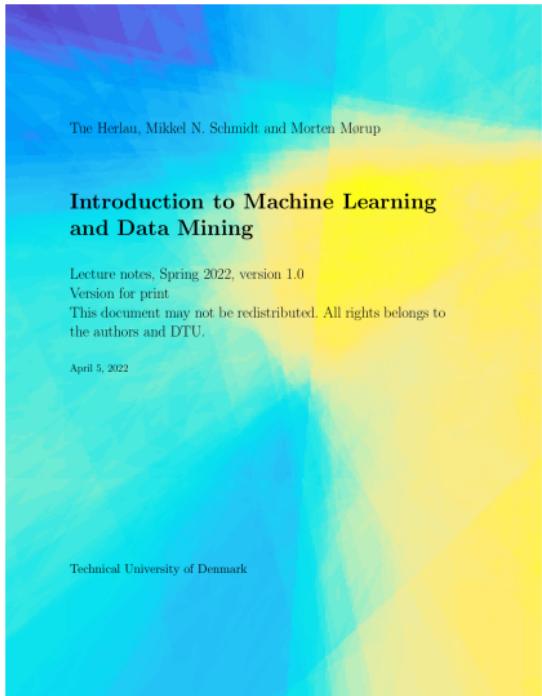
Today

Feedback Groups of the day:

Benedicte Lumby Jessen, Kristine Toft Johansen, Louise Kirstine Jørgensen, Frida Marie Jørgensen, Asbjørn Michael Jørgensen, Ditte Juhler-Nøtstrup, Victor Skaarup Justesen, Yunseo Ka, Dimitri Kaliada, Xuyang Kang, Yusuf Kara, Alland Karim, Malik Schepelern Karrebæk, Hubert Karwowski, Elaheh Kazempour, Yasmine Kennou Filali, Ibrahim Halil Kerpic, Grace Kerr, Dmitri Khlebutin, Philip Kierkegaard, Jegors Kirilovskis, Marios Klavdianos, Matej Kloucek, Dana Theresa Kobinger, William Kock -Andersen, Rasmus Kongsted, Alona Konstantinova, Amalie Martine Jensø Koustrup, Jakob Olund Kristensen, Simon Fogh Kristiansen, Mathias Kronborg, Malia Laurentse Rønne Kuhn, András Kurucsai, Henry Kwan, Raúl Antonio Labarthe Saric, Selma Bundgaard Langvik, Diego Larraguibel Ipinza, Peter Vestereng Larsen, Viktor Dyrby Larsen, Asger Bjørn Larsen, Celina Laungaard, Ying Qi, Tiffanie Leong, Mikkel Christian Rask Levin, Dynel Lewis

Reading material:

Chapter 11, Chapter 13



Lecture Schedule

1 Introduction

30 January: C1

Data: Feature extraction, and visualization

2 Data, feature extraction and PCA

6 February: C2, C3

3 Measures of Similarity, summary statistics and probabilities

13 February: C4, C5

4 Probability densities and data visualization

20 February: C6, C7

Supervised learning: Classification and regression

5 Decision trees and linear regression

27 February: C8, C9 (Project 1 due 29 February at 17:00)

6 Overfitting, cross-validation and Nearest Neighbor

5 March: C10, C12

7 Performance evaluation, Bayes, and Naive Bayes

12 March: C11, C13

Online 24/7 help: Discussion Forum/Piazza
Streaming: Zoom (see link on DTU Learn)
Recordings: <https://panopto.dtu.dk/>
Online exercises: MS Teams

8 Artificial Neural Networks and Bias/Variance

19 March: C14, C15

9 AUC and ensemble methods

2 April: C16, C17

Unsupervised learning: Clustering and density estimation

10 K-means and hierarchical clustering

9 April: C18 (Project 2 due 11 April at 17:00)

11 Mixture models and density estimation

16 April: C19, C20

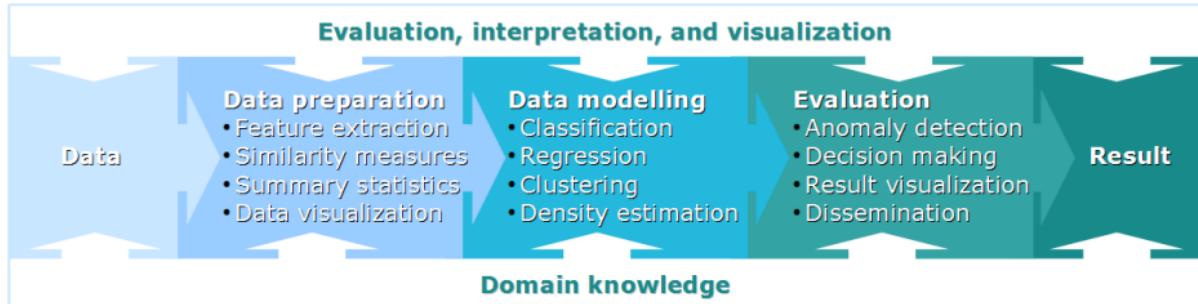
12 Association mining

23 April: C21

Recap

13 Recap and discussion of the exam

30 April: C1-C21



Learning Objectives

- Understand the two different evaluation setups
- Apply appropriate statistical tests to evaluate and compare models
- Account for the assumptions made in Naïve Bayes
- Apply Bayes Theorem to obtain the class posterior likelihood

Statistical testing

- A social media company wish to know if a new ad-placement method increases the click-through rate
- How many customers are likely click adds next month?
- How well can a neural network model learn to distinguish between diseased/non-diseased X-rays?
- Should I recommend my neural network model over a competing method?

All involve induction beyond the dataset

Statistical testing

Tests **can** provide:

- An objective way to choose between methods
- A quantification of model performance which takes uncertainty into account

Tests **do not** provide

- Certain conclusions (Model A is better than B)
- A black-box recipe

Use statistical tests to aid your interpretation of your results not as an argument in itself

Outline

- What is our overall **objective**? What conclusions do we want?
- What tools do we have available?
- What specific test should I use? (classification, regression, etc.)

The objective and evaluation criteria

- Models are compared based on how well they **generalize to future data**
- Suppose we have data $\mathcal{D} = (\mathbf{X}, \mathbf{y})$ and two models $\mathcal{M}_A, \mathcal{M}_B$
- Training on \mathcal{D} , we obtain prediction rules

$$f_{\mathcal{D},A} : \mathbf{x} \rightarrow y, \quad \text{and} \quad f_{\mathcal{D},B} : \mathbf{x} \rightarrow y.$$

- Compared by the **difference in generalization error**:

$$z_{\mathcal{D}} = E_{\mathcal{D},A}^{\text{gen}} - E_{\mathcal{D},B}^{\text{gen}}$$

$$E_{\mathcal{D},A}^{\text{gen}} = \int p(\mathbf{x}, y) L(f_{\mathcal{D},A}(\mathbf{x}), y) d\mathbf{x} dy, \quad E_{\mathcal{D},B}^{\text{gen}} = \int p(\mathbf{x}, y) L(f_{\mathcal{D},B}(\mathbf{x}), y) d\mathbf{x} dy.$$

- If $z_{\mathcal{D}} < 0$, it means **that \mathcal{M}_A is better than \mathcal{M}_Bwhen trained on \mathcal{D}**

Setup I Statistical tests of performance considering the **specific** training set \mathcal{D}

An alternative objective

$$z_{\mathcal{D}} = E_{\mathcal{D},A}^{\text{gen}} - E_{\mathcal{D},B}^{\text{gen}}$$

- If you prove $z_{\mathcal{D}} < 0$, you can't know if this is true for an independent dataset \mathcal{D}'
- Therefore, the conclusion is not independently reproducible
- To overcome this, test if \mathcal{M}_A is better than \mathcal{M}_B when averaging over dataset

$$z = \mathbb{E}_{\mathcal{D}}[z_{\mathcal{D}}] < 0$$

$$E^{\text{gen}} = \int \left[\int L(f_{\mathcal{D}}(\mathbf{x}), y) p(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \right] p(\mathcal{D}) d\mathcal{D}$$

- If $z < 0$, it means \mathcal{M}_A is better than \mathcal{M}_B ... using a typical training set

Setup II Statistical tests of performance considering a dataset of size N

Choices, choices

Setup I Statistical tests of performance considering the **specific** training set \mathcal{D} ?

Setup II *Statistical tests of performance considering a dataset of size N*

Which to choose fundamentally depends on what you want to conclude

- Setup II is a more general (impressive) conclusion
- Setup II is probably what we want in science
- Setup II requires (a lot of) cross-validation
- If you have a single train/test split, use setup I

We will consider **setup I** here

Statistical tasks and tools



Let z be a quantity of interest (for instance $z = E_{\mathcal{A}}^{\text{gen}} - E_{\mathcal{B}}^{\text{gen}}$)

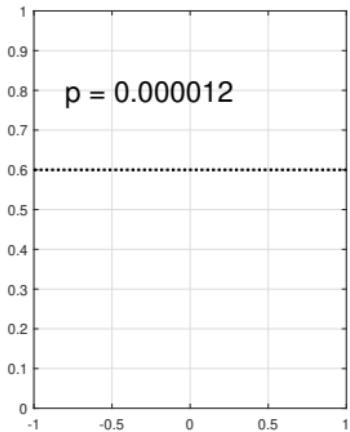
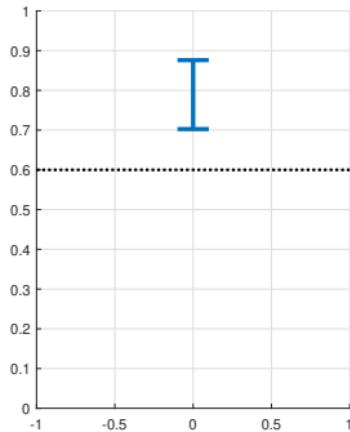
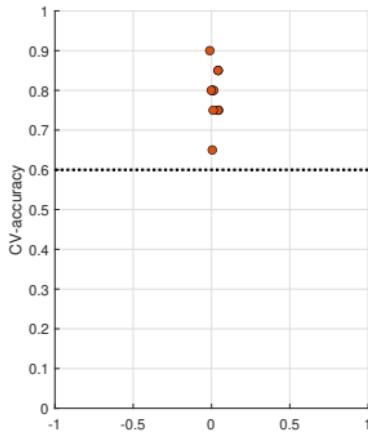
Hypothesis testing Determine **whether there is an effect** by choosing between $H_0 : z = 0$ vs. $H_1 : z \neq 0$

Estimation Determine a likely value $z \approx \hat{z}$ and **an interval** $[z_L, z_U]$ that likely contains z

- Evidence against H_0 is measured by a **p -value** (low p is evidence for an effect $z \neq 0$)
- Estimation of $[z_L, z_U]$ done using an **α -confidence interval** (lower α means a more conservative, wider, interval)

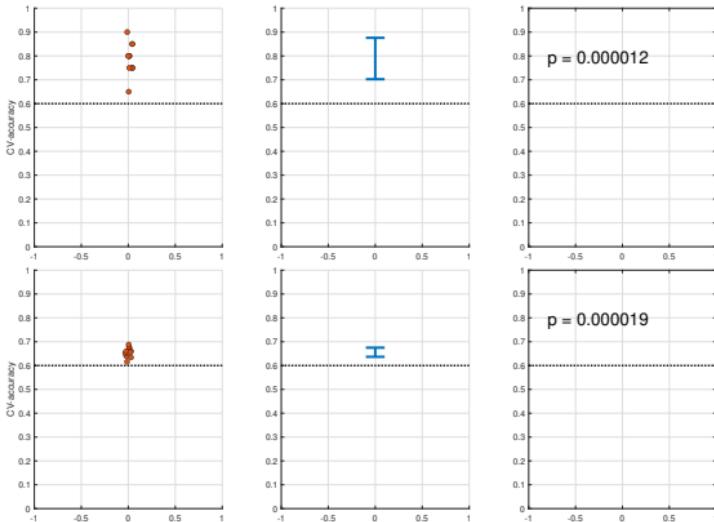
Choosing the right tool

- Consider binary classification using $N = 200$ samples
- We estimate test error using $K = 10$ -fold CV (10 test-error estimates)
- Question: Is accuracy $E_A^{\text{gen}} \approx \frac{1}{K} \sum_{k=1}^K E_i^{\text{test}}$ greater than baseline θ_0 ?
- (Baseline classify everything as maximum class, accuracy 60%)



Which tool to use

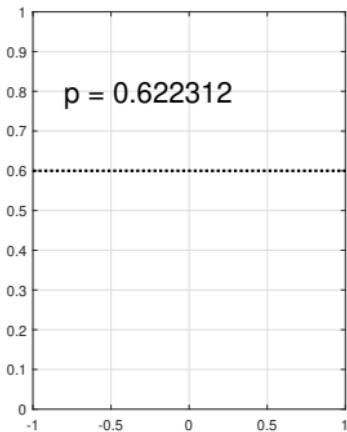
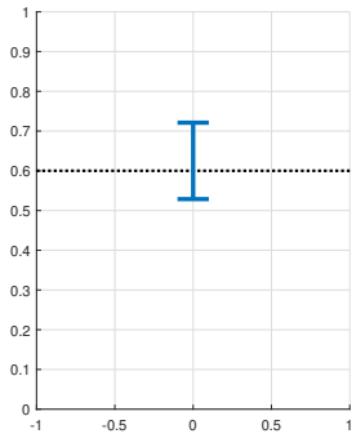
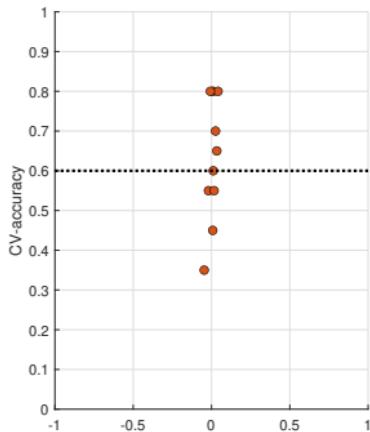
- Top: $N = 200$ sample example
- Bottom: Harder problem using $N = 2000$ samples



- p -value primarily measure of sample size (not **effect size!**)
- Which do **you** think are more informative?

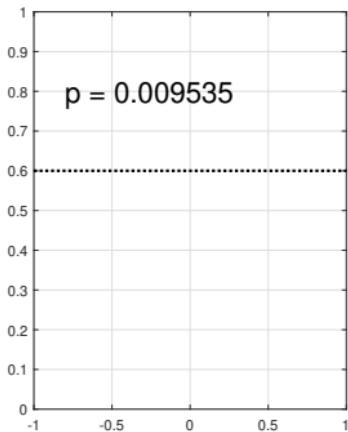
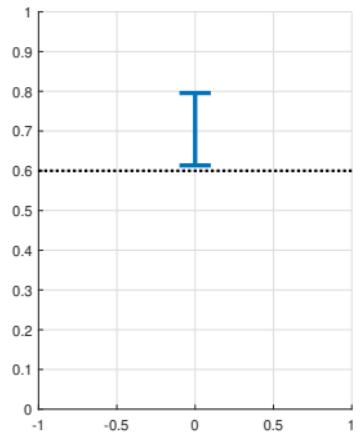
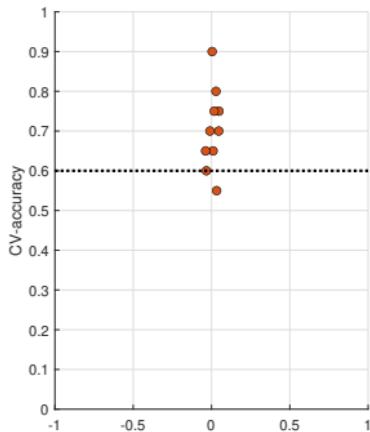
Variability

- New problem using $N = 200$ samples. Is there an effect?

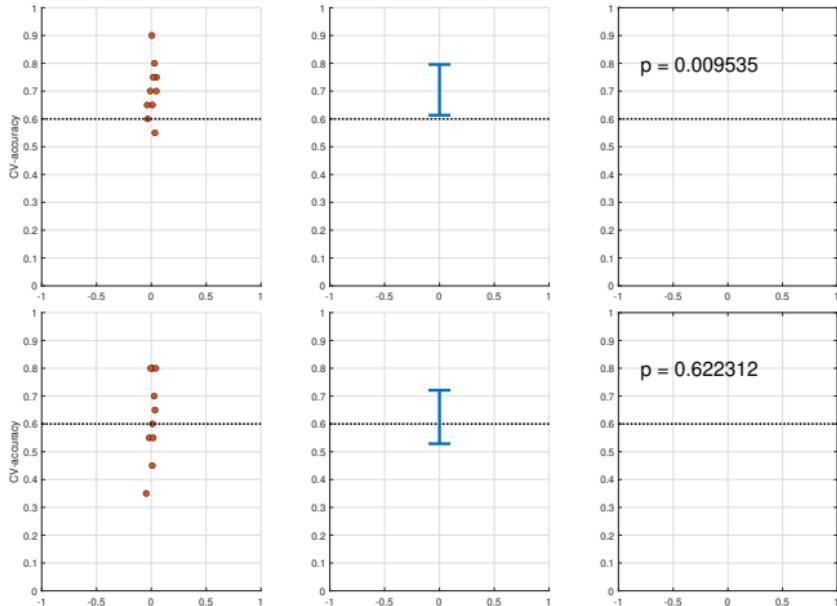


Variability

- Another problem using $N = 200$ samples. Is there an effect?



The nasty bit



- Only difference is random variability in dataset
- Low p -value does **not necessarily** mean reproducible
 - Training many models will lead to false positives
 - Statistics **will not** fix unclear results; probably just lead to false positives

Connecting objective to numbers

- We want to draw conclusions about the difference in performance:

$$z_{\mathcal{D}} = E_{\mathcal{D},A}^{\text{gen}} - E_{\mathcal{D},B}^{\text{gen}}$$

$$E_{\mathcal{D},A}^{\text{gen}} = \int p(\mathbf{x}, y) L(f_{\mathcal{D},A}(\mathbf{x}), y) d\mathbf{x} dy, \quad E_{\mathcal{D},B}^{\text{gen}} = \int p(\mathbf{x}, y) L(f_{\mathcal{D},B}(\mathbf{x}), y) d\mathbf{x} dy.$$

- This can be estimated as

$$\hat{z}_{\mathcal{D}} = \frac{1}{N^{\text{test}}} \sum_{i=1}^{N^{\text{test}}} [L(f_{\mathcal{D},A}(\mathbf{x}_i), y_i) - L(f_{\mathcal{D},B}(\mathbf{x}_i), y_i)]$$

$$= \frac{1}{N^{\text{test}}} \sum_{i=1}^{N^{\text{test}}} z_i, \quad \text{where: } z_i = L(f_{\mathcal{D},A}(\mathbf{x}_i), y_i) - L(f_{\mathcal{D},B}(\mathbf{x}_i), y_i).$$

Abstracting to a statistical question

Consider data as the n numbers

$$D = (z_1, \dots, z_n). \quad (1)$$

General form of the problem: Draw conclusions about

$$\theta = E_{A,D}^{\text{gen}} - E_{B,D}^{\text{gen}}$$

Based on the estimate:

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n z_i. \quad (2)$$

Statistical tools: Parameter

- Assume z_i is a realization of a random variable Z_i

- It has density

$$p(Z_i = z_i | \theta) = p_\theta(z_i)$$

- Density of all dataset

$$p_\theta(D) = \prod_{i=1}^n p_\theta(z_i). \tag{3}$$

- Returning to our goals:

- **estimating plausible ranges of θ**
 - **hypothesis testing such as whether θ takes a particular value**

- Let's look at the statistical tools to accomplish this

Statistical tools: Statistic and estimator

Statistic A statistic is a function of the data D and will be denoted t .

For instance, the mean and variance are both statistics:

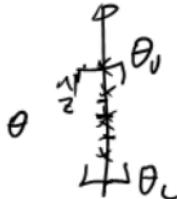
$$t_0(D) = \frac{1}{n} \sum_{i=1}^n Z_i, \text{ or } t_1(D) = \frac{1}{n} \sum_{i=1}^n (Z_i - t_0(D))^2.$$

Estimator An estimator is a statistic t of D such that $t(D)$ is close to θ .

In the examples we will consider the mean

$$t_0(D) = \frac{1}{n} \sum_{i=1}^n Z_i$$

Statistical tools: Confidence interval

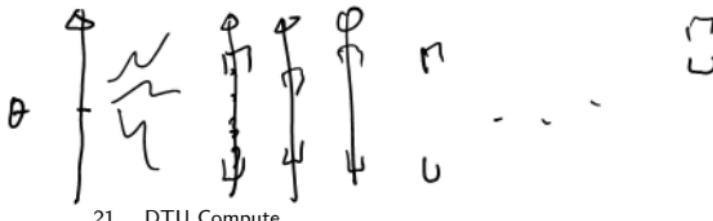


- A **confidence interval** (CI) is an interval $[\theta_L, \theta_U]$ which likely contains θ
- The CI is a function of the data D . θ_L and θ_U are two statistics and for a concrete dataset the interval is computed to be

$$[\theta_L(D), \theta_U(D)]. \quad (4)$$

- With probability $1 - \alpha$, the true value θ should fall within the confidence interval $[\theta_L(D), \theta_U(D)]$ as we randomize over different datasets (gen. by θ)

$$P_\theta(\theta \in [\theta_L, \theta_U]) = 1 - \alpha. \quad (5)$$



Statistical tools: Null hypothesis testing and p -value

- Determining whether a **null hypothesis** H_0 about the parameters is true or false

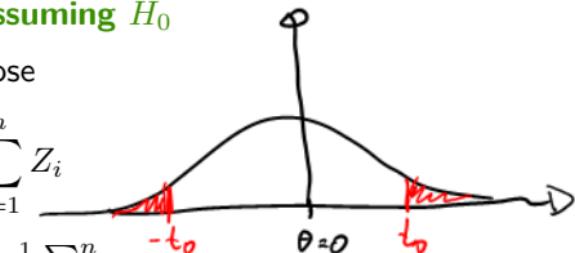
$$H_0 : \theta = 0 \quad \text{vs.} \quad H_1 : \theta \neq 0$$

- Intuitively, if H_0 is true, the data should behave in a certain way

- We test if the data is implausible assuming H_0

- Specifically, let t be a statistic, for our purpose

$$t(D) = \frac{1}{n} \sum_{i=1}^n Z_i$$



On our dataset it has a particular value $t_0 = \frac{1}{n} \sum_{i=1}^n z_i$

- We can compute the density $t(D)$ takes a particular value given H_0 is true:

$$p(t(D) = t | H_0) = p_{\theta=\theta_0}(t(D) = t)$$

- p -value is the chance $t(D)$ is at least as extreme as what we actually observed:

$$p\text{-value} : \quad p = P(t(D) > |t_0| | H_0) = P_{\theta=\theta_0}(t(D) \geq |t_0|). \quad (6)$$

Setup I: Fixed training set

Suppose we carry out cross-validation to obtain:

$$(\mathcal{D}_1^{\text{train}}, \mathcal{D}_1^{\text{test}}), \dots, (\mathcal{D}_K^{\text{train}}, \mathcal{D}_K^{\text{test}}). \quad (7)$$

We collect these into (paired) vectors of predictions and true values:

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_K \end{bmatrix}, \quad \mathbf{y}^{\text{test}} = \begin{bmatrix} y_1^{\text{test}} \\ y_2^{\text{test}} \\ \vdots \\ y_K^{\text{test}} \end{bmatrix}. \quad (8)$$

Evaluation of a single classifier

- Define:

$$c_i = \begin{cases} 1 & \text{if } \hat{y}_i = y_i \\ 0 & \text{if otherwise.} \end{cases}$$

- Number of accurate guesses:

$$m = \sum_{i=1}^n c_i.$$

- Let the chance the classifier is correct be θ . Then, from [Lecture 4](#), we know

$$p(\theta|m, n) = \text{Beta}(\theta|a, b), \quad a = m + \frac{1}{2}, \text{ and } b = n - m + \frac{1}{2}. \quad (9)$$

Evaluating a single classifier (Jeffreys interval)

- If m is the number of accurate guesses, then

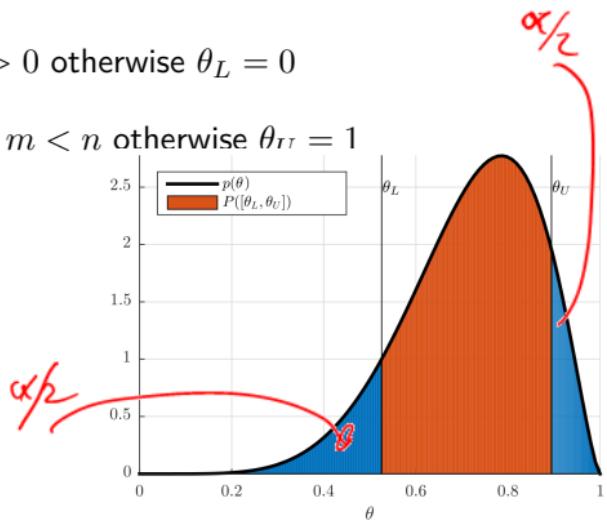
$$p(\theta|m, n) = \text{Beta}(\theta|a, b), \quad a = m + \frac{1}{2}, \text{ and } b = n - m + \frac{1}{2}.$$

- The $1 - \alpha$ confidence interval is given as $[\theta_L, \theta_U]$:

$$\theta_L = \text{cdf}_B^{-1} \left(\frac{\alpha}{2} | a, b \right) \text{ if } m > 0 \text{ otherwise } \theta_L = 0$$

$$\theta_U = \text{cdf}_B^{-1} \left(1 - \frac{\alpha}{2} | a, b \right) \text{ if } m < n \text{ otherwise } \theta_U = 1$$

$$\hat{\theta} = \mathbb{E}[\theta] = \frac{a}{a+b}$$



Comparing two classifiers

- Assume we have predictions from both classifiers:

$$\hat{\mathbf{y}}^A = \hat{y}_1^A, \dots, \hat{y}_n^A, \quad \hat{\mathbf{y}}^B = \hat{y}_1^B, \dots, \hat{y}_n^B.$$

- As before, we want to know if the classifiers are correct or not:

$$c_i^A = \begin{cases} 1 & \text{if } \hat{y}_i^A = y_i \\ 0 & \text{if otherwise.} \end{cases} \quad \text{and} \quad c_i^B = \begin{cases} 1 & \text{if } \hat{y}_i^B = y_i \\ 0 & \text{if otherwise.} \end{cases}$$

- The relevant information is the contingency table:

$$n_{11} = \sum_{i=1}^n c_i^A c_i^B = \{\text{Both classifiers are correct}\}$$

$$n_{12} = \sum_{k=1}^n c_i^A (1 - c_k^B) = \{A \text{ is correct, } B \text{ is wrong}\}$$

$$n_{21} = \sum_{k=1}^n (1 - c_k^A) c_i^B = \{A \text{ is wrong, } B \text{ is correct}\}$$

$$n_{22} = \sum_{k=1}^n (1 - c_k^A)(1 - c_k^B) = \{\text{Both classifiers are wrong}\}$$

Comparing two classifiers: McNemar's test

- We want to compare the accuracy difference: $\theta = \theta_A - \theta_B \in [-1, +1]$
- It is possible to show (approximately)

$$p(\theta | \mathbf{n}) = \frac{1}{2} \text{Beta} \left(\frac{\theta + 1}{2} \mid a = f, b = g \right),$$

$$f = \frac{E_\theta + 1}{2} (Q - 1) \quad g = \frac{1 - E_\theta}{2} (Q - 1)$$

$$E_\theta = \frac{n_{12} - n_{21}}{n}, \quad Q = \frac{n^2(n+1)(E_\theta + 1)(1 - E_\theta)}{n(n_{12} + n_{21}) - (n_{12} - n_{21})^2}.$$

$$\underline{\theta_L = 2\text{cdf}_B^{-1} \left(\frac{\alpha}{2} \mid a = f, b = g \right) - 1, \quad \theta_U = 2\text{cdf}_B^{-1} \left(1 - \frac{\alpha}{2} \mid a = f, b = g \right) - 1}$$
(10)

- For a p -value, note that A is better than B if $n_{12} > n_{21}$
- A p -value can be obtained as:

$$p = 2\text{cdf}_{\text{binom}} \left(m = \min\{n_{12}, n_{21}\} \mid \theta = \frac{1}{2}, N = n_{12} + n_{21} \right)$$

Confidence interval for a regression model

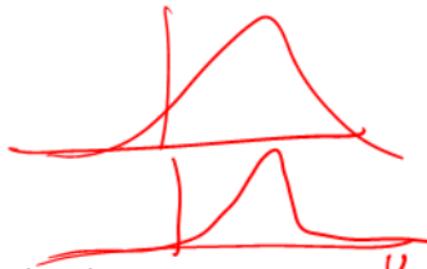
- Use cross-validation to obtain predictions \hat{y}_i and true values y_i . Select loss

$$z_i = |\hat{y}_i - y_i| \quad \text{or} \quad z_i = (\hat{y}_i - y_i)^2 \quad (11)$$

- Estimated error is: $\hat{z} = \frac{1}{n} \sum_{i=1}^n z_i$.
- Assume each error is normally distributed (**warning!**)

$p(m|\mathcal{D})$

$$p(D|u, \sigma^2) = \prod_{i=1}^n \mathcal{N}(z_i|u, \sigma^2)$$



- It is possible to show u follows a generalized Student's t -distribution:

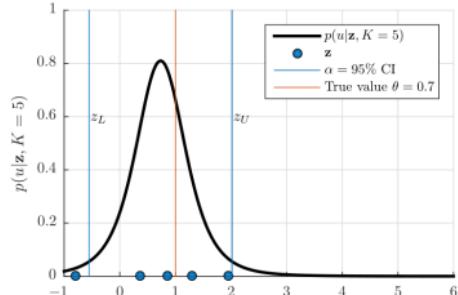
$$\underline{p(u|D)} = p_T(u|\nu = n-1, \mu = \hat{z}, \sigma = \tilde{\sigma})$$

with parameters $\hat{z} = \frac{1}{n} \sum_{i=1}^n z_i$ and $\tilde{\sigma} = \sqrt{\sum_{i=1}^n \frac{(z_i - \hat{z})^2}{n(n-1)}}$.

- The Student's t -distribution has density

$$\text{Student } t\text{-distribution} \quad p_T(x|\nu, \mu, \sigma) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2}) \sqrt{\pi\nu\sigma^2}} \left(1 + \frac{1}{\nu} \left[\frac{x-\mu}{\sigma}\right]^2\right)^{-\frac{\nu+1}{2}}.$$

Confidence interval for a regression model



- Step back: Assuming $z_i = L(y_i, \hat{y}_i)$ and

$$z_i \sim \mathcal{N}(z_i | \mu = u, \sigma^2)$$

- In this case u is the average error rate. Since we have shown:

$$p(u|D) = p_{\mathcal{T}}(u|\nu = n-1, \mu = \hat{z}, \sigma = \tilde{\sigma})$$

- An approximate $1 - \alpha$ confidence interval is:

$$z_L = \text{cdf}_{\mathcal{T}}^{-1} \left(\frac{\alpha}{2} \mid \nu, \hat{z}, \tilde{\sigma} \right), \quad z_U = \text{cdf}_{\mathcal{T}}^{-1} \left(1 - \frac{\alpha}{2} \mid \nu, \hat{z}, \tilde{\sigma} \right). \quad (12)$$

Comparing two regression models

- Use cross-validation to obtain (paired) predictions along with true values y_i

$$\hat{y}_1^A, \dots, \hat{y}_n^A, \quad \text{and} \quad \hat{y}_1^B, \dots, \hat{y}_n^B. \quad (13)$$

- Select a loss-function to compute the per-observation losses as in

$$z_1^A, \dots, z_n^A, \quad \text{and} \quad z_1^B, \dots, z_n^B.$$

- Note that

$$\begin{aligned} z &= E_{A,\mathcal{D}}^{\text{gen}} - E_{B,\mathcal{D}}^{\text{gen}} \approx \hat{z} = \left(\frac{1}{n} \sum_{i=1}^n z_i^A \right) - \left(\frac{1}{n} \sum_{i=1}^n z_i^B \right) \\ &= \frac{1}{n} \sum_{i=1}^n z_i, \quad \text{where } z_i = z_i^A - z_i^B \end{aligned}$$

- Assume $z_i \sim \mathcal{N}(z_i | \mu = u, \sigma^2)$
- Compute a $1 - \alpha$ CI using methods on previous slide

Comparing two regression models: p -values

$$z = E_A^{\text{gen}} - E_B^{\text{gen}} \approx \hat{z} = \frac{1}{n} \sum_{i=1}^n z_i, \quad \text{where } z_i = z_i^A - z_i^B$$

- Assuming

$$z_i \sim \mathcal{N}(z_i | \mu = u, \sigma^2)$$

where u is the true difference in error function we have shown:

$$p(u|D) = p_{\mathcal{T}}(u|\nu = n - 1, \mu = \hat{z}, \sigma = \tilde{\sigma})$$

- Therefore, we can test the hypothesis

$$H_0 : \text{Model } \mathcal{M}_A \text{ and } \mathcal{M}_B \text{ have the same performance, } u = 0 \quad (14)$$

$$H_1 : \text{Model } \mathcal{M}_A \text{ and } \mathcal{M}_B \text{ have different performance, } u \neq 0. \quad (15)$$

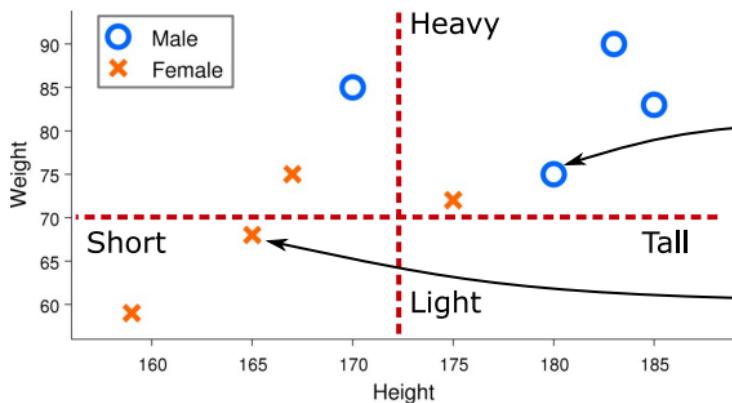
- A p -value can be computed as

$$p = 2\text{cdf}_{\mathcal{T}}(-|\hat{z}| \mid \nu = n - 1, \mu = 0, \sigma = \tilde{\sigma}). \quad (16)$$

Which type of cross-validation?

- When using **setup I** choose K as large as feasible (leave-one-out)
- Hold-out has the benefit the training/test data is fixed
- Results will be significant with enough data
 - Focus on estimated an effect size
 - Multiple-comparison problem
 - **Transparency, availability of datasets/code, breadth of testing, self-criticism** guarantees reprehensibility, not a sophisticated test
- In **setup II**, correlation of training data is taken into account and K -fold is optimal
 - Your **setup I** results do not generalize beyond your training data

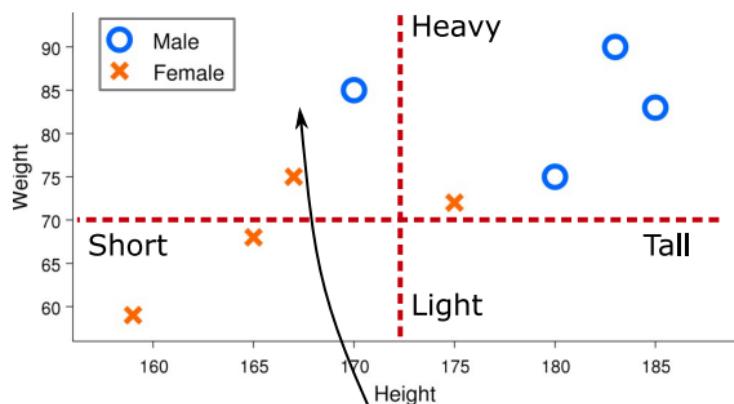
Bayes and Naive-Bayes



| Gender y | Tall x_1 | Heavy x_2 |
|------------|------------|-------------|
| 0 | 1 | 1 |
| 0 | 1 | 1 |
| 0 | 1 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |

$$p(y|x_1, x_2) = \frac{p(x_1, x_2|y)p(y)}{\sum_{k=0}^1 p(x_1, x_2|y=k)p(y=k)}$$

Example 1: Normal Bayes



| Gender y | Tall x_1 | Heavy x_2 |
|------------|------------|-------------|
| 0 | 1 | 1 |
| 0 | 1 | 1 |
| 0 | 1 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |

Probability a short, heavy person is male:

$$P(y = 0|x_1 = 0, x_2 = 1) = \frac{p(x_1 = 0, x_2 = 1|y = 0)p(y = 0)}{\sum_{k=0}^1 p(x_1 = 0, x_2 = 1|y = k)p(y = k)}$$

$$= \frac{\frac{1}{4} \cdot \frac{4}{8}}{\frac{1}{4} \cdot \frac{4}{8} + \frac{1}{4} \cdot \frac{4}{8}} = \frac{1}{2}$$

Example 1: Solution

Probability a short, heavy person is male:

$$\begin{aligned} P(y = 0|x_1 = 0, x_2 = 1) &= \frac{p(x_1 = 0, x_2 = 1|y = 0)p(y = 0)}{\sum_{k=0}^1 p(x_1 = 0, x_2 = 1|y = k)p(y = k)} \\ &= \frac{\frac{1}{4} \cdot \frac{4}{8}}{\frac{1}{4} \cdot \frac{4}{8} + \frac{1}{4} \cdot \frac{4}{8}} = \frac{1}{2} \end{aligned}$$

A practical problem with Bayesian classifier

- In general:

$$p(y|x_1, x_2, \dots, x_M) = \frac{p(x_1, x_2, \dots, x_M|y)p(y)}{\sum_{k=0}^{K-1} p(x_1, x_2, \dots, x_M|y=k)p(y=k)}$$
$$p(x_1, \dots, x_M|y=k) = \frac{\text{Nr. obs where } y=k \text{ and we measure } x_1, \dots, x_M}{\text{Observations where } y=k}$$

A practical problem with Bayesian classifier

- In general:

$$p(y|x_1, x_2, \dots, x_M) = \frac{p(x_1, x_2, \dots, x_M|y)p(y)}{\sum_{k=0}^{K-1} p(x_1, x_2, \dots, x_M|y=k)p(y=k)}$$
$$p(x_1, \dots, x_M|y=k) = \frac{\text{Nr. obs where } y=k \text{ and we measure } x_1, \dots, x_M}{\text{Observations where } y=k}$$

- Naive Bayes assumption

$$p(x_1, x_2, \dots, x_M|y) = p(x_1|y)p(x_2|y) \times \dots \times p(x_M|y)$$

A practical problem with Bayesian classifier

- In general:

$$p(y|x_1, x_2, \dots, x_M) = \frac{p(x_1, x_2, \dots, x_M|y)p(y)}{\sum_{k=0}^{K-1} p(x_1, x_2, \dots, x_M|y=k)p(y=k)}$$
$$p(x_1, \dots, x_M|y=k) = \frac{\text{Nr. obs where } y=k \text{ and we measure } x_1, \dots, x_M}{\text{Observations where } y=k}$$

- Naive Bayes assumption

$$p(x_1, x_2, \dots, x_M|y) = p(x_1|y)p(x_2|y) \times \dots \times p(x_M|y)$$

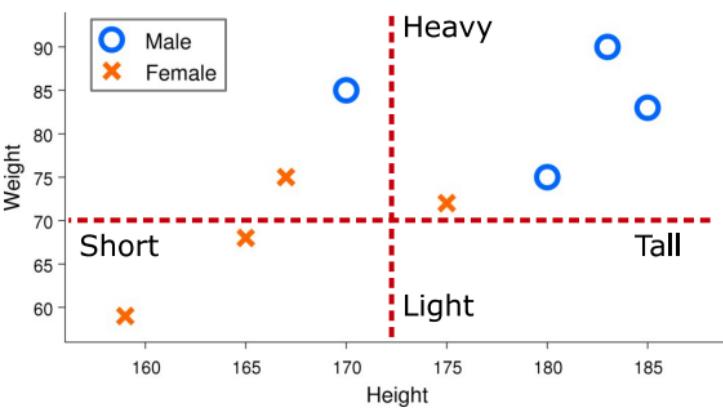
- Naive Bayes classifier

$$p(y|x_1, x_2, \dots, x_M) = \frac{p(x_1, x_2, \dots, x_M|y)p(y)}{\sum_{k=0}^1 p(x_1, x_2, \dots, x_M|y=k)p(y=k)}$$
$$= \frac{p(x_1|y)p(x_2|y) \times \dots \times p(x_M|y)p(y)}{\sum_{k=0}^1 p(x_1|y=k)p(x_2|y=k) \times \dots \times p(x_M|y=k)p(y=k)}$$

Example 2:

- Naive Bayes classifier (Probability someone is a female given they are heavy and tall)

$$p(y = 1|x_1 = 1, x_2 = 1) = \frac{p(x_1|y)p(x_2|y)p(y)}{\sum_{k=0}^1 p(x_1|y=k)p(x_2|y=k)p(y=k)}$$

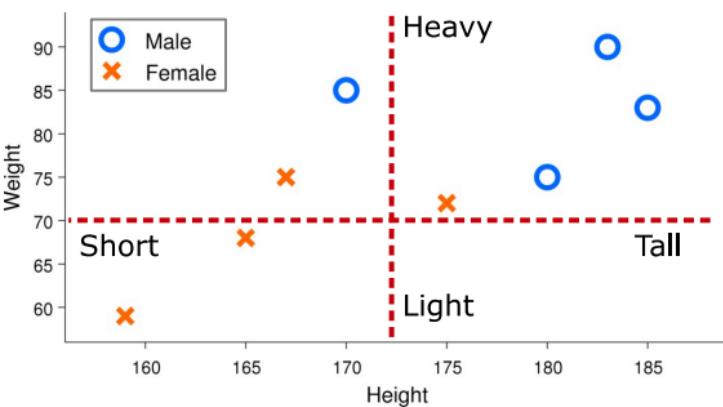


| Gender y | Tall x_1 | Heavy x_2 |
|------------|------------|-------------|
| 0 | 1 | 1 |
| 0 | 1 | 1 |
| 0 | 1 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |

Example 2: Solution

- Naive Bayes classifier (Probability someone is a female given they are heavy and tall)

$$\begin{aligned}
 p(y = 1|x_1 = 1, x_2 = 1) &= \frac{p(x_1|y)p(x_2|y)p(y)}{\sum_{k=0}^1 p(x_1|y=k)p(x_2|y=k)p(y=k)} \\
 &= \frac{\frac{1}{4} \frac{2}{4} \frac{1}{2}}{\frac{1}{4} \frac{2}{4} \frac{1}{2} + \frac{3}{4} \frac{4}{4} \frac{1}{2}} = \frac{2}{2+12} = \frac{1}{7}
 \end{aligned}$$



| Gender y | Tall x_1 | Heavy x_2 |
|------------|------------|-------------|
| 0 | 1 | 1 |
| 0 | 1 | 1 |
| 0 | 1 | 1 |
| 0 | 0 | 1 |
| 1 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |

Quiz 1, Naive-Bayes (Spring 2012)

| | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 |
|----|----|----|----|----|----|----|----|----|----|-----|
| P1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| P2 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 |
| P3 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| P4 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| P5 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |
| P6 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |

Table 1: Table indicating whether 10 songs denoted S1–S10 are downloaded to 6 different phones denoted P1–P6. P1 and P2 given in red are phones that belong to females whereas P3, P4, P5, and P6 given in blue belong to males.

The phones P1 and P2 are owned by females whereas P3, P4, P5 and P6 are owned by males (this is indicated in red and blue respectively in Table 1). We would like to predict whether a phone is owned by a male based on whether or not the songs S1, S2 and S3 have been downloaded. We will therefore classify whether the phone belongs to a male or female considering only the attributes S1, S2 and S3 and the data in Table 1. We will apply a Naïve Bayes classifier that assumes independence between these attributes. Given that a phone has installed songs 1, 2 and 3 (i.e., S1=1, S2=1 and S3=1) What is the probability that the phone is owned by a male according to the Naïve Bayes classifier?

- A. 1/12
- B. 1/6
- C. 2/3
- D. 1
- E. Don't know.

$$p(y|x_1, x_2, \dots, x_M) = \frac{p(x_1|y) \times \dots \times p(x_M|y)p(y)}{\sum_{k=0}^1 p(x_1|y=k) \times \dots \times p(x_M|y=k)p(y=k)}$$

According to the Naïve Bayes classifier we have

$$P(Male|S1 = 1, S2 = 1, S3 = 1) =$$

$$\frac{\begin{pmatrix} P(S1 = 1|Male) \times \\ P(S2 = 1|Male) \times \\ P(S3 = 1|Male) \times \\ P(Male) \end{pmatrix}}{\begin{pmatrix} P(S1 = 1|Female) \times \\ P(S2 = 1|Female) \times \\ P(S3 = 1|Female) \times \\ P(Female) \end{pmatrix} + \begin{pmatrix} P(S1 = 1|Male) \times \\ P(S2 = 1|Male) \times \\ P(S3 = 1|Male) \times \\ P(Male) \end{pmatrix}}$$
$$= \frac{2/4 \cdot 2/4 \cdot 2/4 \cdot 4/6}{2/2 \cdot 0/2 \cdot 1/2 \cdot 2/6 + 2/4 \cdot 2/4 \cdot 2/4 \cdot 4/6} = 1.$$

Robust estimation and non-binary data

Assume

$$p(x_1, \dots, x_M | y) = \prod_{k=1}^M p(x_k | y)$$

Defining $n_{x_j=k|y=c} = \sum_{i=1}^N \delta_{X_{ij}, k} \delta_{y, c}$ we have more generally:

Binary case: $p(x_j = 1 | y = c) = \frac{n_{x_j=1|y=c} + \alpha}{N_c + 2\alpha}.$

Categorical case: $p(x_j = k | y = c) = \frac{n_{x_j=k|y=c} + \alpha}{N_c + K\alpha}.$

Continuous case: $p(x_j = x | y = c) = \mathcal{N}(x | \mu = \mu_{j|c}, \sigma^2 = (\sigma_{j|c} + \alpha)^2)$

$$\mu_{j|c} = \mathbb{E}_{y=c}[x_j] = \frac{1}{N_c} \sum_{i=1}^N \delta_{y_i, c} X_{ij},$$

$$\sigma_{j|c} = \text{std}_{y=c}[x_j] = \sqrt{\frac{1}{N_c - 1} \sum_{i=1}^N \delta_{y_i, c} (X_{ij} - \mu_c)^2}$$

Select these parameters using cross-validation.

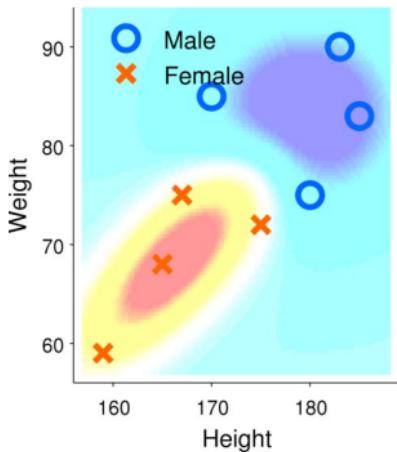
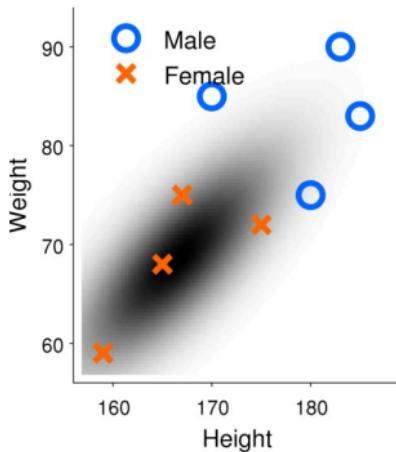
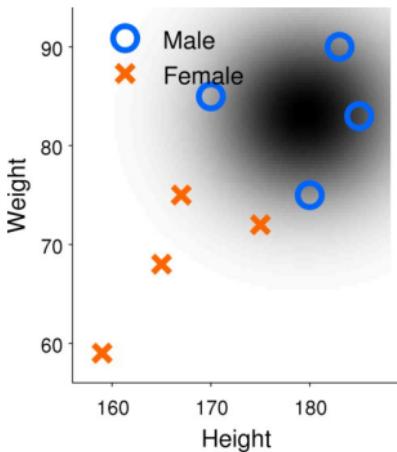
Bayesian classification by the multivariate normal distribution

Continuous density estimation

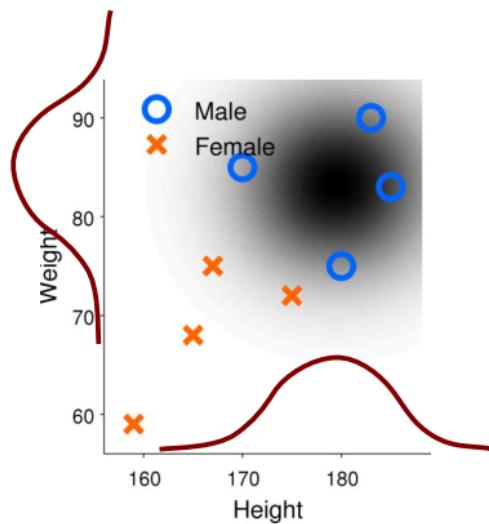
- Fit a Normal distribution to each class
 - Compute class mean and covariance
- Classify using Bayes rule as before

$$P(\mathbf{x}|y = c) = \frac{1}{(2\pi)^{M/2} \det(\boldsymbol{\Sigma}_c)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c^{-1} (\mathbf{x} - \boldsymbol{\mu}_c)\right)$$

$$P(y = c|\mathbf{x}) = \frac{P(\mathbf{x}|y = c)P(y = c)}{\sum_{c'} P(\mathbf{x}|y = c')P(y = c')}$$



- What does the Naive Bayes assumption of independence of the attributes correspond to in terms of the parameters of the multivariate normal distribution?



Midterm practice test

Look at the test on DTU Learn. Note the test is not part of your evaluation.

Midterm question 1

In the analysis of house prices the following attributes were collected for a house: The year the house was built (denoted YEAR), the size of the house given in square meters (denoted SIZE) the county in which the house is located (denoted LOCATION). Which statement about the three attributes is correct?

- A. YEAR is ratio, SIZE is interval and LOCATION is nominal
- B. YEAR is interval, SIZE is ratio and LOCATION is nominal
- C. YEAR is interval, SIZE is ratio and LOCATION is ordinal
- D. YEAR is interval, SIZE is ratio and LOCATION is interval
- E. Don't know.

Midterm question 2

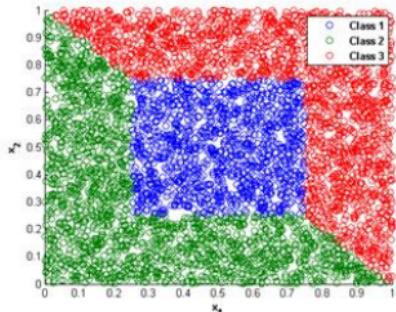
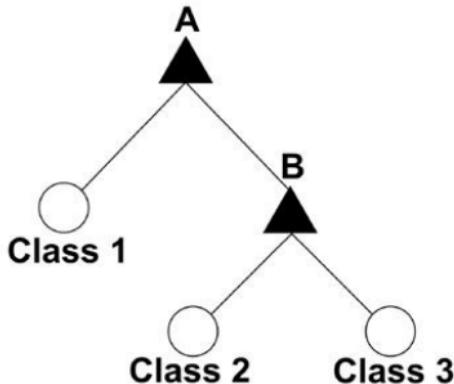


Figure 1

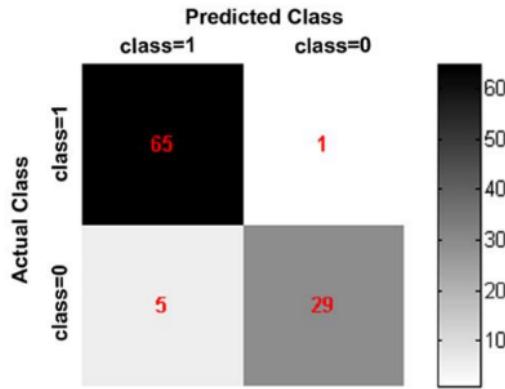


Consider the classification problem given in figure 1 and the Decision Tree shown below it with two decision nodes denoted A and B . We will let $\mathbf{x}_n = (x, y)$ denote a 2-dimensional observation such that $\mathbf{x}_n - 0.5 \cdot \mathbf{1}$ denotes the subtraction of 0.5 from each of the two coordinates of \mathbf{x}_n . Which one of the following classification rules would lead to a correct classification of the data?

- A. A: $\|\mathbf{x}_n - 0.5 \cdot \mathbf{1}\|_1 \leq 0.25$, B : $\|\mathbf{x}_n\|_\infty \leq 1$
- B. A: $\|\mathbf{x}_n\|_1 \leq 1$, B: $\|\mathbf{x}_n - 0.5 \cdot \mathbf{1}\|_2 \leq \infty$
- C. A: $\|\mathbf{x}_n - 0.5 \cdot \mathbf{1}\|_2 \leq 0.25$, B : $\|\mathbf{x}_n\|_\infty \leq 1$
- D. A: $\|\mathbf{x}_n - 0.5 \cdot \mathbf{1}\|_\infty \leq 0.25$, B : $\|\mathbf{x}_n\|_1 \leq 1$
- E. Don't know.

Midterm question 3

A classifier has the confusion matrix given in the figure below. Which statement about the classifier is correct?



- A. The Accuracy is 94% and the Error rate is 6%
- B. The Accuracy is 6% and the Error rate is 94%
- C. The Accuracy is 65% and the Error rate is 35%
- D. There is insufficient information in the confusion matrix to determine the Accuracy and Error rate.
- E. Don't know.

Midterm question 4

Which statement about crossvalidation is wrong?

- A. Cross-validation can be used to estimate the generalization error.
- B. Leave one out cross-validation is more computationally expensive than 10 fold crossvalidation.
- C. Holding out one third of the data for validation is faster but less accurate than performing 10 fold cross-validation.
- D. The same test set can be used for model selection as well as evaluation of the generalization performance of the model.
- E. Don't know.

Midterm question 5

Consider a data set of four features: A , B , C , and D that are applied in a classification algorithm. The table below shows the cross-validated Error rate when using different combinations of the features.

| Feature(s) | Error rate |
|---------------------|------------|
| A | 0.40 |
| B | 0.45 |
| C | 0.33 |
| D | 0.42 |
| A and B | 0.20 |
| A and C | 0.25 |
| A and D | 0.34 |
| B and C | 0.29 |
| B and D | 0.42 |
| C and D | 0.40 |
| A and B and C | 0.13 |
| A and B and D | 0.17 |
| B and C and D | 0.10 |
| A and C and D | 0.15 |
| A and B and C and D | 0.28 |

We will apply a forward feature selection algorithm. Which feature set will the selection algorithm choose?

- A. C
- B. B and C and D
- C. A and B
- D. A and B and C
- E. Don't know.

Midterm question 6

When training a decision tree we will use the classification error as impurity measure $I(t)$ given by $I(t) = 1 - \max_i [p(i|t)]$ where $p(i|t)$ denotes the fraction of data objects belonging to class i at a given node t . We will use Hunt's algorithm to grow the tree and recall that the purity gain is given by:

$$\Delta = I(\text{Parent}) - \sum_{j=1}^k \frac{N(v_j)}{N} I(v_j)$$

where N is the total number of data objects at the parent node, k is the number of child nodes and $N(v_j)$ is the number of data objects associated with the child node, v_j . We will consider classification of Iris flowers into Iris-Setosa, Iris-Virginica and Iris-Versicolor. At a potential split we have:

- Before the split: 5 Iris-Setosa, 10 Iris-Virginica and 10 Iris Versicolor.

After the split

- 0 Iris-Setosa, 8 Iris-Virginica and 2 Iris-Versicolor in the left node.
- 5 Iris-Setosa, 2 Iris-Virginica and 8 Iris-Versicolor in the right node.

Which statement is correct?

- A. The purity gain is $\Delta = \frac{3}{5}$
- B. The purity gain is $\Delta = \frac{3}{15}$
- C. The purity gain is $\Delta = \frac{6}{25}$
- D. The purity gain is $\Delta = \frac{7}{15}$
- E. Don't know.

Midterm question 7

When people are well rested and take an exam their chance of passing the exam is 90%, however, when people are not well rested there chance of passing the exam is only 40%. On any given day 80% of people are well-rested. What is the chance that a person passing

the test is well rested?

- A. $\frac{4}{10}$
- B. $\frac{8}{10}$
- C. $\frac{9}{10}$
- D. $\frac{10}{11}$
- E. Don't know.

Midterm question 8

When carrying out a principal component analysis of a dataset with four attributes we obtain the following singular values $\sigma_1 = 4$, $\sigma_2 = 2$, $\sigma_3 = 1$, and $\sigma_4 = 0$.

Which one of the following statements is wrong?

- A. The first principal component accounts for more than 60% of the variation in the data.
- B. The third principal component accounts for less than 5% of the variation in the data.
- C. The second principal component accounts for more than 20% of the variation in the data.
- D. The data can be perfectly represented in a three dimensional sub-space.
- E. Don't know.

Midterm question 9

Consider the following sequence of numbers

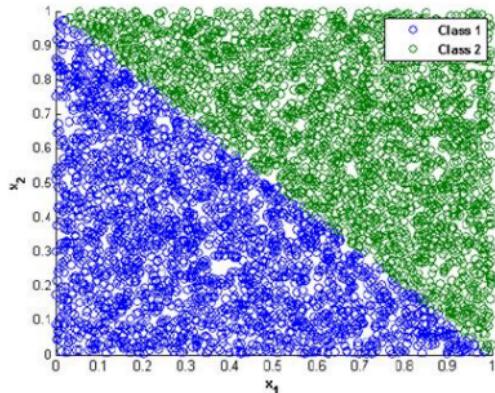
$$x = [0 \ 1 \ 1 \ 1 \ 2 \ 3 \ 4 \ 4 \ 5 \ 14].$$

What is the sum of the mean, the median and the mode of these numbers, i.e. what is the value: $y =$

$\text{mean}(x) + \text{median}(x) + \text{mode}(x)$?

- A. $y = 1$
- B. $y = 6$
- C. $y = 7$
- D. $y = 11$
- E. Don't know.

Midterm question 10



Consider the classification problem given in the figure below where x_1 and x_2 are used as features for a logistic regression classifier and a decision tree. The considered logistic regression models all include the constant term w_0 . Which one of the following

statements is **wrong**?

- A. The two classes can be perfectly separated by a logistic regression model using x_1 and x_2 as features.
- B. A decision tree with less than five nodes, all of the usual axis-aligned form $x_1 > a$ or $x_2 > b$ for different values of a, b , can perfectly separate the classes using only x_1 and x_2 as features.
- C. A logistic regression model can perfectly separate the two classes using only the feature z given by $z = x_1 + x_2$.
- D. In logistic regression the probability that each observation belong to the two classes can be derived from the logistic function.
- E. Don't know.

Resources

<https://www.youtube.com> Video explaining Naive Bayes

(<https://www.youtube.com/watch?v=8yvBqhm92xA>)

<https://machinelearningmastery.com> Statistical comparison of the cross-validation estimate of the generalization error is not a solved problem. This reference provides an overview of various issues and proposed solutions. Note no simple solution exists.

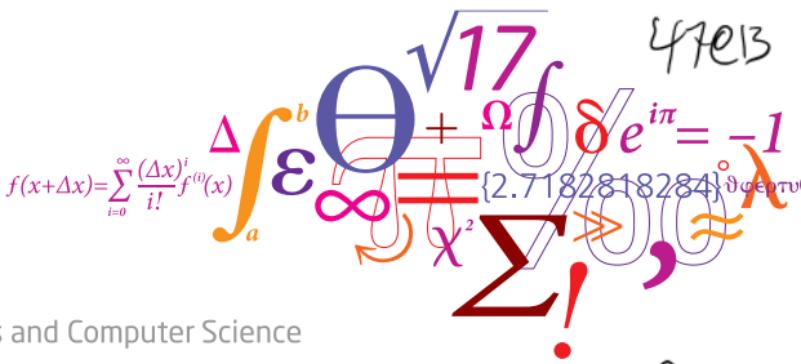
(<https://machinelearningmastery.com/statistical-significance-tests-for-comparing-machine-learning-algorithms/>)

02450: Introduction to Machine Learning and Data Mining

Overfitting, cross-validation and Nearest Neighbor

Bjørn Sand Jensen

DTU Compute, Technical University of Denmark (DTU)



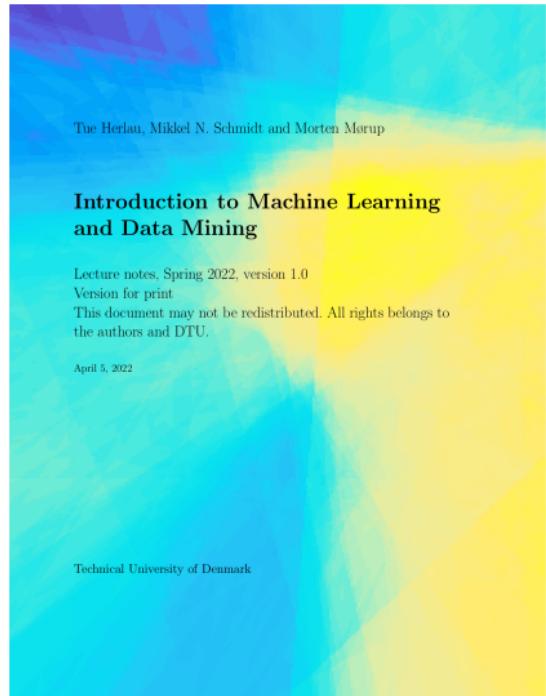
Today

Feedback Groups of the day:

Gustav Broe Hansen, Clara Marie Zacho Hansen,
Monica Diaz Hansen, Martin Moos Hansen, Christian
Alexander Harborg, Rune Daugaard Harlyk, Micki
Kanaiya Harning, Casper Holm Harreby, Julie Alsing
Haugaard, Martin Vennick Haugbølle, Mari Helgesen,
Gerard Hernández Gil, David Herrera Manzanedo,
Hugo André Michel HERRERO, Oleg Bjørn Klimenko
Hertoft, Liam Hinze, Benjamin Aastrup Hoch,
Isabelle Hoch-Nielsen, Mads Helle Højgaard, William
Ludvig Holberg, Mikkel Bjarke Horn, Gabriel Sejr
Hornstrup, Nicolai Hornung, Hróbjartur Höskuldsson,
Mattias Benjamin Houen, Nichlas Høst Husted,
Philip Alexander Mebrahtu Hviid, Csaba-Róbert
Ilyés, Verena Vanessa Irmng-Pedersen, Erkam Isbilir,
Md Jahidul Islam, Taner Tahir Ismet, Christoffer Emil
Iversen, Sverri Jacobsen, Nikolaj Holst Jakobsen,
Hector Helt Jakobsen, Oliver Stenberg Jakobsen, Pál
Jámbor, Saad Rizwan Jaura, Søren Serup Jensen,
Marcus Leander Jensen, Jeppe Elias Ekberg Jensen,
Mia Fredensborg Jensen, Mapendo Jeremia

Reading material:

Chapter 10, Chapter 12



Lecture Schedule

1 Introduction

30 January: C1

Data: Feature extraction, and visualization

2 Data, feature extraction and PCA

6 February: C2, C3

3 Measures of Similarity, summary statistics and probabilities

13 February: C4, C5

4 Probability densities and data visualization

20 February: C6, C7

Supervised learning: Classification and regression

5 Decision trees and linear regression

27 February: C8, C9 (Project 1 due 29 February at 17:00)

6 Overfitting, cross-validation and Nearest Neighbor

5 March: C10, C12

7 Performance evaluation, Bayes, and Naive Bayes

12 March: C11, C13

Online 24/7 help: Discussion Forum/Piazza
Streaming: Zoom (see link on DTU Learn)
Recordings: <https://panopto.dtu.dk/>
Online exercises: MS Teams

8 Artificial Neural Networks and Bias/Variance

19 March: C14, C15

9 AUC and ensemble methods

2 April: C16, C17

Unsupervised learning: Clustering and density estimation

10 K-means and hierarchical clustering

9 April: C18 (Project 2 due 11 April at 17:00)

11 Mixture models and density estimation

16 April: C19, C20

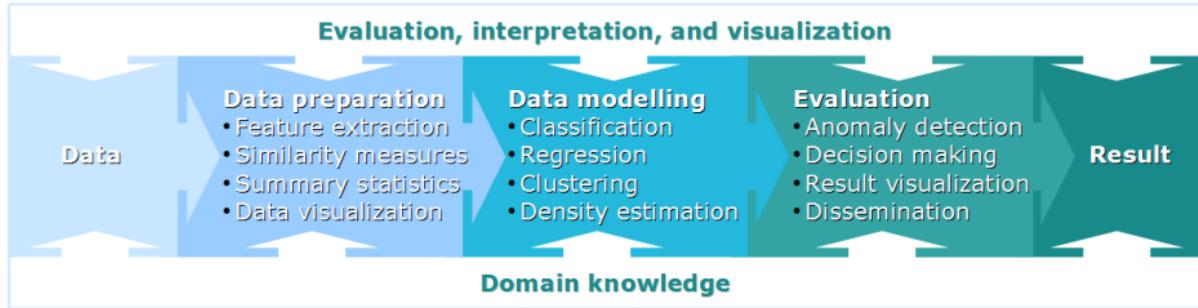
12 Association mining

23 April: C21

Recap

13 Recap and discussion of the exam

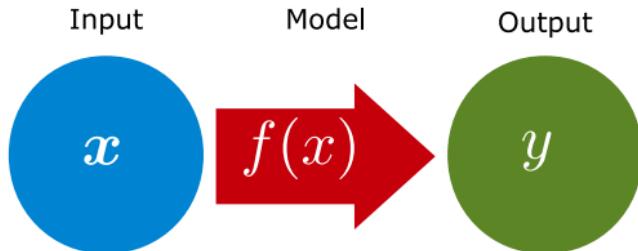
30 April: C1-C21



Learning Objectives

- Explain the difference between training, test and generalization error
- Explain how cross-validation can be used for (i) performance evaluation (ii) model selection
- Apply forward and backward selection
- Explain how K-Nearest Neighbours can be used to classify data

Supervised learning



- **Mapping between domains**
 - Classification: Discrete output
 - Regression: Continuous output

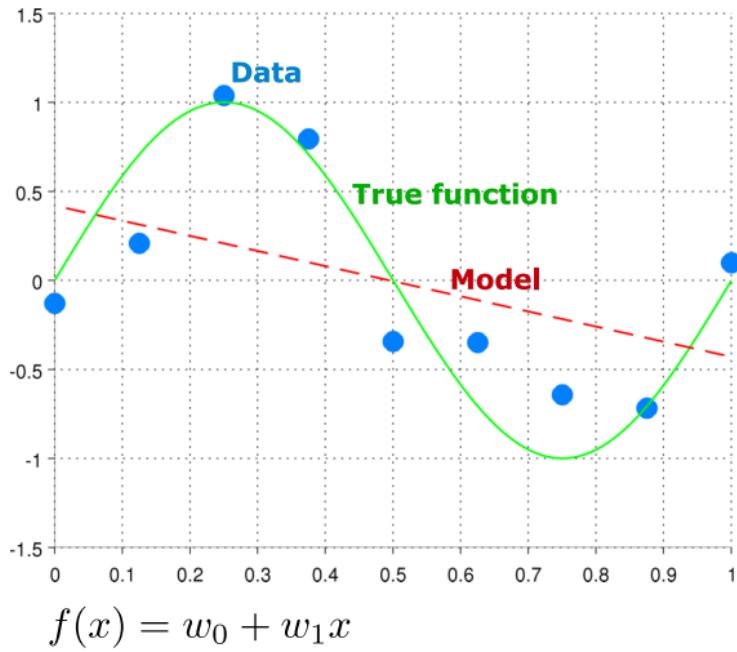
Roadmap

- Introduce errors
 - Training error
 - Test error
 - Generalization error
- Introduce cross-validation
 - **Basic cross validation** for **performance evaluation**
 - **Cross-validation** for **model selection**
 - **Two-level cross-validation** for **model selection and performance evaluation**
- Nearest Neighbor methods

Why are there “multiple models”?

Example: Linear regression

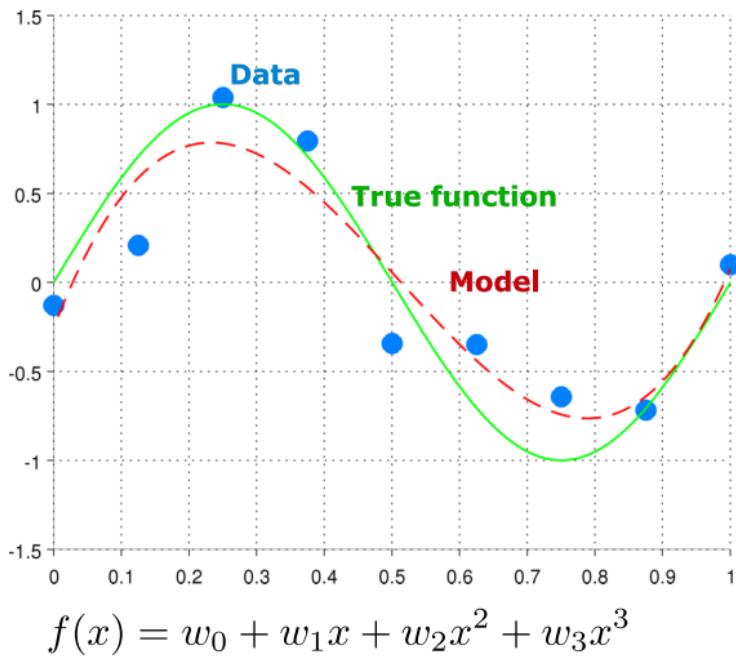
- Bad fit
- Too simple model



Why are there “multiple models”?

Example: Linear regression

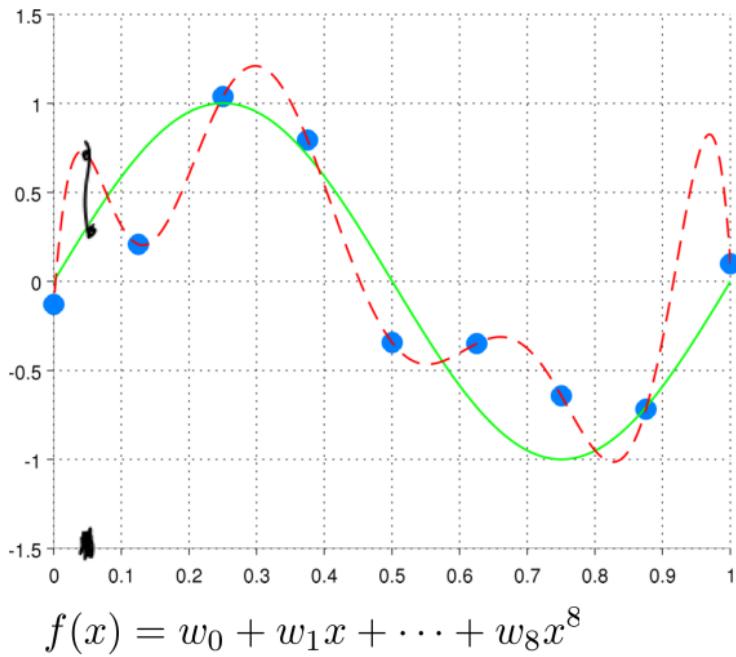
- Reasonable fit
- **Reasonable model**



Why are there “multiple models”?

Example: Linear regression

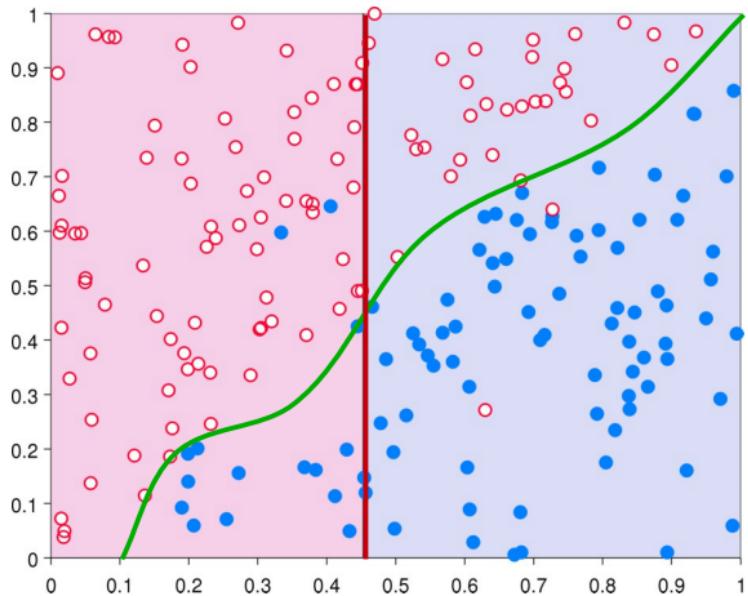
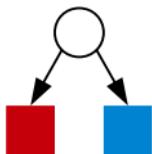
- Perfect fit
- **Too complex model**



Why are there “multiple models”?

Example: Classification tree

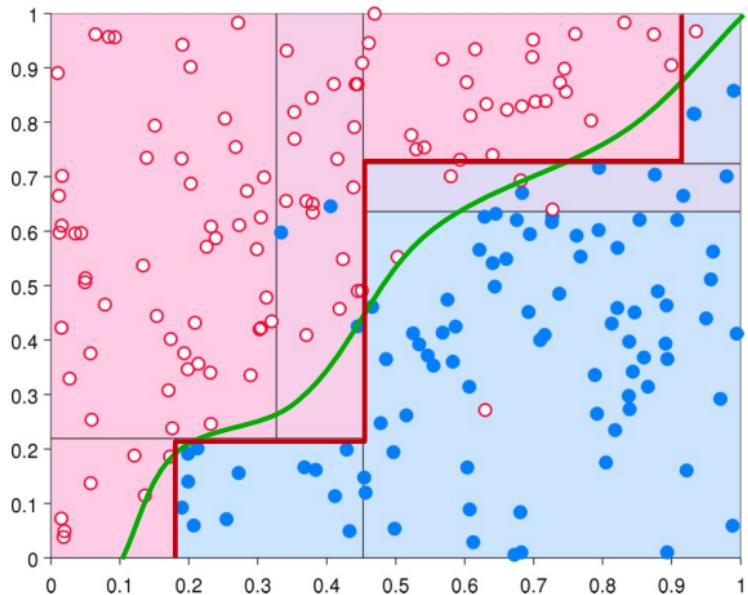
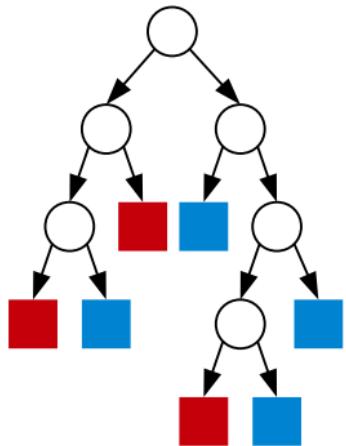
- Bad fit
- **Too simple model**



Why are there “multiple models”?

Example: Classification tree

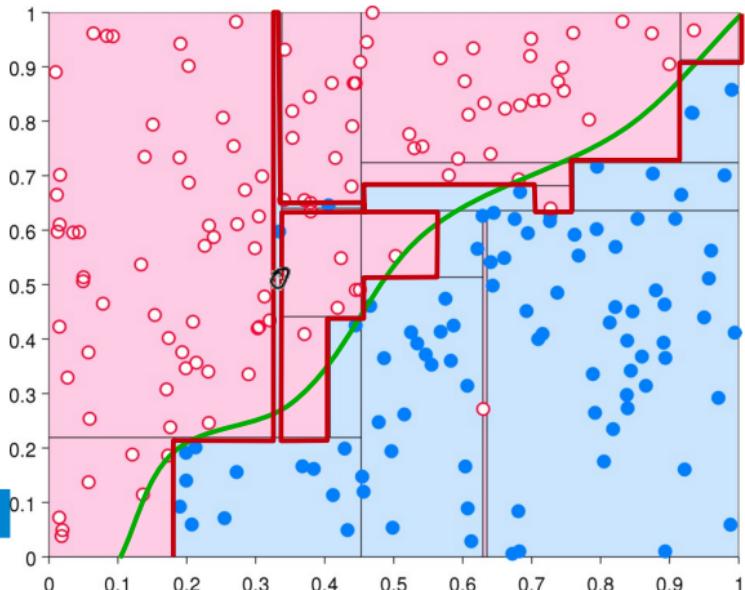
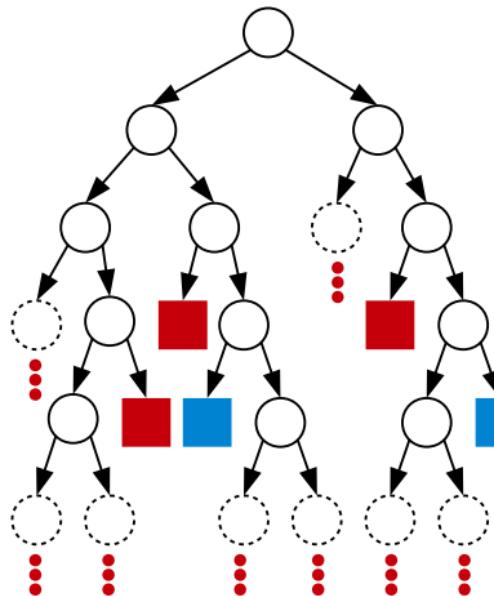
- Reasonable fit
- **Reasonable model**



Why are there “multiple models”?

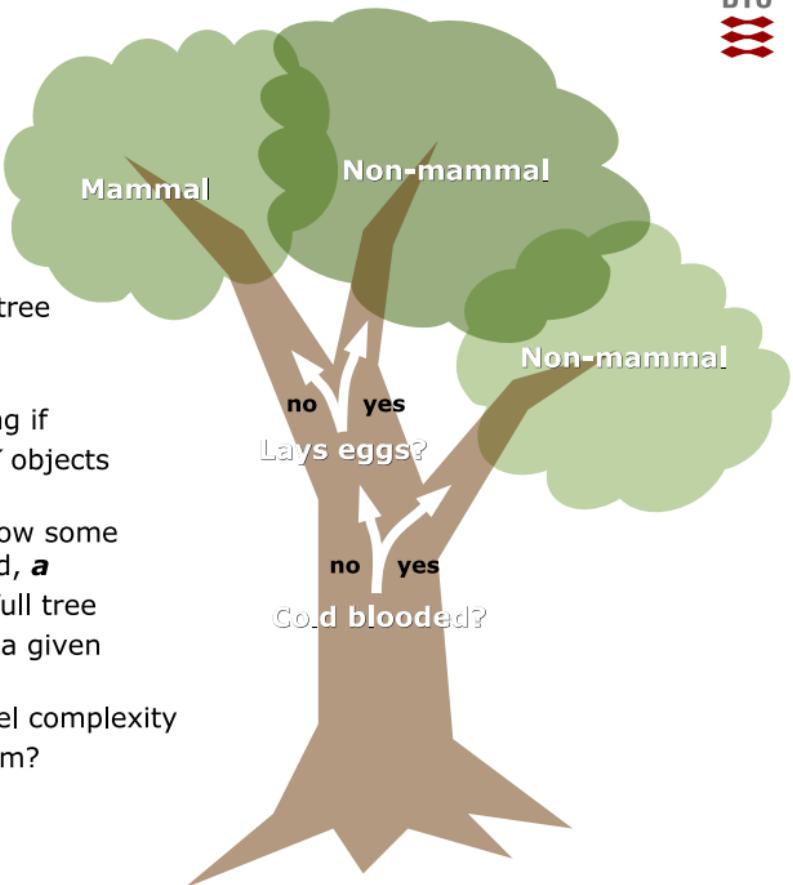
Example: Classification tree

- Perfect fit
- **Too complex model**

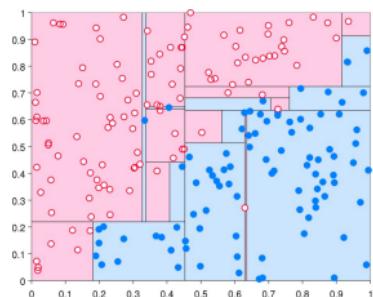
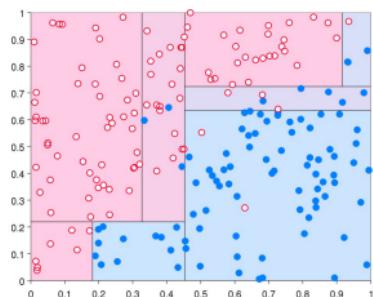
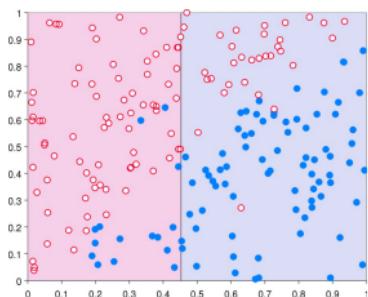
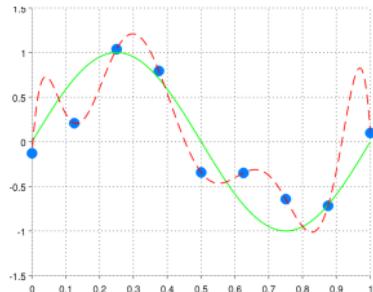
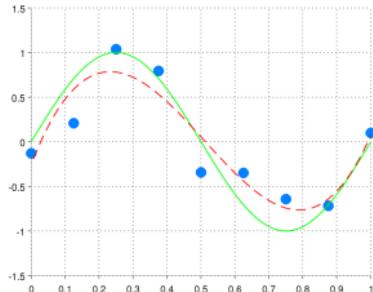
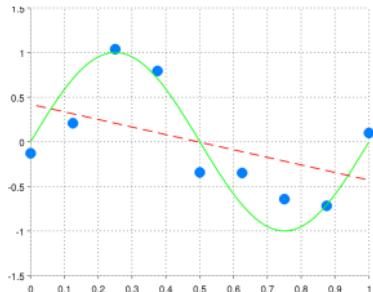


Decision trees

- Hunt's algorithm
 - Continue splitting until each node is pure
 - Results in a very complex tree (overfitting)
- **Control complexity**
 - **Pre-pruning:** Stop splitting if
 - There is less than K objects on the branch
 - Impurity gain is below some predefined threshold, a
 - **Post-pruning:** Generate full tree
 - Cut off branches to a given pruning level, c
- K , a , and/or c determine model complexity
 - How should we choose them?

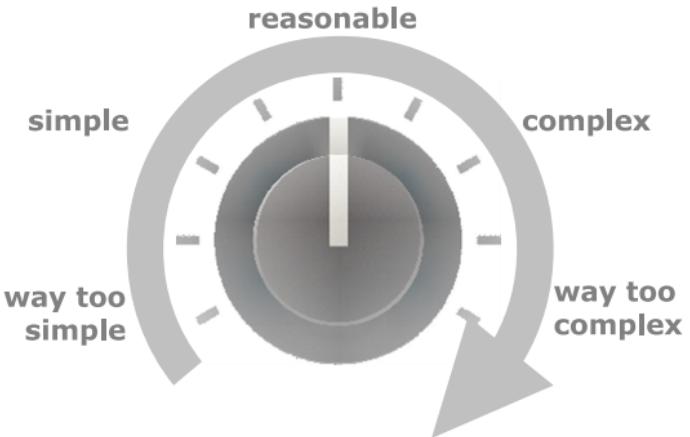


Model overfitting



Control the model complexity

- Find **parameter** or **mechanism** in model that controls complexity



Lex Parsimoniae, Law of parsimony



Given two models with same predictive performance, the simpler model is preferred over the more complex model
- William of Ockham (1288-1347)
(paraphrased)

https://commons.wikimedia.org/wiki/File:William_of_Ockham.png



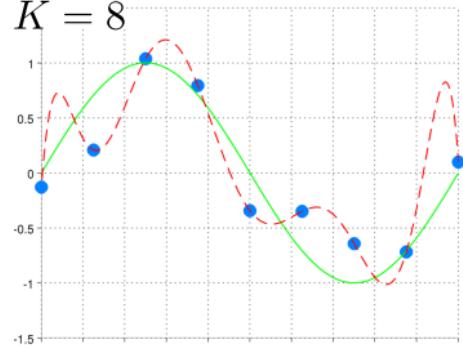
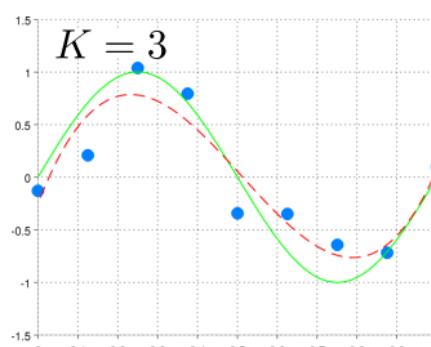
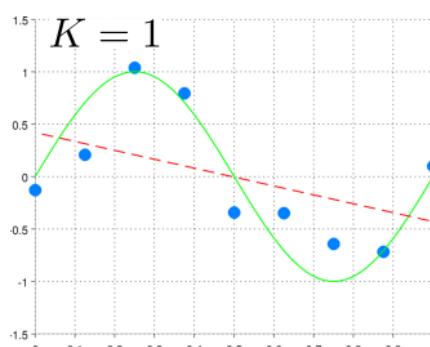
"Everything should be made as simple as possible, but not simpler" - Einstein

Linear regression

- Linear regression on non-linearly transformed inputs (polynomials)

$$f(x) = w_0 + w_1x + \cdots + w_8x^8$$

– **Control complexity:** Choose a suitable value for K



Solution:
Assess model performance correctly and select best model

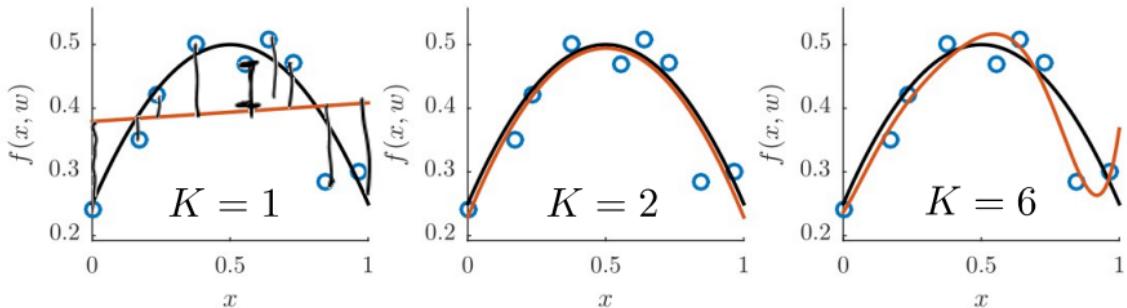
Training error

- Suppose we train 3 models on a dataset of 9 observations

$\mathcal{M}_1 = \{\text{1'st order polynomial}\}$

$\mathcal{M}_2 = \{\text{2'nd order polynomial}\}$

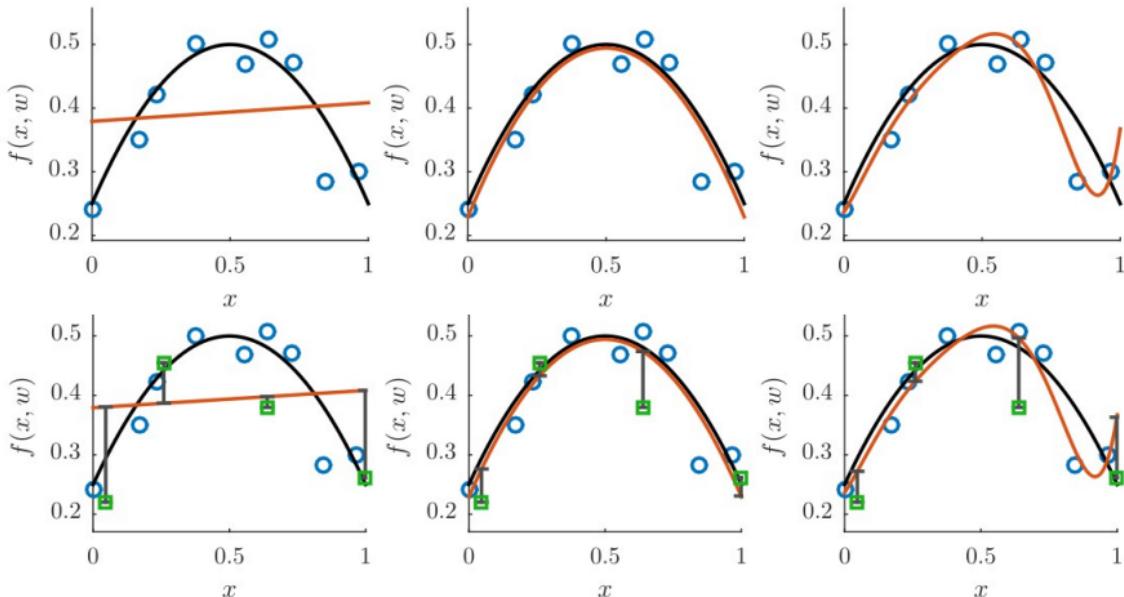
$\mathcal{M}_3 = \{\text{6'th order polynomial}\}$



$$E_{\mathcal{M}_k}^{\text{train}} = \frac{1}{N^{\text{train}}} \sum_{i \in \mathcal{D}^{\text{train}}} (y_i - f_{\mathcal{M}_k}(x_i, \mathbf{w}))^2.$$

Test error error

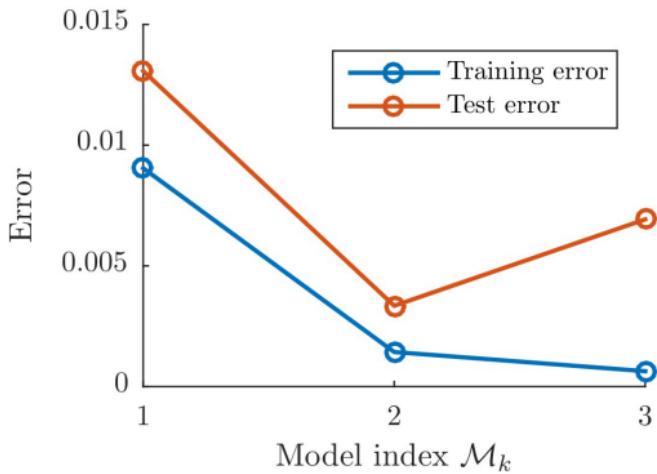
- Test error is obtained by testing the trained models on new data



$$E_{\mathcal{M}_k}^{\text{train}} = \frac{1}{N^{\text{train}}} \sum_{i \in \mathcal{D}^{\text{train}}} (y_i - f_{\mathcal{M}_k}(x_i, \mathbf{w}))^2.$$

$$E_{\mathcal{M}_k}^{\text{test}} = \frac{1}{N^{\text{test}}} \sum_{i \in \mathcal{D}^{\text{test}}} (y_i - f_{\mathcal{M}_k}(x_i, \mathbf{w}))^2$$

Overfitting

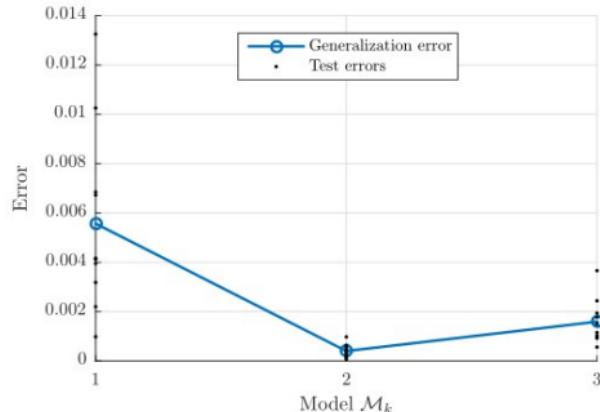
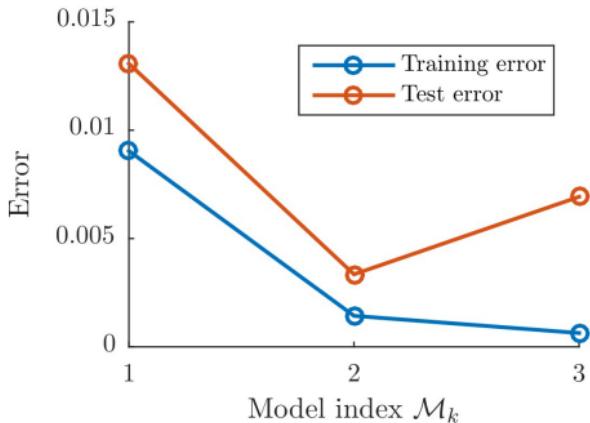


- **Overfitting** is that the training error usually decreases for overly complex models while the test error increases
- Test error is the more true error
- **Never, ever validate a model on the same data it was trained upon**

Generalization error

- The generalization error is the test error evaluated over an infinitely large test set
- The generalization error is the "true performance" of the trained model
 - Train model \mathcal{M} on the available dataset \mathcal{D} to get prediction rule $f_{\mathcal{M}}$
 - Compute $E_{\mathcal{M}}^{\text{gen}} = \mathbb{E}_{(\mathbf{x}, \mathbf{y})} [L(\mathbf{y}, f_{\mathcal{M}}(\mathbf{x}))] = \int p(x, y) L(y, f_{\mathcal{M}}(x)) dx dy$
- If we somehow had many test sets $\mathcal{D}_1, \dots, \mathcal{D}_k$

$$E_{\mathcal{M}}^{\text{gen}} \approx \frac{1}{K} \sum_{k=1}^K E_{\mathcal{M}, \mathcal{D}_k}^{\text{test}}$$



Basic cross-validation

- Purpose: Estimate the generalization error

Basic cross-validation

- **Purpose:** Estimate the generalization error

- 3 variants:

- **Holdout:** Partitions dataset in two (training, test), approximate the generalization error based on the generated test set

$$\mathcal{D} = \mathcal{D}^{\text{train}} \cup \mathcal{D}^{\text{test}}$$

$$E_{\mathcal{M}}^{\text{gen}} \approx E_{\mathcal{M}}^{\text{test}}$$

Holdout method

1/3 x N

Test
Training

2/3 x N

Basic cross-validation

- **Purpose:** Estimate the generalization error

- 3 variants:

- **Holdout:** Partitions dataset in two (training, test), approximate the generalization error based on the generated test set

$$\mathcal{D} = \mathcal{D}^{\text{train}} \cup \mathcal{D}^{\text{test}}$$

$$E_{\mathcal{M}}^{\text{gen}} \approx E_{\mathcal{M}}^{\text{test}}$$

- **K-fold:** Partitions dataset in K parts. Each part is a test set and the other K-1 training sets

$$\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \dots \cup \mathcal{D}_K$$
$$E_{\mathcal{M}}^{\text{gen}} \approx \frac{1}{K} \sum_{k=1}^K E_{\mathcal{M},k}^{\text{test}}$$

Holdout method

1/3 x N

Test
Training

2/3 x N

K-fold cross-validation (3-fold)



Basic cross-validation

- **Purpose:** Estimate the generalization error

- 3 variants:

- **Holdout:** Partitions dataset in two (training, test), approximate the generalization error based on the generated test set

$$\mathcal{D} = \mathcal{D}^{\text{train}} \cup \mathcal{D}^{\text{test}}$$

$$E_{\mathcal{M}}^{\text{gen}} \approx E_{\mathcal{M}}^{\text{test}}$$

- **K-fold:** Partitions dataset in K parts. Each part is a test set and the other K-1 training sets

$$\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \dots \cup \mathcal{D}_K$$

$$E_{\mathcal{M}}^{\text{gen}} \approx \frac{1}{K} \sum_{k=1}^K E_{\mathcal{M},k}^{\text{test}}$$

- **Leave-one-out:** Partitions dataset into N parts. Let each observation be a test set and the other N-1 training sets (K-fold with K=N)

$$\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \dots \cup \mathcal{D}_N$$

$$E_{\mathcal{M}}^{\text{gen}} \approx \frac{1}{N} \sum_{k=1}^N E_{\mathcal{M},k}^{\text{test}}$$

Holdout method

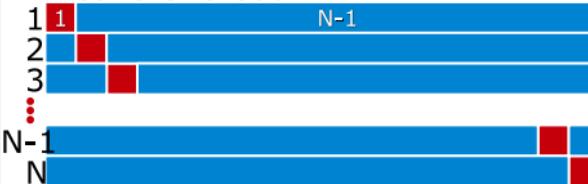


Test
Training

K-fold cross-validation (3-fold)



Leave-one-out

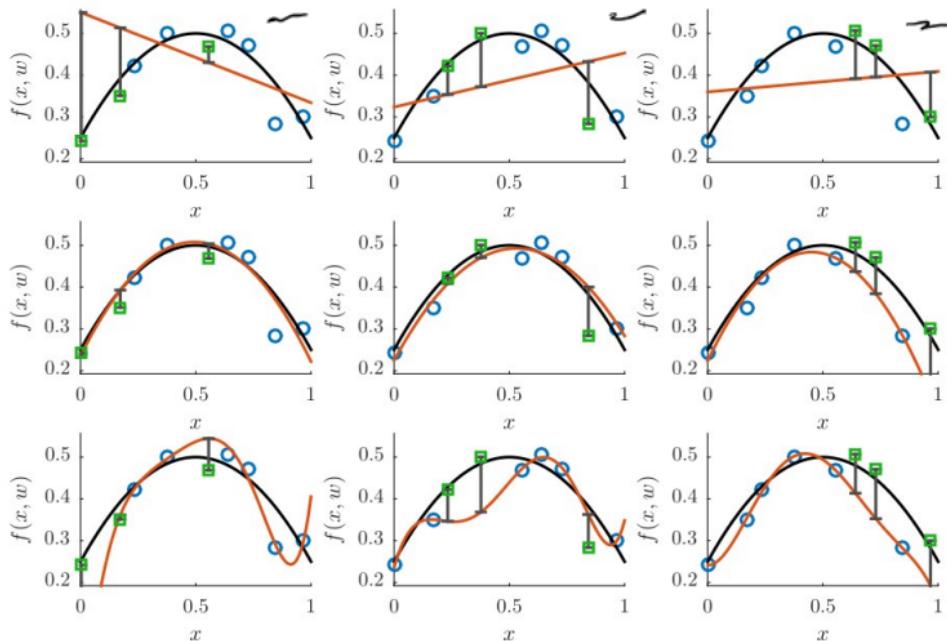


Cross-validation (1-layer)

- K=3 fold cross-validation for the three Linear-regression models

Vertically: The three models

Horizontally: The three cross-validation folds

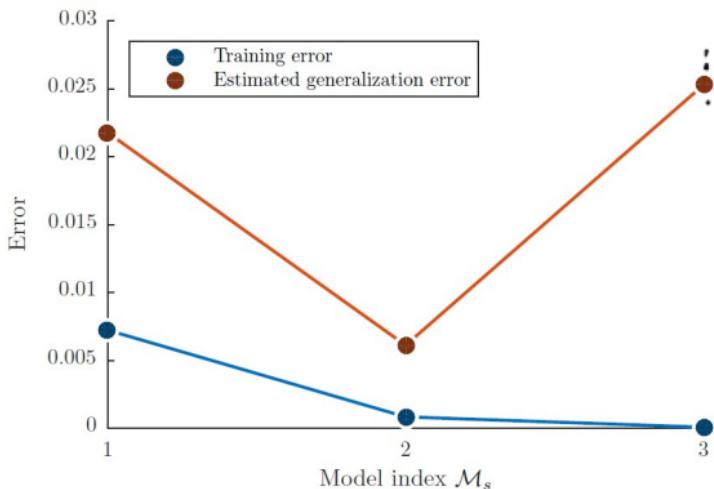


Cross-validation for model selection (1-layer)

- Purpose: Select the best of S models
- The idea:

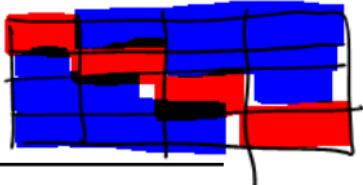
- For each model, estimate the cross-validation error $\hat{E}_{\mathcal{M}_1}^{\text{gen}}, \dots, \hat{E}_{\mathcal{M}_S}^{\text{gen}}$ using basic cross-validation.
- Select the optimal model \mathcal{M}_{s^*} as that with the lowest error:

$$s^* = \arg \min_s \hat{E}_{\mathcal{M}_s}^{\text{gen}}$$



Cross-validation (1-layer)

- K-fold cross-validation for model selection, the algorithm



Algorithm 4: K-fold cross-validation for model selection

Require: K , the number of folds in the cross-validation loop

Require: $\mathcal{M}_1, \dots, \mathcal{M}_S$. The S different models to select between

Ensure: \mathcal{M}_{s^*} the optimal model suggested by cross-validation

for $k = 1, \dots, K$ splits do

 Let $\mathcal{D}_k^{\text{train}}, \mathcal{D}_k^{\text{test}}$ the k 'th split of \mathcal{D}

 for $s = 1, \dots, S$ models do

 Train model \mathcal{M}_s on the data $\mathcal{D}_k^{\text{train}}$

 Let $E_{\mathcal{M}_s, k}^{\text{test}}$ be the *test error* of the model \mathcal{M}_s when it is *tested* on $\mathcal{D}_k^{\text{test}}$

 end for

end for

For each s compute: $\hat{E}_{\mathcal{M}_s}^{\text{gen}} = \sum_{k=1}^K \frac{N_k^{\text{test}}}{N} E_{\mathcal{M}_s, k}^{\text{test}}$

Select the optimal model: $s^* = \arg \min_s \hat{E}_{\mathcal{M}_s}^{\text{gen}}$

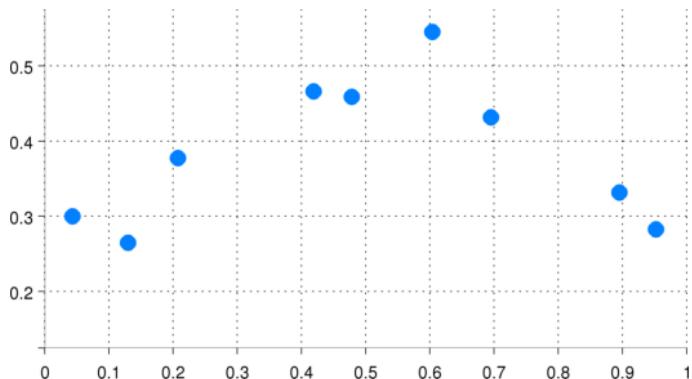
\mathcal{M}_{s^*} is now the optimal model suggested by cross-validation



| | A | B | C |
|---|---|---|---|
| 1 | - | - | - |
| 2 | - | - | - |
| 3 | - | - | - |
| 4 | - | - | - |
| | - | - | - |

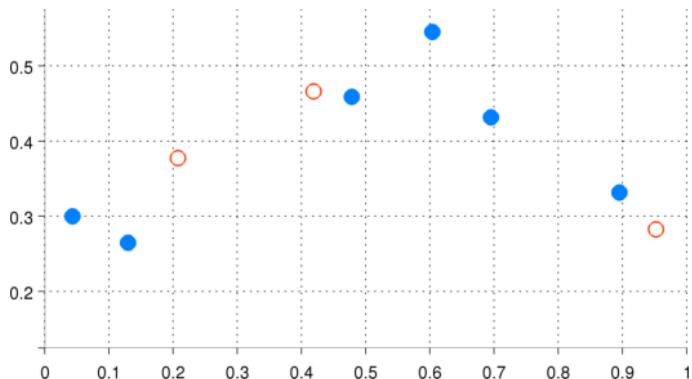
Holdout method

- Randomly choose a subset of data points to be in a **test set**
 - For example choose 1/3 of the points
- The rest is the **training set**



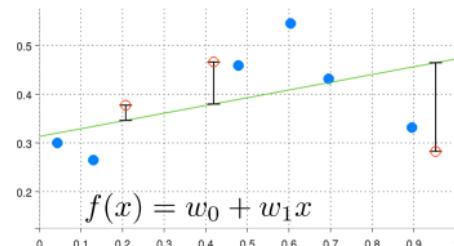
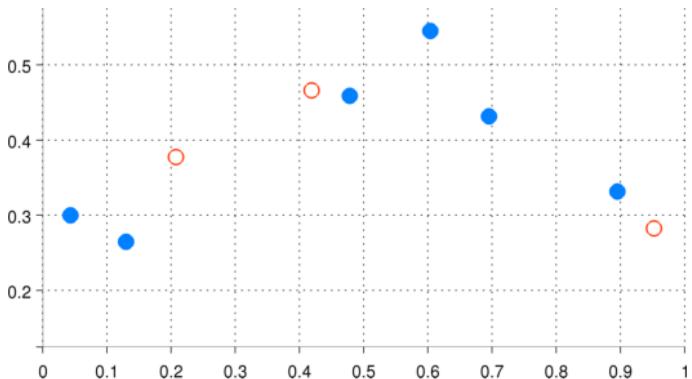
Holdout method

- Randomly choose a subset of data point to be in a **test set**
 - For example choose 1/3 of the points
- The rest is the **training set**

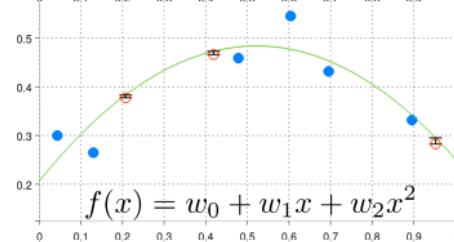


Holdout method

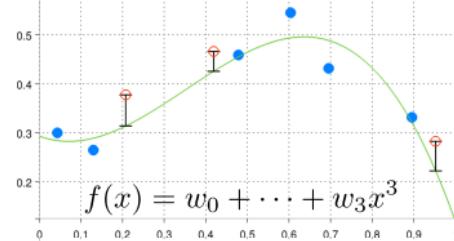
- Using the **training set**
 - Train the model for different complexities
- Using the **test set**
 - Compute the test error
- Choose the model with lowest **test error**



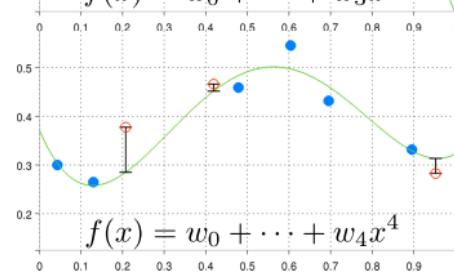
$$f(x) = w_0 + w_1x$$



$$f(x) = w_0 + w_1x + w_2x^2$$



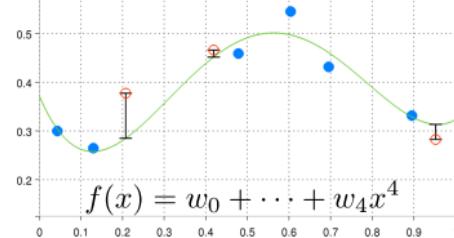
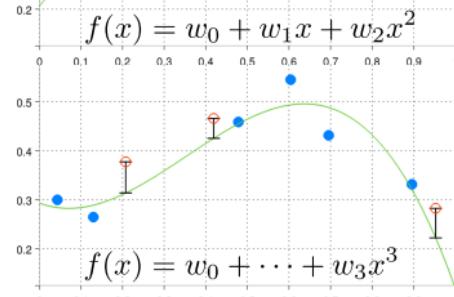
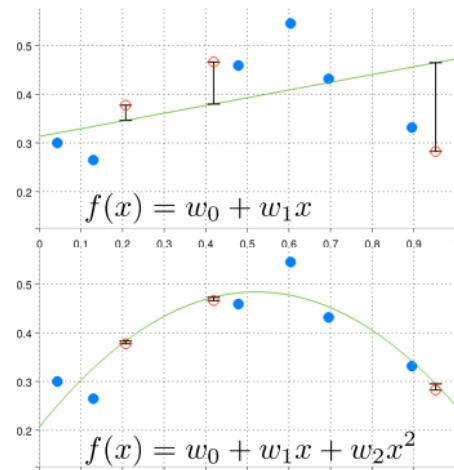
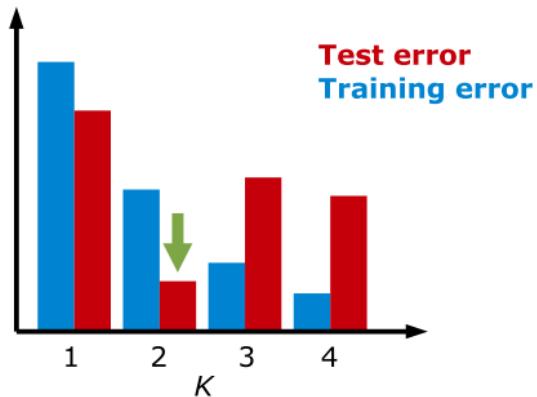
$$f(x) = w_0 + \dots + w_3x^3$$



$$f(x) = w_0 + \dots + w_4x^4$$

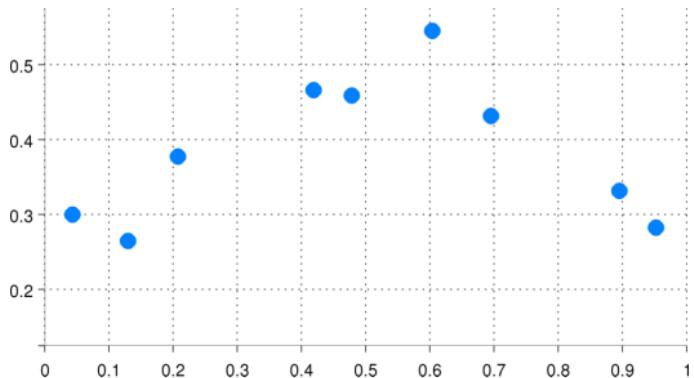
Holdout method

- Using the **training set**
 - Train the model for different complexities
- Using the **test set**
 - Compute the test error
- Choose the model with lowest **test error**



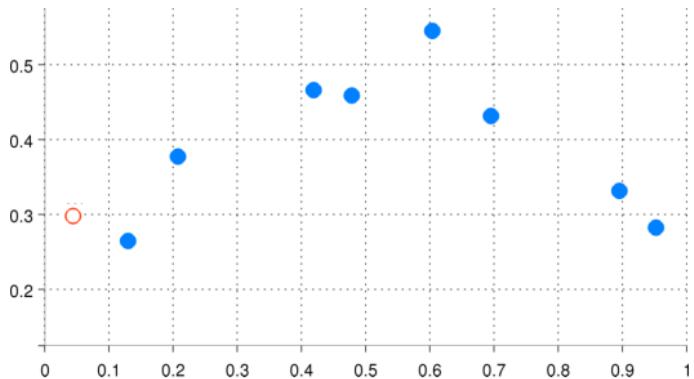
Leave-one-out

- Choose the first data point as a **test set**
- The rest is the **training set**



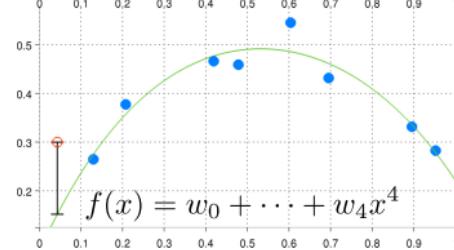
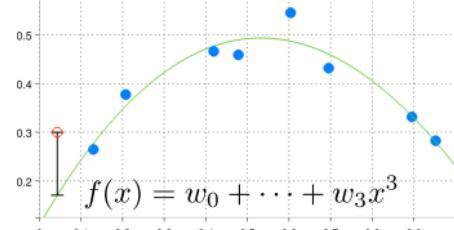
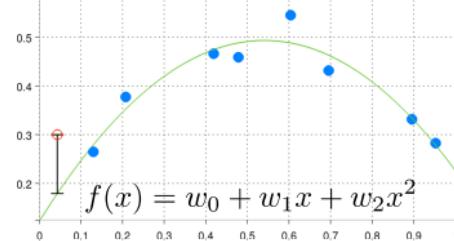
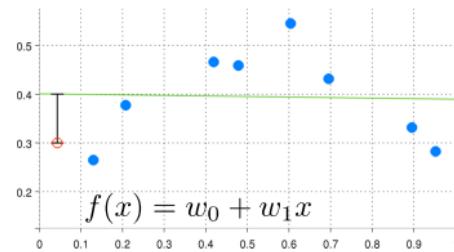
Leave-one-out

- Choose the first data point as a **test set**
- The rest is the **training set**



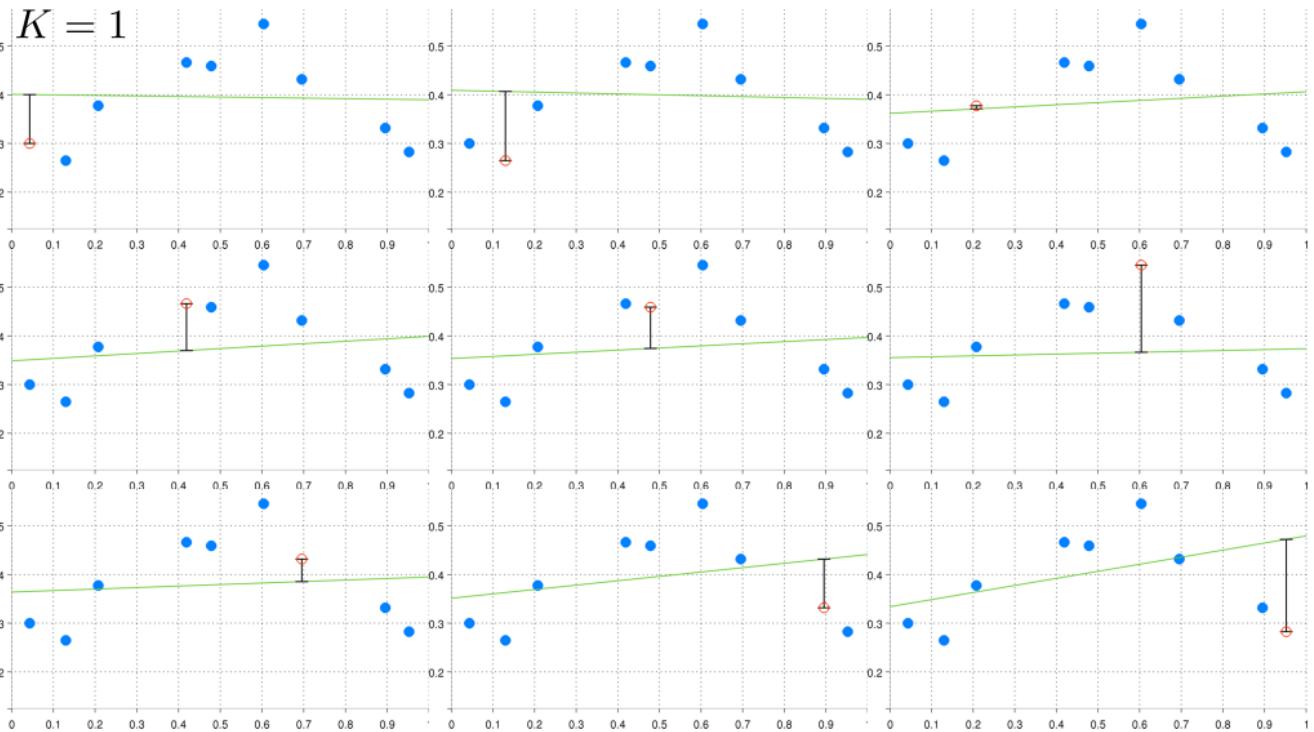
Leave-one-out

- Using the **training set**
 - Train the model for different complexities
- Using the **test set**
 - Compute the test error
- **Repeat for all data points**
 - All data points get to be test set
 - Compute **average test error**



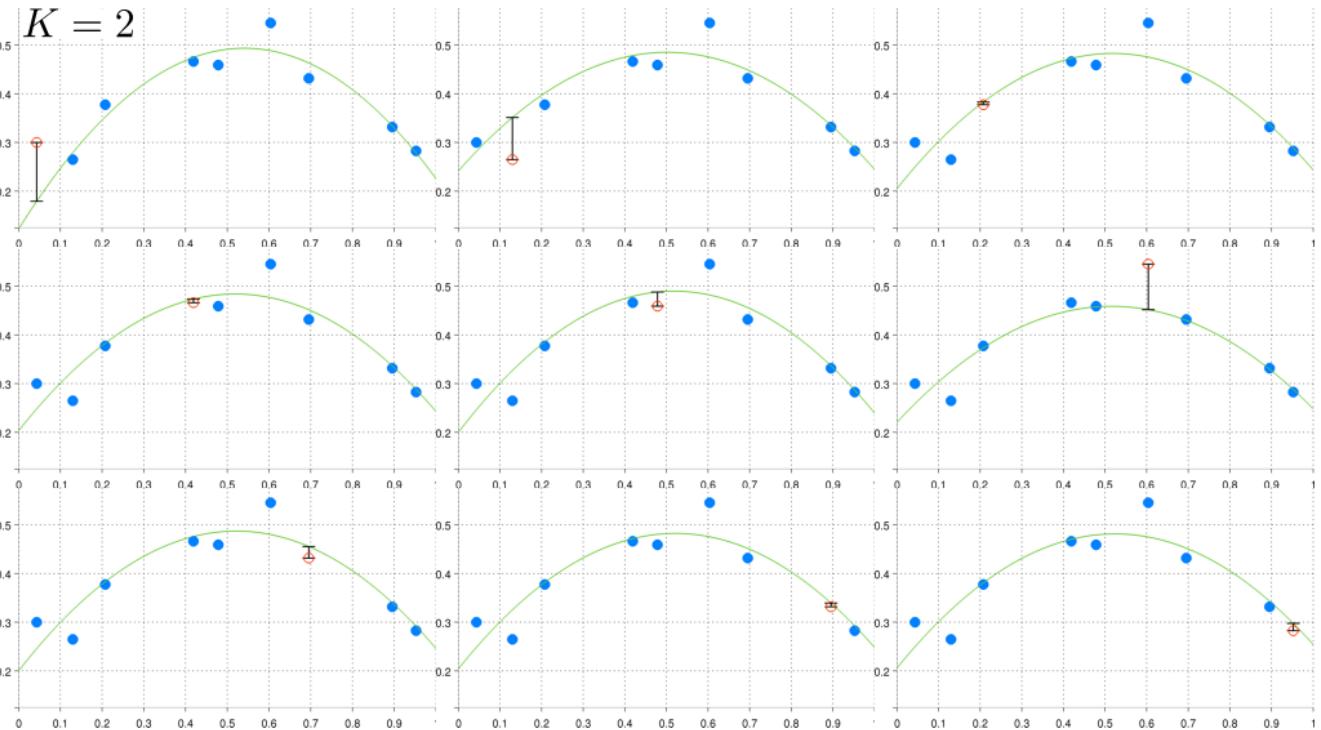
Leave-one-out

$K = 1$

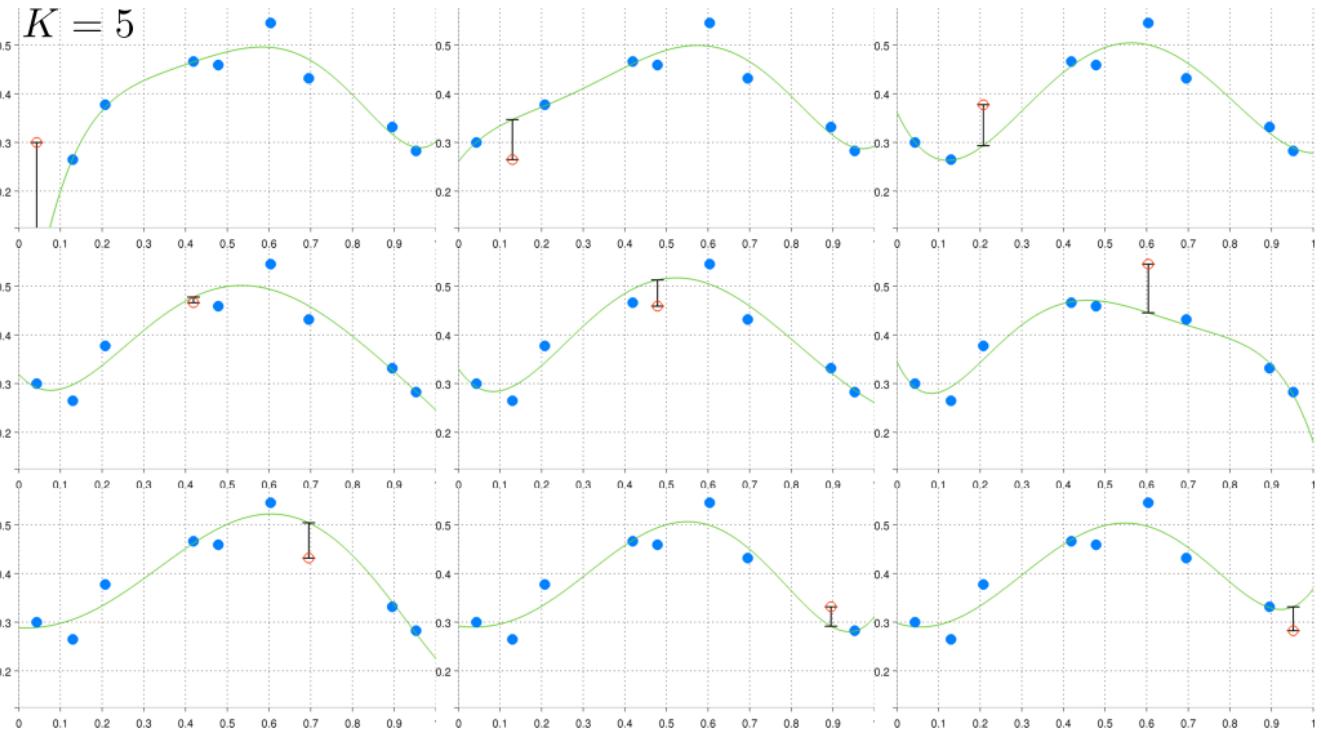


Leave-one-out

$K = 2$

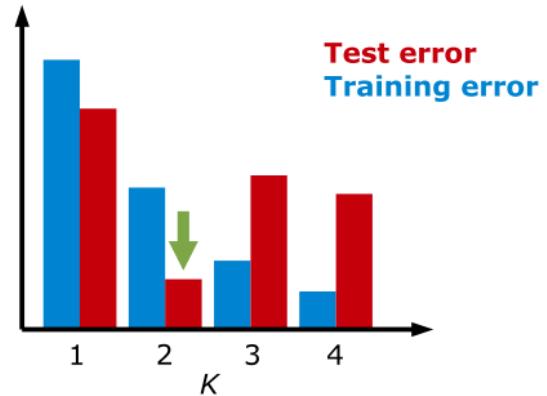


Leave-one-out cross-validation



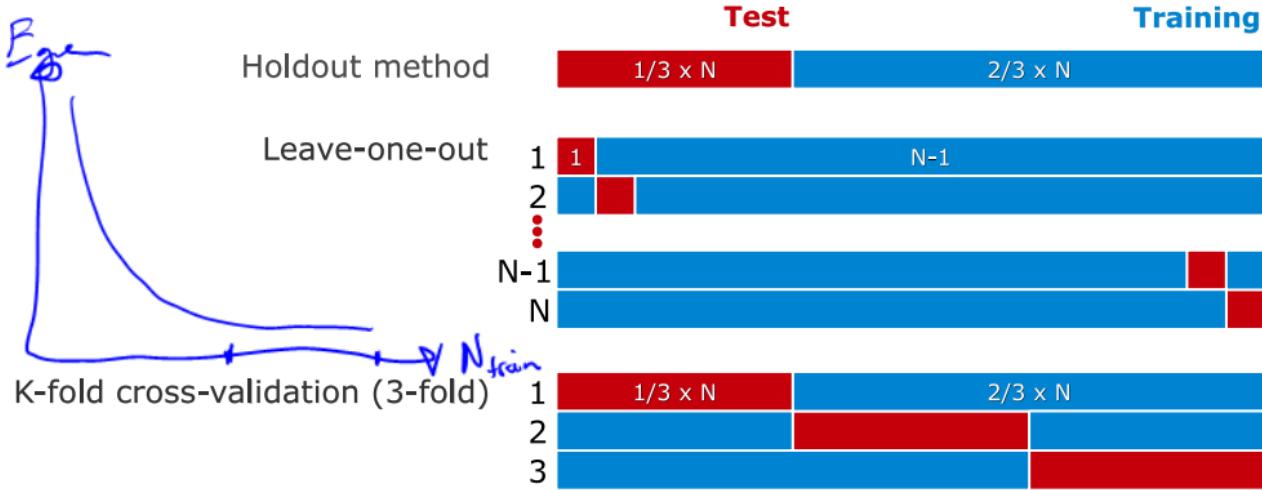
Leave-one-out

- Using the **training set**
 - Train the model for different complexities
- Using the **test set**
 - Compute the test error
- **Repeat for all data points**
 - All data points get to be test set
 - Compute **average test error**



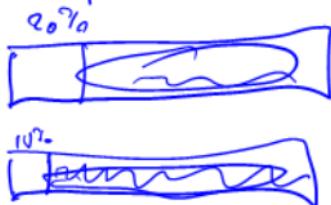
So which method should I choose?

- Holdout is computationally least intensive, but not very data efficient
 - Leave-one-out is very computationally intensive, and the estimates are highly correlated
 - Recommendation: K -fold for $K = 5, 10$ or holdout if problem very large.



Quiz 1, Cross validation (Spring 2012)

| Feature(s) | Training MSE | Test MSE |
|---------------------|--------------|----------|
| A | 2.0 | 2.2 |
| B | 1.8 | 1.9 |
| C | 1.6 | 1.7 |
| D | 1.9 | 2.1 |
| A and B | 1.7 | 2.0 |
| A and C | 1.3 | 1.8 |
| A and D | 1.4 | 1.5 |
| B and C | 1.5 | 1.6 |
| B and D | 1.7 | 1.8 |
| C and D | 1.5 | 2.0 |
| A and B and C | 1.2 | 2.1 |
| A and B and D | 1.1 | 2.0 |
| A and C and D | 1.0 | 2.3 |
| B and C and D | 1.2 | 2.5 |
| A and B and C and D | 0.9 | 2.8 |



Lin. reg.

Consider a neural network regression problem with four attributes denoted A, B, C and D. A neural network with two hidden units is trained using different combinations of the attributes. The neural network is trained on 50% of the data and tested on the remaining 50% of the data using the hold-out method. In the table is given the training and test performance of the neural network for the different combinations of attributes. Which of the following statements is incorrect

- A. Hold-out 50% of the data is more computationally efficient than 5 fold cross-validation. ✓
- B. Leave one-out cross-validation gives a poor estimate of the generalization error as only one observation is part of the test set at a time.
- C. The size of the training set in 10 fold cross-validation is larger than the size of the training set in 5 fold cross-validation.
- D. Not all observations are used for testing using the hold-out method. ✓
- E. Don't know.

In the hold-out method a given percentage, here 50% are removed for testing and the data trained on the remaining $100\%-50\% = 50\%$ of the data. This is more computationally efficient than 5 fold cross-validation as we only need to estimate one model for each combination of attributes. Leave-one-out cross-validation would however give a more precise

estimate of the generalization error as all observations are used in the test set once while keeping as many observations as possible for training, thus, second answer is incorrect. The size of the training set is larger in 10 fold cross-validation as 10 % of data is removed for testing whereas 20% of data is removed for testing for each fold of 5 fold cross-validation.

Forward selection



- Suppose we want to do linear regression
- As usual, we have M attributes

$$f(x) = w_0$$

$$f(x) = w_0 + w_1 x_1 + w_2 x_{27} + w_3 x_{88}$$

$$f(x) = w_0 + w_1 x_{19} + w_2 x_{76}$$

$$f(x) = w_0 + w_1 x_{19} + w_2 x_{76} + w_3 x_{88}$$

$$f(x) = w_0 + w_1 x_1 + w_2 x_{27} + w_3 x_{19}$$

$$f(x) = w_0 + w_1 x_{27} + w_2 x_{88}$$

$$x_1, x_2, \dots, x_M$$

⋮

- We can control model complexity by using a subset of attributes
 - Large subset: Complex model; hard to interpret
 - Small subset: Too simple model
- In general, we can construct 2^M models; often far too many
- Sequential feature selection allow us to efficiently search the model space

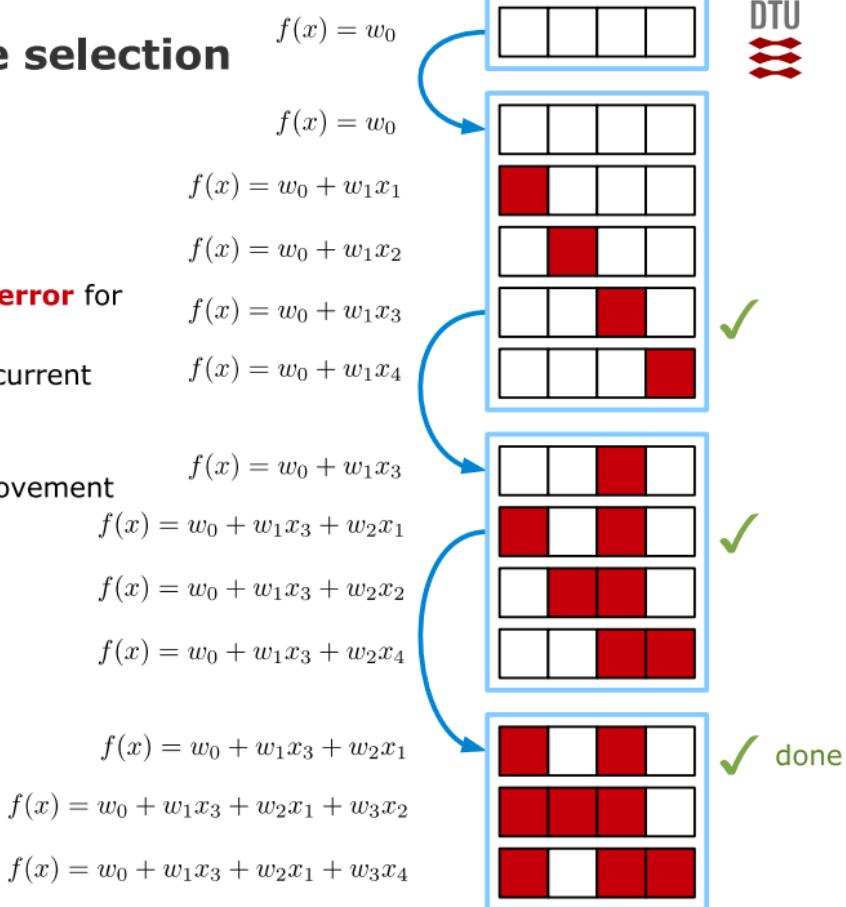
Sequential feature selection

Forward selection

- Start with no features
- Compute **cross-validation error** for
 - Current feature subset
 - All subsets equal to the current + one added feature
- Choose best subset
- Repeat until no further improvement

$$\frac{(M^2 + M)}{2}$$

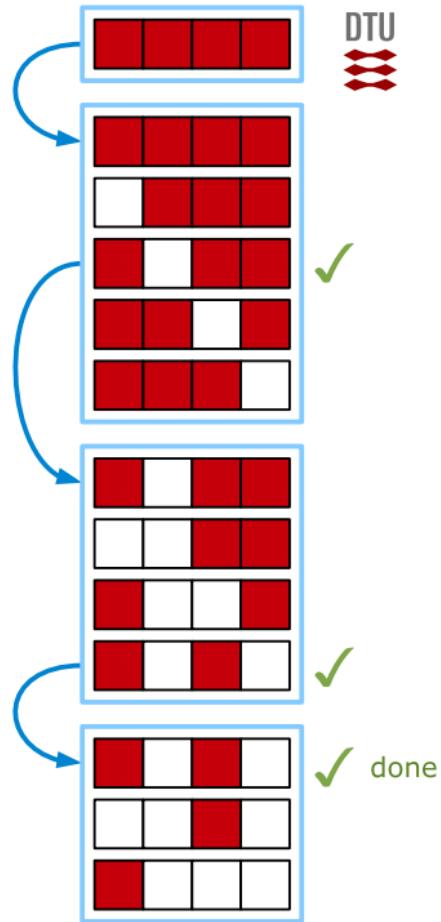
$$2^M$$



Sequential feature selection

Backward selection

- Start with all features
- Compute **cross-validation error** for
 - Current feature subset
 - All subsets equal to the current
 - one removed feature
- Choose best subset
- Repeat until no further improvement



Quiz 2, Forward selection (Spring 2012)

C, B

| Feature(s) | Training MSE | Test MSE |
|------------------------|--------------|----------|
| A | 2.0 | 2.2 |
| B | 1.8 | 1.9 |
| C | 1.6 | 1.7 |
| D | 1.9 | 2.1 |
| A and B | 1.7 | 2.0 |
| A and C <i>B</i> | 1.3 | 1.8 |
| A and D <i>B</i> | 1.4 | 1.5 |
| B and C <i>B</i> | 1.5 | 1.6 |
| B and D | 1.7 | 1.8 |
| C and D <i>B</i> | 1.5 | 2.0 |
| A and B and C <i>B</i> | 1.2 | 2.1 |
| A and B and D | 1.1 | 2.0 |
| A and C and D | 1.0 | 2.3 |
| B and C and D <i>B</i> | 1.2 | 2.5 |
| A and B and C and D | 0.9 | 2.8 |

Lin. reg.

Consider a neural network regression problem with four attributes denoted A, B, C and D where a neural network with two hidden units is trained using different combinations of the attributes. The table gives the training and test performance of the neural network for different combinations of attributes. Which of the following statements is *correct*?

- C*
- A. Using a forward selection strategy feature B and C would be selected as the optimal model.
 - B. Using a forward selection strategy features A and D would be selected as the optimal model.
 - C. Using a forward selection strategy features A and C and D would be selected as the optimal model.
 - D. Using a forward selection strategy features A and B and C would be selected as the optimal model.
 - E. Don't know.

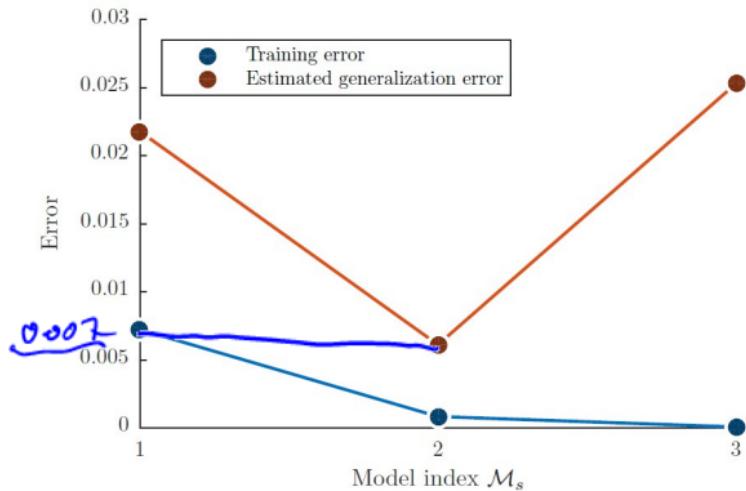
Using forward selection we would like to minimize the test error. Thus, the Forward selection would first select C and subsequently B terminating at the

solution B and C. Selecting additional attributes will not improve on the test error.

A problem with 1 level cross-validation?

- For each model, estimate the cross-validation error $\hat{E}_{\mathcal{M}_1}^{\text{gen}}, \dots, \hat{E}_{\mathcal{M}_S}^{\text{gen}}$ using basic cross-validation.
- Select the optimal model \mathcal{M}_{s^*} as that with the lowest error:

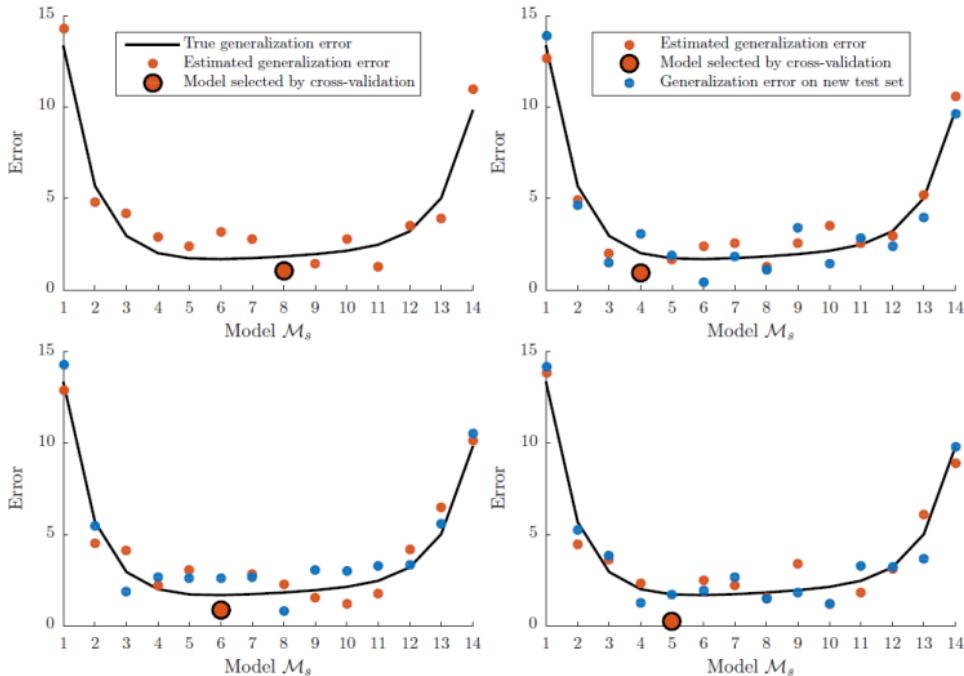
$$s^* = \arg \min_s \hat{E}_{\mathcal{M}_s}^{\text{gen}}$$



- Is the generalization error the selected model ($k=2$) about 0.007?

A problem with 1 level cross-validation?

- Same as before, just with more models. Is the error of the red dot a fair estimate of the generalization error?



Two-layer cross-validation

- Purpose: Select optimal model and estimate generalization error of optimal model

Two-layer cross-validation

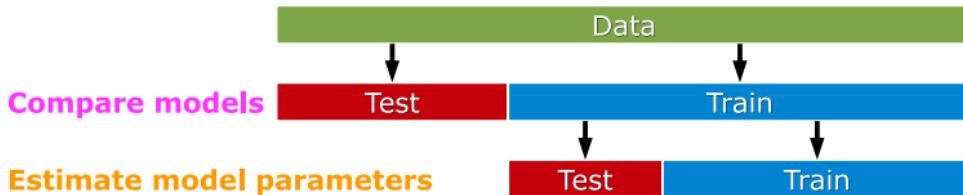
- Purpose: Select optimal model and estimate generalization error of optimal model
- How?
 - Recall "**one layer cross-validation for model selection**"
 - This method returns a model (the best model)
 - We can consider "**one-layer cross-validation for model selection**" as a single model

Two-layer cross-validation

- Purpose: Select optimal model and estimate generalization error of optimal model
- How?
 - Recall "**one layer cross-validation for model selection**"
 - This method returns a model (the best model)
 - We can consider "**one-layer cross-validation for model selection**" as a single model
- Recall:
 - "**Basic cross-validation for performance evaluation**" estimates the generalization error of a model

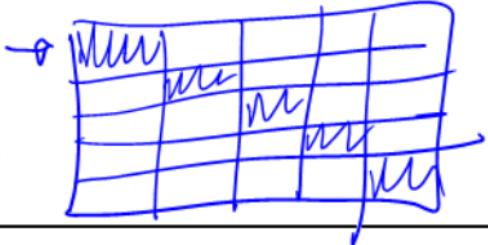
Two-layer cross-validation

- Purpose: Select optimal model and estimate generalization error of optimal model
- How?
 - Recall "**one layer cross-validation for model selection**"
 - This method returns a model (the best model)
 - We can consider "**one-layer cross-validation for model selection**" as a single model
- Recall:
 - "**Basic cross-validation for performance evaluation**" estimates the generalization error of a model
- Idea: Apply "**basic cross-validation for performance evaluation**" on the "**one-layer cross-validation for model selection**"-model to estimate it's generalization error



Cross-validation (2-layer)

- Two-layer cross-validation, the algorithm



Algorithm 5: Two-level cross-validation

Require: K_1, K_2 , folds in outer, and inner cross-validation loop respectively

Require: $\mathcal{M}_1, \dots, \mathcal{M}_S$: The S different models to cross-validate

Ensure: \hat{E}^{gen} , the estimate of the generalization error

→ **for** $i = 1, \dots, K_1$ **do**

Outer cross-validation loop. First make the outer split into K_1 folds

Let $\mathcal{D}_i^{\text{par}}, \mathcal{D}_i^{\text{test}}$ be the i 'th split of \mathcal{D}

for $j = 1, \dots, K_2$ **do**

Inner cross-validation loop. Use cross-validation to select optimal model

Let $\mathcal{D}_j^{\text{train}}, \mathcal{D}_j^{\text{val}}$ be the j 'th split of $\mathcal{D}_i^{\text{par}}$

for $s = 1, \dots, S$ **do**

Train \mathcal{M}_s on $\mathcal{D}_j^{\text{train}}$

Let $E_{\mathcal{M}_s, j}^{\text{val}}$ be the validation error of the model \mathcal{M}_s when it is tested on $\mathcal{D}_j^{\text{val}}$

end for

end for

For each s compute: $\hat{E}_s^{\text{gen}} = \sum_{j=1}^{K_2} \frac{|\mathcal{D}_j^{\text{val}}|}{|\mathcal{D}_i^{\text{par}}|} E_{\mathcal{M}_s, j}^{\text{val}}$

Select the optimal model $\mathcal{M}^* = \mathcal{M}_{s^*}$ where $s^* = \arg \min_s \hat{E}_s^{\text{gen}}$

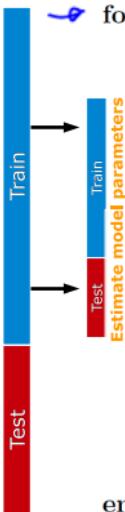
Train \mathcal{M}^* on $\mathcal{D}_i^{\text{par}}$

Let E_i^{test} be the test error of the model \mathcal{M}^* when it is tested on $\mathcal{D}_i^{\text{test}}$

end for

Compute the estimate of the generalization error: $\hat{E}^{\text{gen}} = \sum_{i=1}^{K_1} \frac{|\mathcal{D}_i^{\text{test}}|}{N} E_i^{\text{test}}$

Compare models



Quiz 3, two-level cross-validation (Spring 2016)

Consider a classification tree model applied to a dataset of $N = 1000$ observations. Suppose we wish to both select the optimal pruning level and estimate the generalization error of the classification tree model by cross-validation. To simplify the problem, we only consider 3 possible pruning levels:

$$3, 4, 5.$$

We opt for a two-level cross-validation strategy in which we use an inner loop of $K_2 = 5$ -fold cross-validation to estimate the optimal pruning level and an outer loop of $K_1 = 10$ fold cross-validation to estimate the generalization error. That is, for each of the K_1 outer folds, the dataset is divided into

a validation set and a parameter estimation set on which K_2 -fold cross-validation is used to select the optimal pruning level for this outer fold.

How many models do we *train* using 2-level cross-validation?

- A. 50
- B. 150
- C. 160
- D. 180
- E. Don't know.

This can easily be obtained noting for each of the K_1 outer folds we must both (i) train K_2 models on the $L = 3$ different settings of pruning level (ii) train a single new model to estimate the generalization

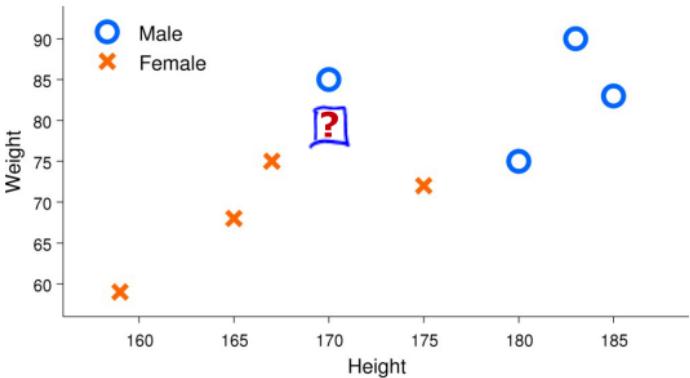
error for this fold. Accordingly the number of trained models is

$$K_1(K_2L + 1).$$

Therefore, option C is correct.

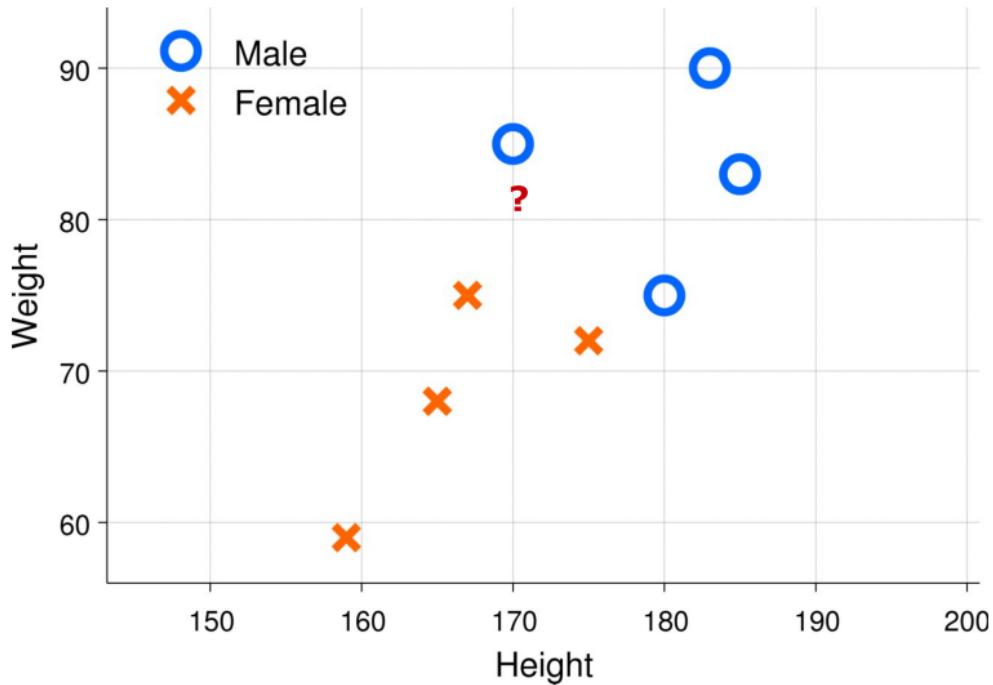
Classify gender based on height and weight

| | Height | Weight | Gender |
|---|--------|--------|--------|
| 1 | 183 | 90 | Male |
| 2 | 180 | 75 | Male |
| 3 | 170 | 85 | Male |
| 4 | 185 | 83 | Male |
| 5 | 159 | 59 | Female |
| 6 | 167 | 75 | Female |
| 7 | 165 | 68 | Female |
| 8 | 175 | 72 | Female |
| 9 | 171 | 82 | ? |



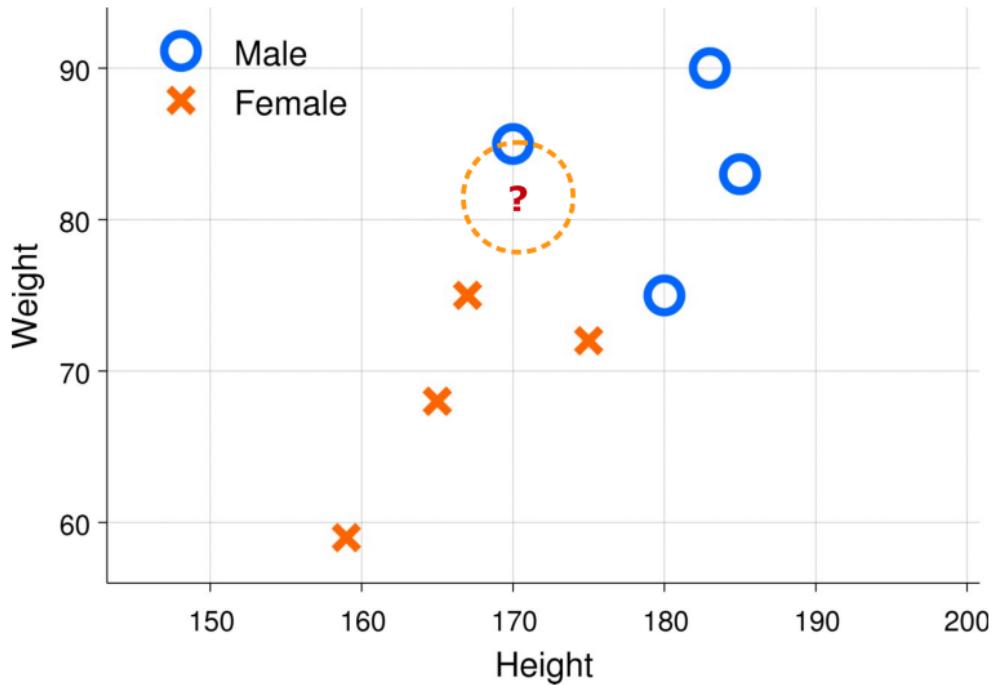
Nearest neighbor classifier

- 1 nearest neighbor



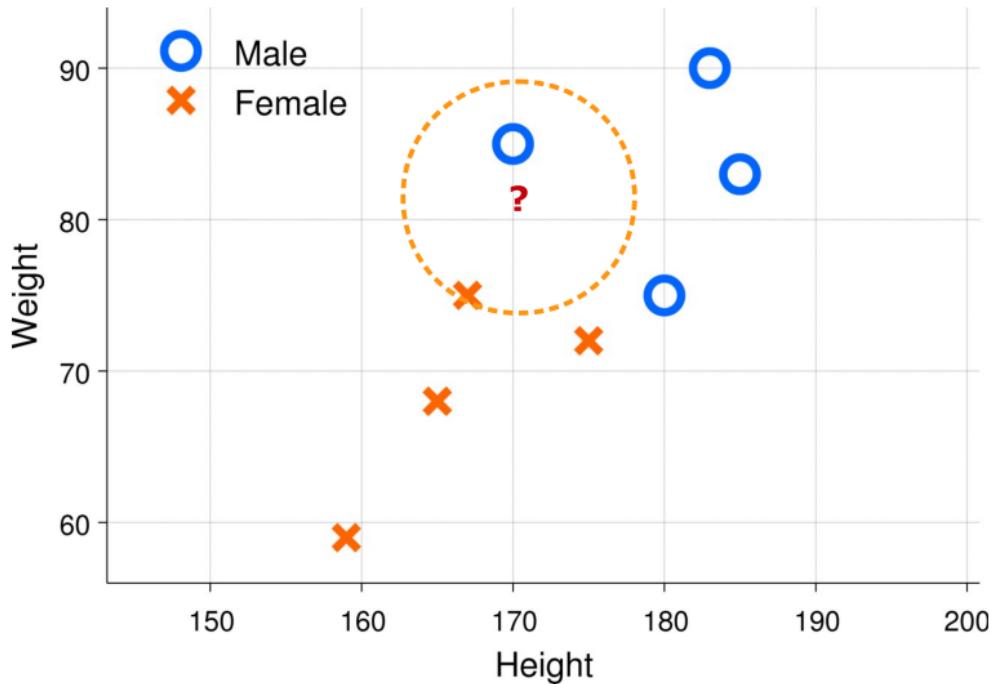
Nearest neighbor classifier

- 1 nearest neighbor



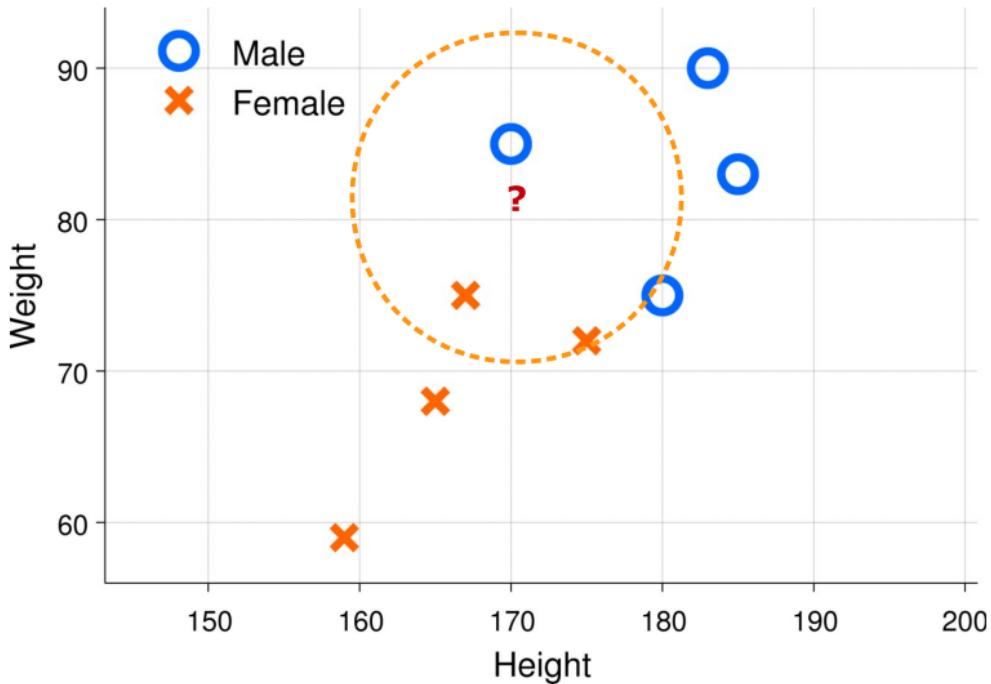
Nearest neighbor classifier

- 2 nearest neighbors



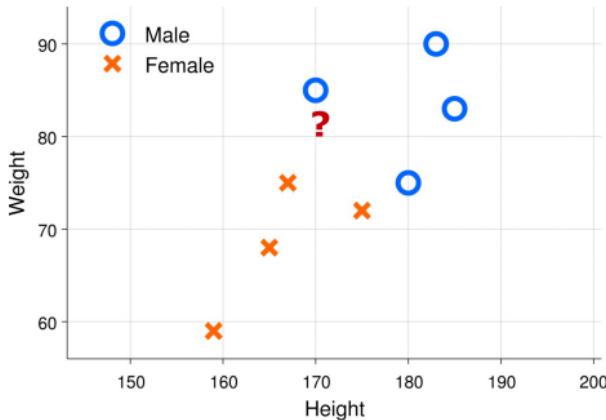
Nearest neighbor classifier

- 3 nearest neighbors

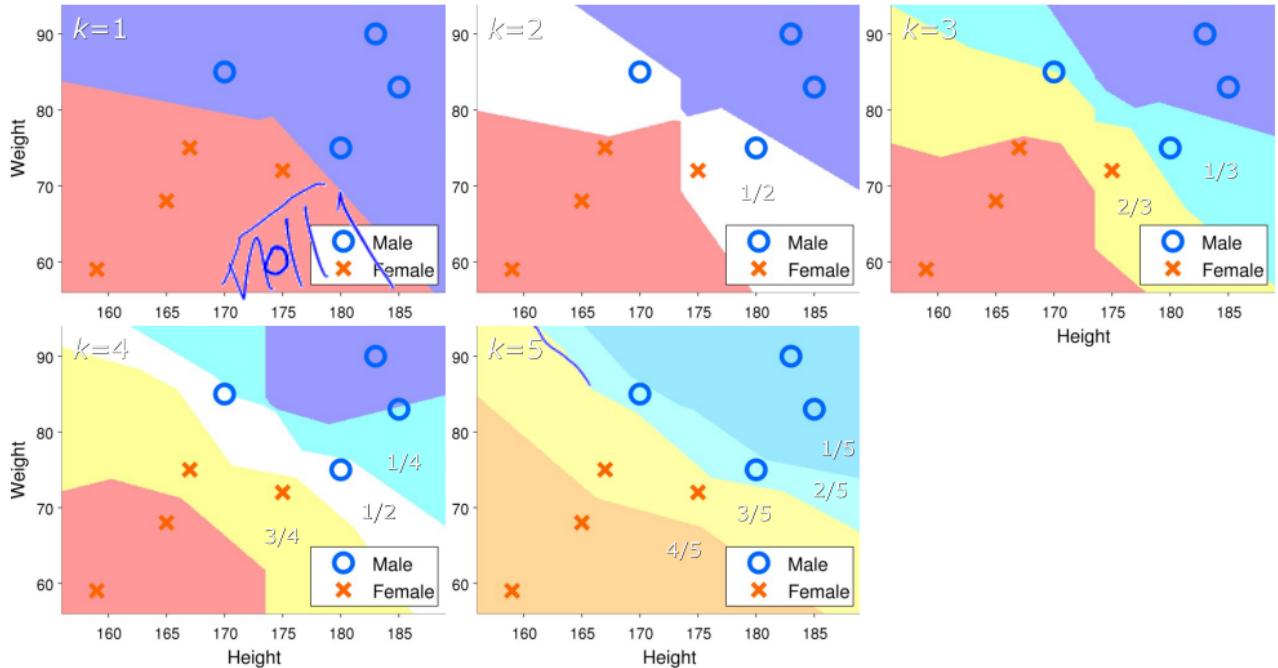


Nearest neighbor classifier

- Choose
 - The number of neighbors, k
 - A distance measure
1. Compute distance to all other data objects
 2. Find the k nearest data objects
 3. Classify according to majority of neighbors



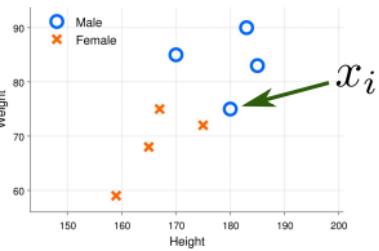
Nearest neighbor decision surface



KNN with leave-one-out CV

Leave-one-out CV is convenient with KNN

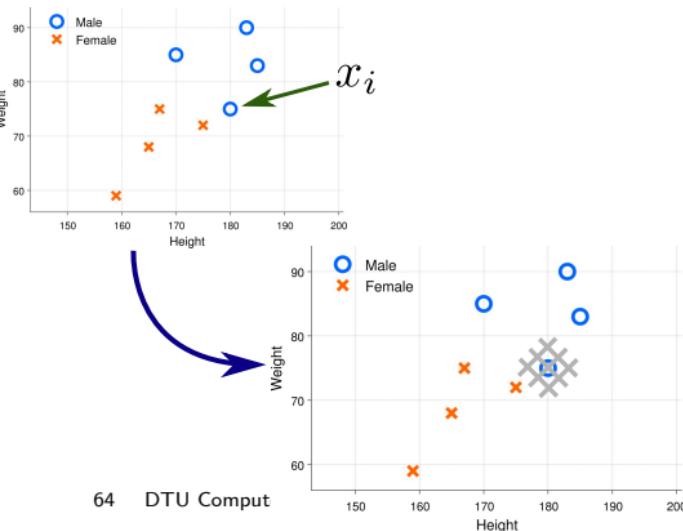
- For each observation x_i



KNN with leave-one-out CV

Leave-one-out CV is convenient with KNN

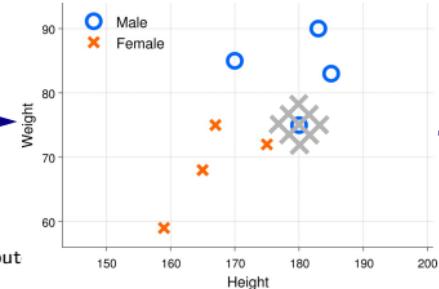
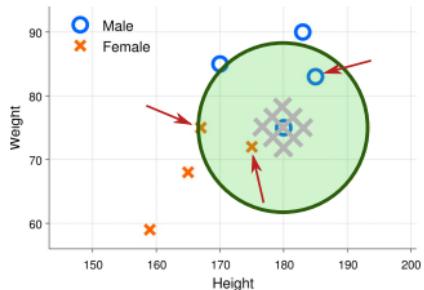
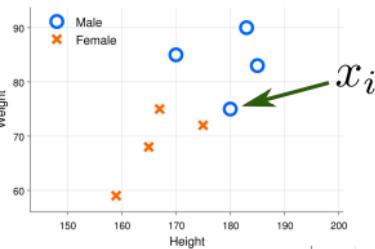
- For each observation x_i
 - Temporarily remove x_i



KNN with leave-one-out CV

Leave-one-out CV is convenient with KNN

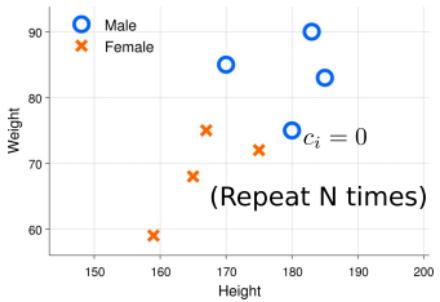
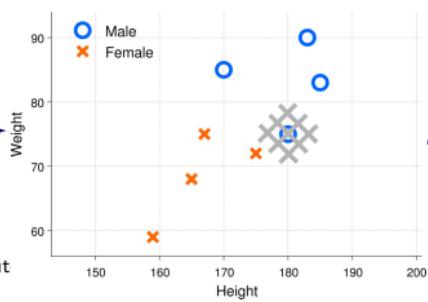
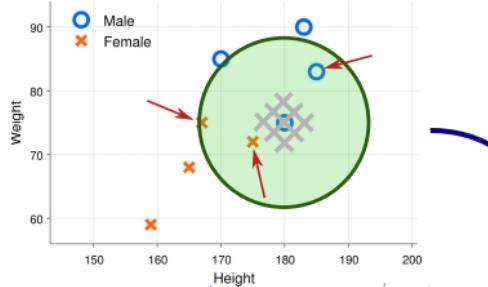
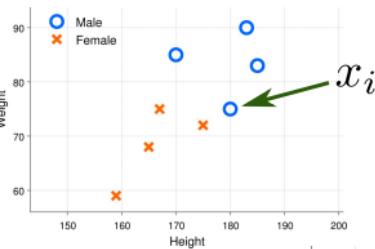
- For each observation x_i
 - Temporarily remove x_i
 - Find K nearest neighbors around x_i (not x_i itself)



KNN with leave-one-out CV

Leave-one-out CV is convenient with KNN

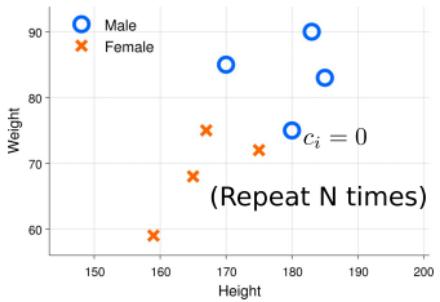
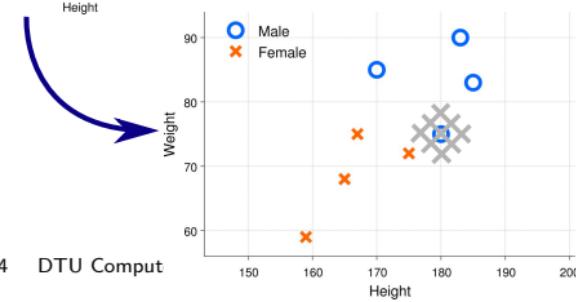
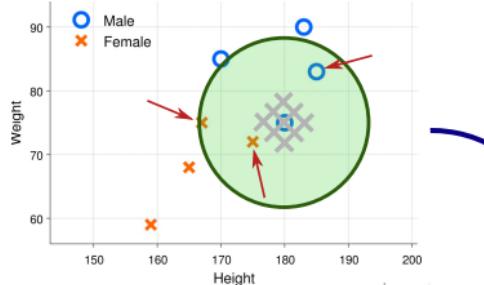
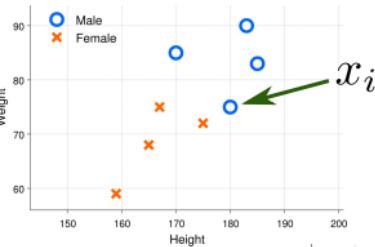
- For each observation x_i
 - Temporarily remove x_i
 - Find K nearest neighbors around x_i (not x_i itself)
 - Determine whether x_i is classified correctly based on this neighborhood, $c_i = 0, 1$



KNN with leave-one-out CV

Leave-one-out CV is convenient with KNN

- For each observation x_i
 - Temporarily remove x_i
 - Find K nearest neighbors around x_i (not x_i itself)
 - Determine whether x_i is classified correctly based on this neighborhood, $c_i = 0, 1$
- Compute accuracy as $\frac{1}{N} \sum_{i=1}^N c_i$



Quiz 4, KNN (Spring 2011)

- For each observation x_i
 - Temporarily remove x_i
 - Find K nearest neighbors around x_i (not x_i itself)
 - Determine whether x_i is classified correctly based on this neighborhood

| | | Diabetic | | | | Normal | | | |
|----------|----|----------|------|------|------|--------|------|------|------|
| | | D1 | D2 | D3 | D4 | N1 | N2 | N3 | N4 |
| Diabetic | D1 | 0 | 58.5 | 51.6 | 18.1 | 38.0 | 52.5 | 71.7 | 50.7 |
| | D2 | 58.5 | 0 | 32.1 | 72.6 | 50.5 | 65.0 | 13.2 | 63.8 |
| | D3 | 51.6 | 32.1 | 0 | 60.5 | 28.4 | 32.9 | 45.3 | 56.3 |
| | D4 | 18.1 | 72.6 | 60.5 | 0 | 45.9 | 60.4 | 79.8 | 56.8 |
| Normal | N1 | 38.0 | 50.5 | 28.4 | 45.9 | 0 | 17.5 | 63.7 | 50.7 |
| | N2 | 52.5 | 65.0 | 32.9 | 60.4 | 17.5 | 0 | 78.2 | 57.2 |
| | N3 | 71.7 | 13.2 | 45.3 | 79.8 | 63.7 | 78.2 | 0 | 71.0 |
| | N4 | 50.7 | 63.8 | 56.3 | 56.8 | 50.7 | 57.2 | 71.0 | 0 |

The figure shows the distance between the first four diabetic (D1–D4) and normal (N1–N4) women. What are the number of misclassified observations for leave-

one-out cross validation based on 3-nearest neighbor classification when only considering the 8 observations (i.e., D1–D4 and N1–N4) in the figure?

- A. None of the observations will be misclassified.
- B. 2 of the observations will be misclassified.
- C. 6 of the observations will be misclassified.
- D. All of the observations will be misclassified.

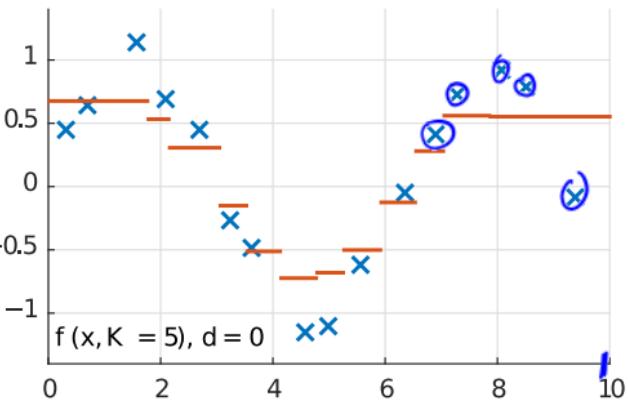
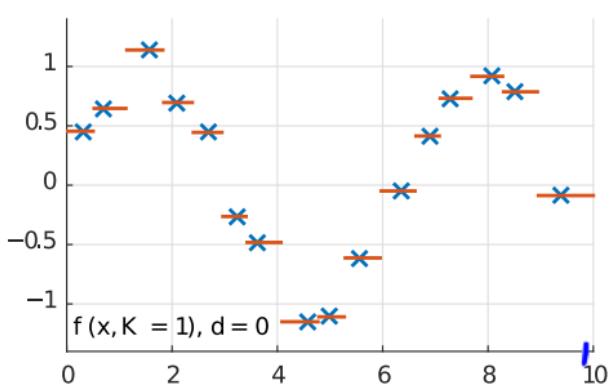
Based on 3-nearest neighbor classification all the observations will be misclassified, for instance D1's three nearest neighbors are D4, N1 and N4 thus there

are two normal woman and one diabetic woman as nearest neighbors, thus, a majority of normal observations such that D1 will be classified as normal.

KNN Regression

- Given a training set \mathbf{X}, \mathbf{y}
- For a test observation x predict the average y -value in the neighbourhood

$$\hat{y} = f(\mathbf{x}, K) = \frac{1}{K} \sum_{i \in N_{\mathbf{X}}(\mathbf{x}, K)} y_i$$



Resources

<https://towardsdatascience.com> Alternative introduction to cross-validation

(<https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85>)



02450: Introduction to Machine Learning and Data Mining

Decision trees and linear regression

Bjørn Sand Jensen

DTU Compute, Technical University of Denmark (DTU)

A collage of mathematical symbols including integrals, summation, infinity, and various numbers.

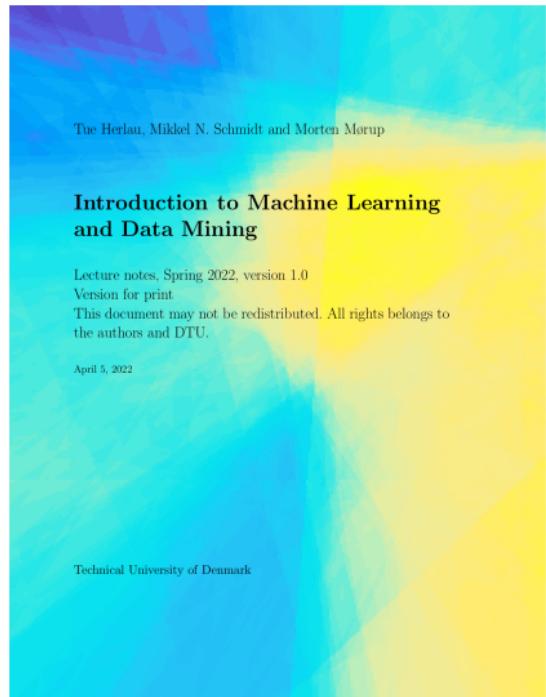
Today

Feedback Groups of the day:

Réka Farkas, Åiax Faura Vilalta, Max Sébastien Favrot, Farras Muhammad Fikry, Jørgen Finsveen, Jakob Astrup Fischer, Eduardo Frances Torres, Jacob Tækker Fredensborg, Konstantina Freri, Benedikte Friis, Rasmus Juel Friis, Eva Frossard, Hugh Fulton, Julia Gabriela Makulec, Natasha Garibay, Onofre Javier Garrido Mansoa, Philip Bentkjær Gaub, Georgios Georgiou, Antonia Sophie Gerstenberg, Omid Ghaiby, Mouna Ghiyati Ibn Ziyad, Shaghayegh Gholami Hatkehlouei, Aditya Vijendra Girase, Myrsini Gkolemi, Jonas Lolk Glymov, Adrien Goldszal, Søren Rønnekær Holgreen Graae, Christoffer Grauballe, Isabelle Marie Grimaldi, Julius Ellegård Grønager, Aleks Laith Gryn, Yikai Gu, Alma Kvist Gude, Arnór Gunnarsson, Artur Habuda, Harris Nielsen Hadzimahovic, Bjørn Hagbarth, Oliver Alexander Hagel, Frederik Olsen Halling, Malene Ræbild Hamburger, Chunxu Han, Lukas Hanisch, David Hansen, Alex Richard Ejby Hansen

Reading material:

Chapter 8, Chapter 9



Lecture Schedule

1 Introduction

30 January: C1

Data: Feature extraction, and visualization

2 Data, feature extraction and PCA

6 February: C2, C3

3 Measures of Similarity, summary statistics and probabilities

13 February: C4, C5

4 Probability densities and data visualization

20 February: C6, C7

Supervised learning: Classification and regression

5 Decision trees and linear regression

27 February: C8, C9 (Project 1 due 29 February at 17:00)

6 Overfitting, cross-validation and Nearest Neighbor

5 March: C10, C12

7 Performance evaluation, Bayes, and Naive Bayes

12 March: C11, C13

8 Artificial Neural Networks and Bias/Variance

19 March: C14, C15

9 AUC and ensemble methods

2 April: C16, C17

Unsupervised learning: Clustering and density estimation

10 K-means and hierarchical clustering

9 April: C18 (Project 2 due 11 April at 17:00)

11 Mixture models and density estimation

16 April: C19, C20

12 Association mining

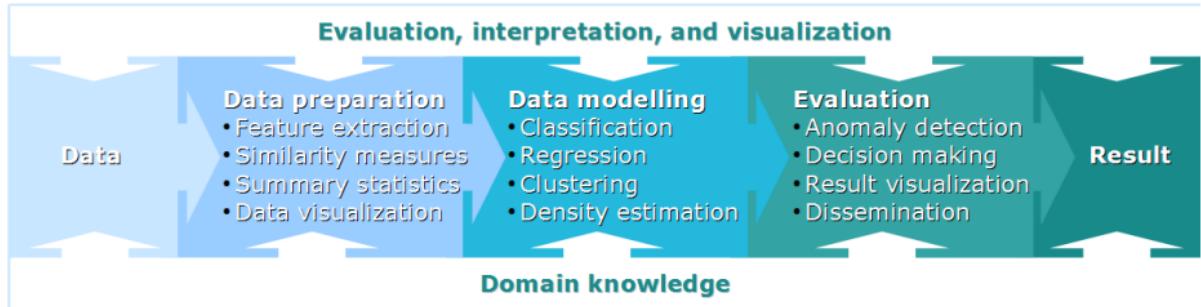
23 April: C21

Recap

13 Recap and discussion of the exam

30 April: C1-C21

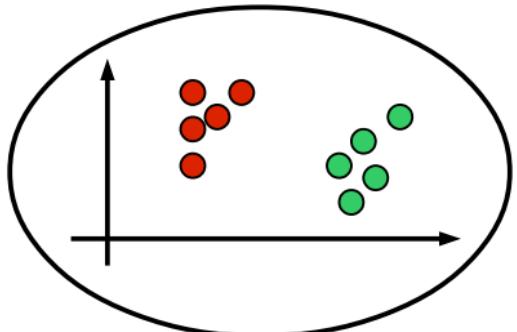
Online 24/7 help: Discussion Forum/Piazza
Streaming: Zoom (see link on DTU Learn)
Recordings: <https://panopto.dtu.dk/>



Learning Objectives

- Explain what supervised learning is
- Explain the difference between classification and regression
- Be able to evaluate classifiers in terms of the confusion matrix, error rate and accuracy
- Understand the principle behind decision trees and Hunt's algorithm
- Apply and interpret decision trees, linear regression and logistic regression

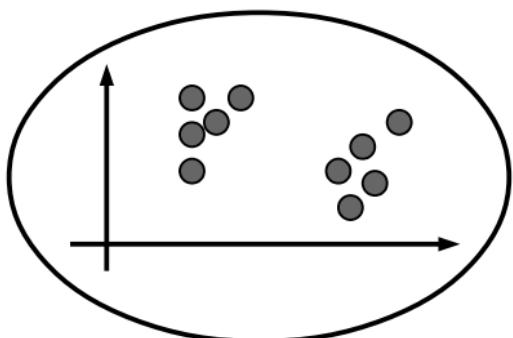
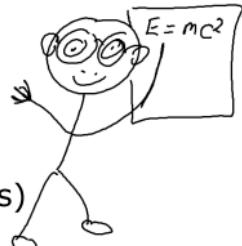
Supervised and Unsupervised learning



Supervised Learning

Input data x_n and output y_n

(Generalize from known examples)



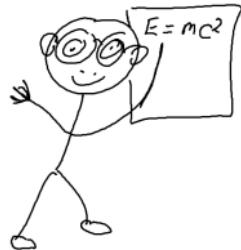
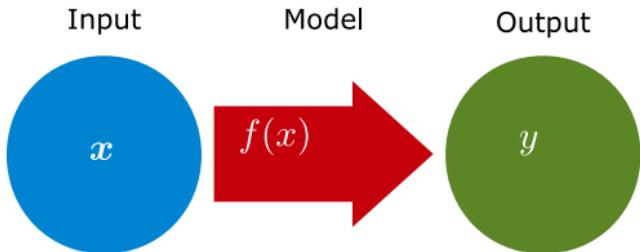
Unsupervised Learning

Input data x_n alone

(Exploratory analysis)



Supervised learning



- **Data**
 - Inputs and outputs (*this is what we are given*)
- **Model**
 - Function that maps inputs to outputs (*what we are trying to determine*)
$$\{x_n, y_n\}_{n=1}^N$$

$$f(\mathbf{x})$$
- **Cost function**
 - Dissimilarity measure between observation and prediction (*how we tell if a model is good or bad*)
$$d(y, f(\mathbf{x}))$$
- **Types of supervised learning**
 - Regression: Continuous output \mathbf{y}
 - Classification: Discrete output \mathbf{y}

Classification

- **Definition:** Learning a function that maps a data object to a discrete class
- **Why classify?**
 - Descriptive modeling
 - Explain / understand the relation between attributes and class
 - Predictive modeling
 - Predict the class of a new data object

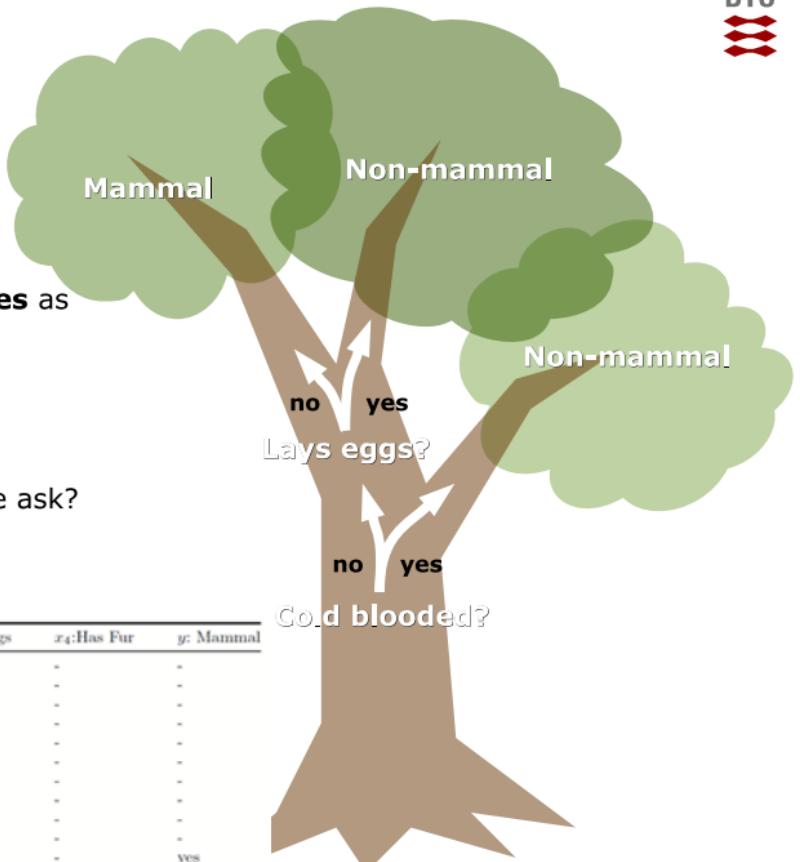
Decision trees

- Remember the game “20 questions to the professor”? (see also www.20q.net)

- Q1. Is it an Animal? Yes.
- Q2. Can you hold it? No.
- Q3. Does it live in groups (gregarious)? Yes.
- Q4. Are there many different sorts of it? No.
- Q5. Can it jump? Yes.
- Q6. Does it eat seeds? No.
- Q7. Is it white? Sometimes.
- Q8. Is it black and white? No.
- Q9. Does it have paws? Yes.
- Q10. Can you see it in a zoo? Yes.
- Q11. Does it roar? Yes.
- Q12. Is it worth a lot of money? Yes.
- Q13. Does it have spots? Yes.
- Q14. Is it multicoloured? Yes.
- Q15. Can you make money by selling it? Yes.
- Q16. Does it live in the jungle? Yes.
- Q17. I guessed that it was a leopard? Wrong.
- Q18. Does it like to play? Yes.
- Q19. I guessed that it was a cheetah? Wrong.
- Q20. I am guessing that it is a siberian tiger? Correct.

Decision trees

- Ask a series of questions until a conclusion is reached
- Example:** Classify **vertebrates** as
 - Mammal or
 - Non-mammal
- Learning task**
 - Which questions should we ask?



| Name | x_1 : Cold Blooded | x_2 : Has Legs | x_3 : Lay Eggs | x_4 : Has Fur | y : Mammal |
|------------|----------------------|------------------|------------------|-----------------|--------------|
| Snake | yes | - | yes | - | - |
| Starfish | yes | - | yes | - | - |
| Bluebird | - | yes | yes | - | - |
| Blackbird | - | yes | yes | - | - |
| Earthworm | yes | - | yes | - | - |
| Chameleon | yes | - | yes | - | - |
| Ant | yes | - | yes | - | - |
| Jellyfish | yes | - | yes | - | - |
| Snail | yes | - | yes | - | - |
| Sea Urchin | yes | - | yes | - | - |
| Dolphin | - | - | - | - | yes |
| Rat | - | yes | - | yes | yes |
| Dog | - | yes | - | yes | yes |
| Monkey | - | yes | - | yes | yes |
| Lion | - | yes | - | yes | yes |

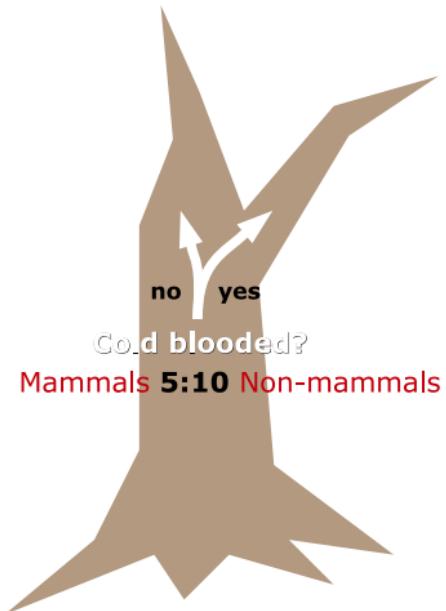
Hunts algorithm

- Assign all data objects to the root



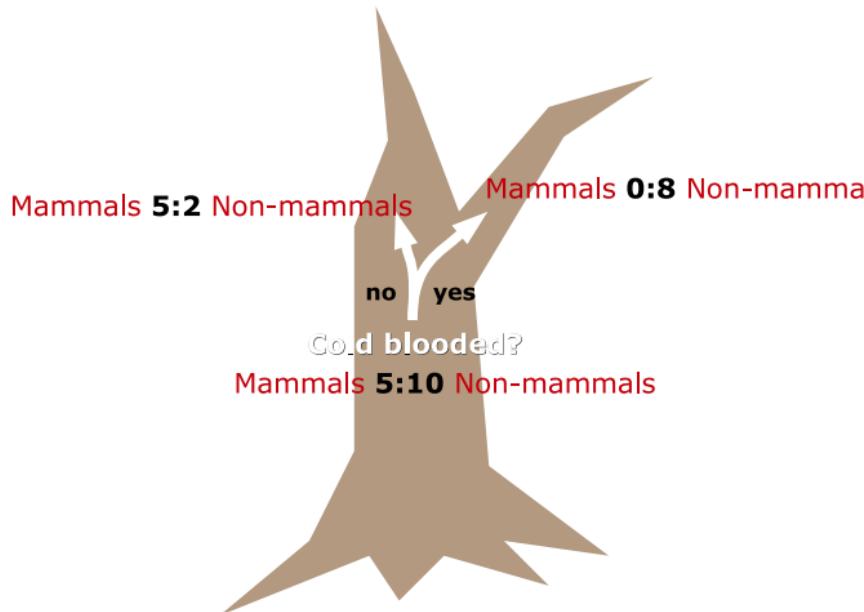
Hunts algorithm

- Select an attribute test condition
 - Find a good question to ask



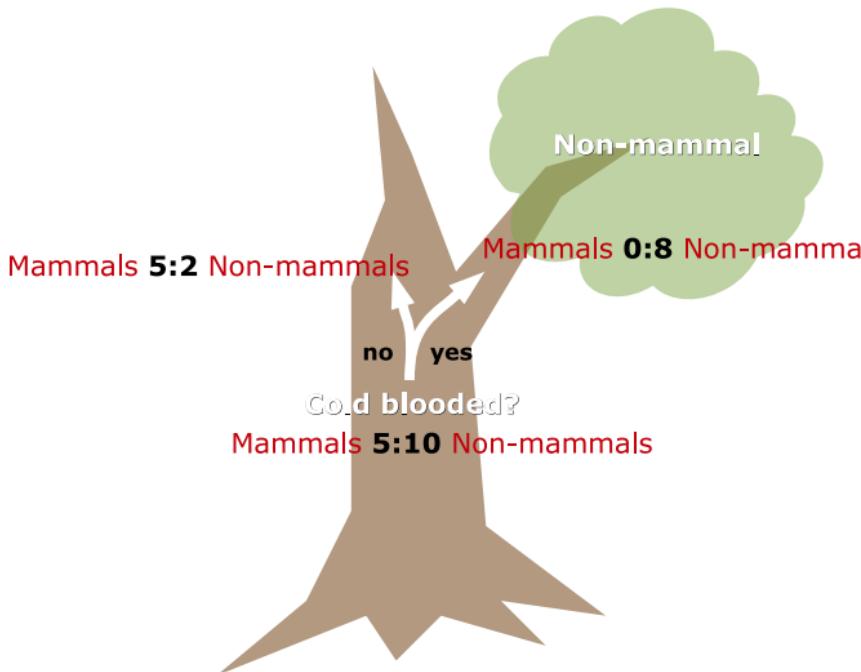
Hunt's Algorithm

- Partition the data objects into subsets according to the test condition



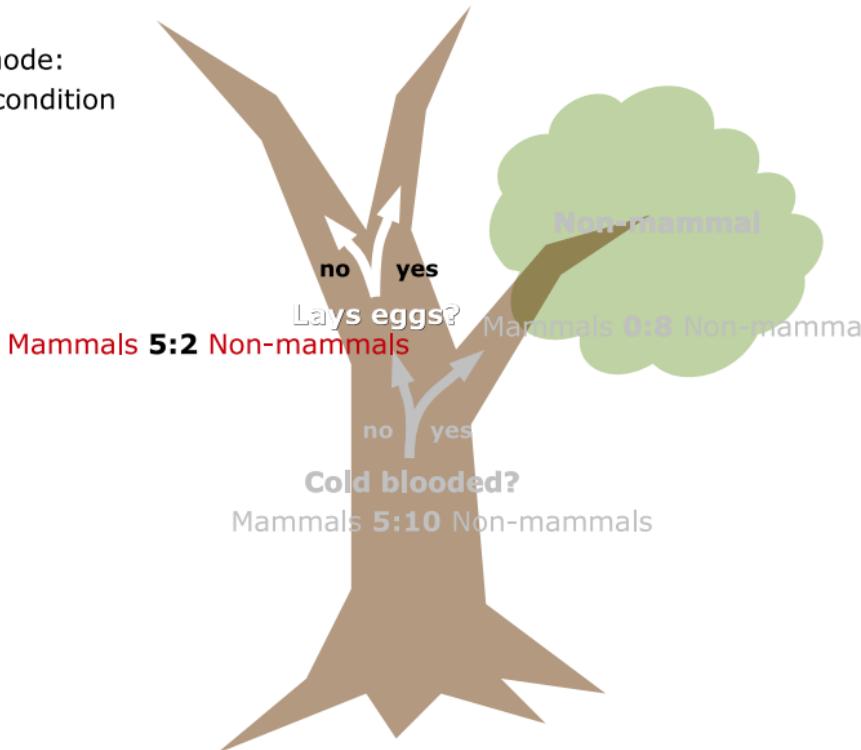
Hunts algorithm

- If all data objects belong to the same class
 - Create a leaf node



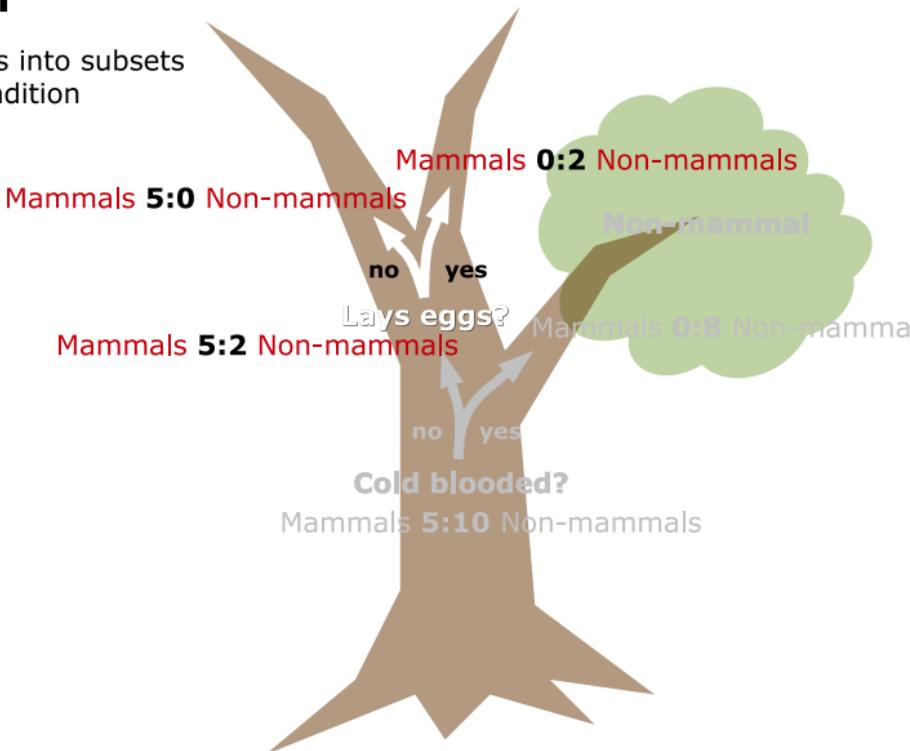
Hunts algorithm

- Repeat for each non-leave node:
 - Select an attribute test condition



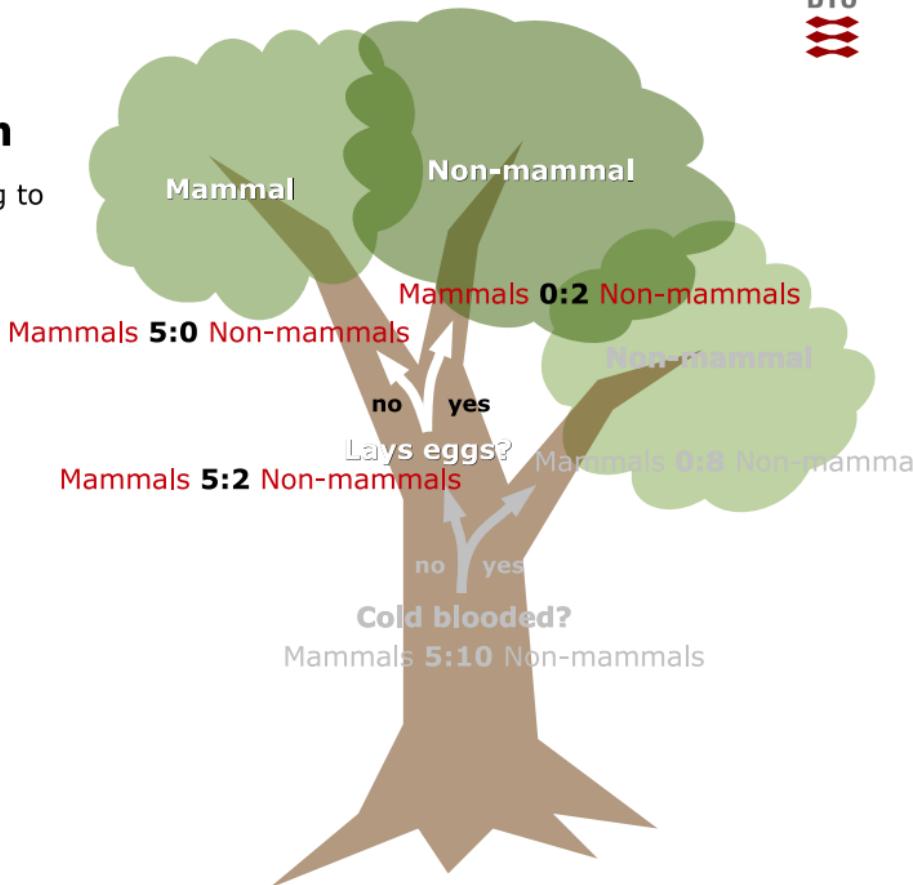
Hunts algorithm

- Partition the data objects into subsets according to the test condition



Hunts algorithm

- If all data objects belong to the same class
 - Create a leaf node



Hunts algorithm

- But how do we find the **best question** at each step?

Algorithm 2: Hunt's algorithm for decision trees

Require: Initial tree T only containing the root node

Require: D_r : Dataset associated with the current branch.

Initially just the full dataset

if The **stop criterion** is met **then**

Add a leaf node to the tree which assigns every observation to the most prevalent class in D_r

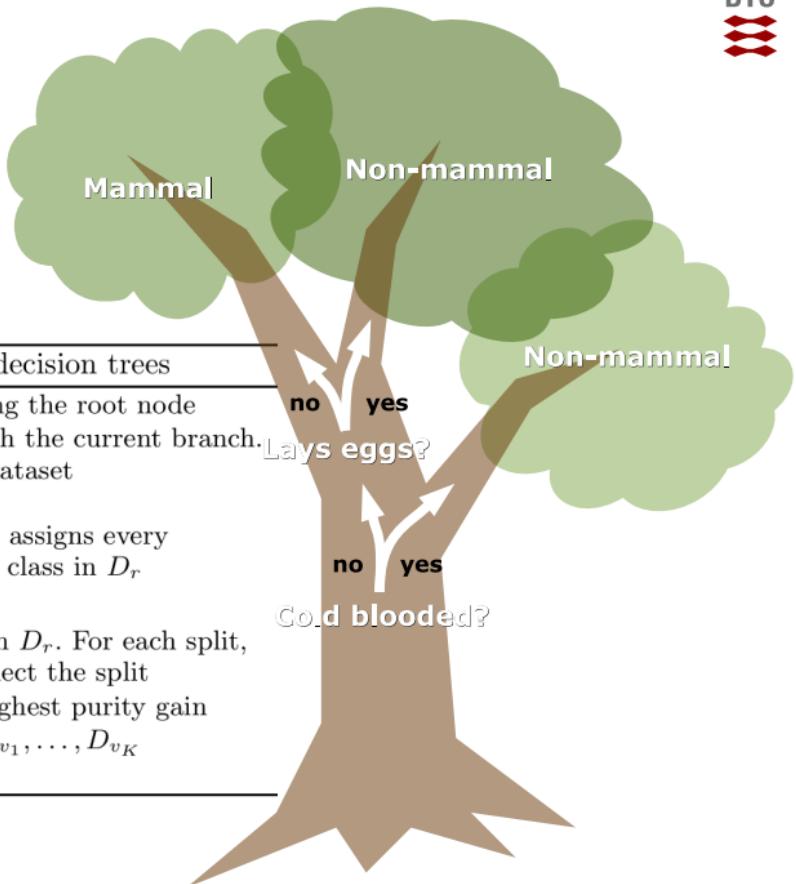
else

Try a number of different splits on D_r . For each split, compute the **purity gain** and select the split

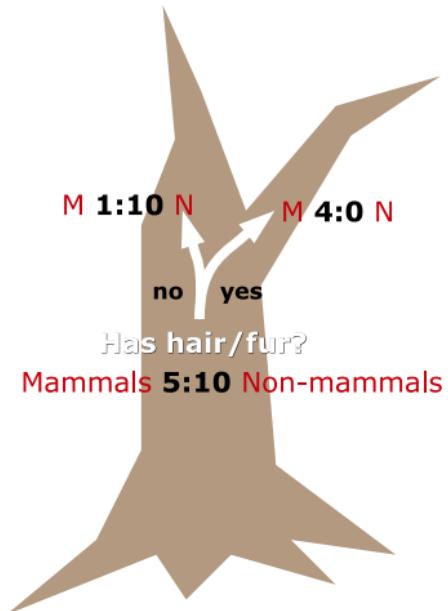
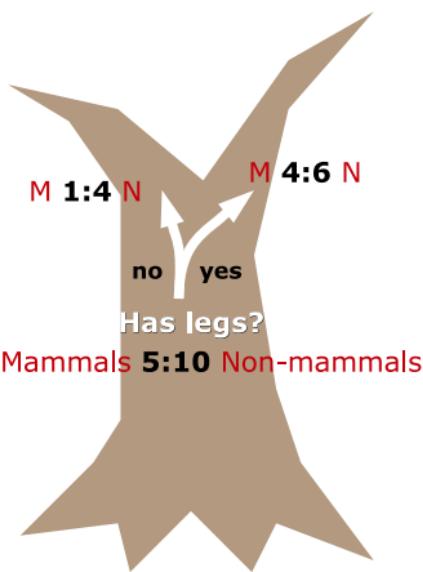
$D_r = \{D_{v_1}, \dots, D_{v_K}\}$ with the highest purity gain

Recursively call the method on D_{v_1}, \dots, D_{v_K}

end if



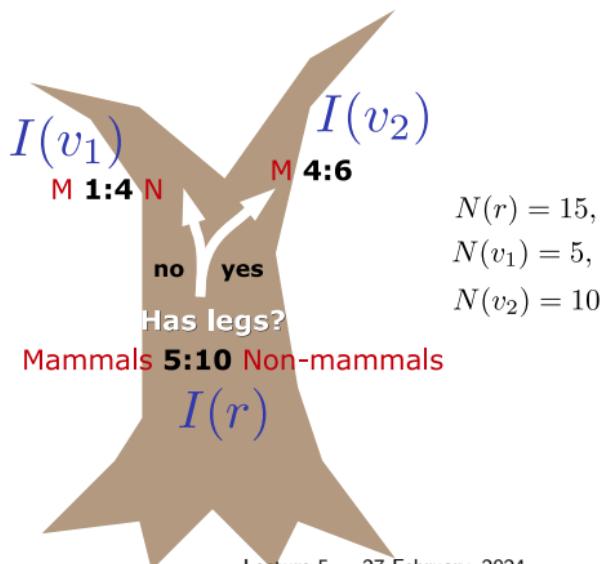
Which split is best?



Which split is best?

- Create a measure Δ (**the purity gain**) of how good a split is
- A binary split creates 3 partitions: the root r and the right/left branches v_1, v_2 .
- For each partition, we compute $I(r), I(v_1), I(v_2)$ (**the impurity**)
- Purity gain is the **weighted reduction in impurity**:

$$\Delta = I(r) - \sum_{k=1}^{K=2} \frac{N(v_k)}{N(r)} I(v_k)$$



Which split is best?

- Create a measure Δ (**the purity gain**) of how good a split is
- A binary split creates 3 partitions: the root r and the right/left branches v_1, v_2 .
- For each partition, we compute $I(r), I(v_1), I(v_2)$ (**the impurity**) of each partition
- Purity gain is the **weighted reduction in impurity**:

$$\Delta = I(r) - \sum_{k=1}^{K=2} \frac{N(v_k)}{N(r)} I(v_k)$$

$$P(M|v_1) = \frac{1}{5}$$

$$P(NM|v_1) = 1 - \frac{1}{5} = \frac{4}{5}$$

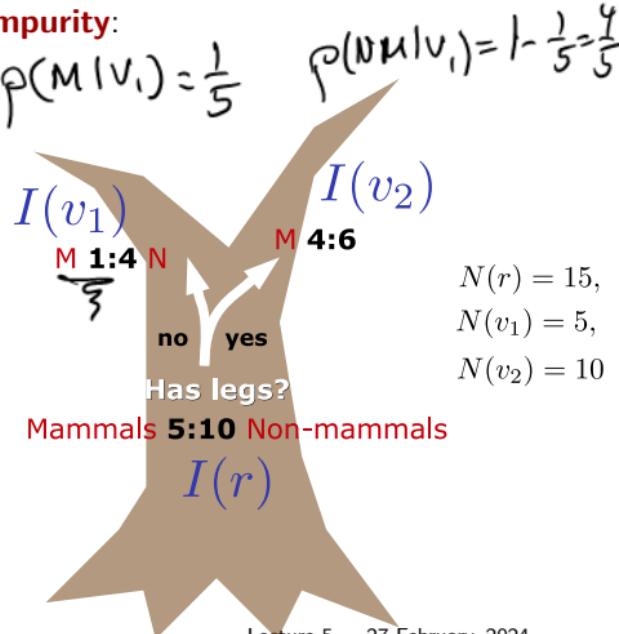
The impurity measure $I()$ can be one of the following

$$\text{Entropy}(v) = - \sum_{c=1}^C p(c|v) \log_2 p(c|v),$$

$$\text{Gini}(v) = 1 - \sum_{c=1}^C \underline{p(c|v)^2},$$

$$\text{ClassError}(v) = 1 - \max_c p(c|v).$$

$$p(c|v) = \frac{\{\text{Nr. in class } c \text{ in branch } v\}}{N(v)}$$



Quiz 1: Impurity gain

If we use the Gini index as impurity measure I , what is the purity gain Δ for the split indicated by the tree?

$$\Delta = I(r) - \sum_{k=1}^{K=2} \frac{N(v_k)}{N(r)} I(v_k)$$

The impurity measure $I()$ can be one of the following

$$\text{Entropy}(v) = - \sum_{c=1}^C p(c|v) \log_2 p(c|v),$$

$$\text{Gini}(v) = 1 - \sum_{c=1}^C \underline{\overbrace{p(c|v)^2}},$$

$$\text{ClassError}(v) = 1 - \max_c p(c|v).$$

$$p(c|v) = \frac{\text{Nr. in class } c \text{ in branch } v}{N(v)}$$

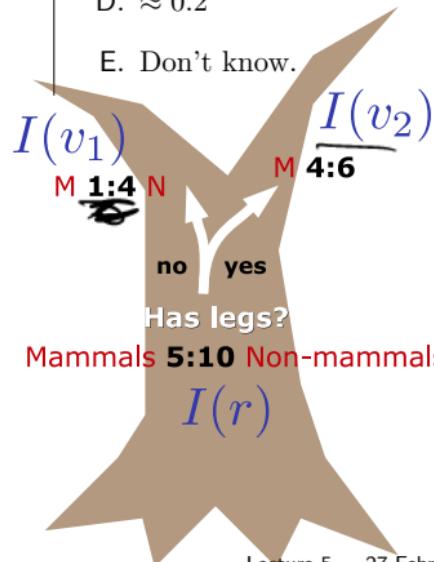
A. ≈ 0.0177

B. ≈ 0.104

C. ≈ 0.129

D. ≈ 0.2

E. Don't know.



$$P(M|r) = \frac{5}{15}$$

$$P(M|v_1) = \frac{1}{5}$$

$$P(M|v_2) = \frac{4}{10}$$

$$P(NM|r) = \frac{10}{15}$$

$$P(NM|v_1) = \frac{4}{5}$$

$$P(NM|v_2) = \frac{6}{10}$$

$$I(r) = 1 - \left(\frac{5}{15}\right)^2 - \left(\frac{10}{15}\right)^2$$

$$I(v_1) = \underbrace{\quad\quad\quad}_{\sim} \approx$$

$$I(v_2) = \underbrace{\quad\quad\quad}_{\sim} \approx$$

$$\Delta = I(r) - \frac{5}{15} I(v_1) - \frac{10}{15} I(v_2) =$$

Using Gini impurity we get:

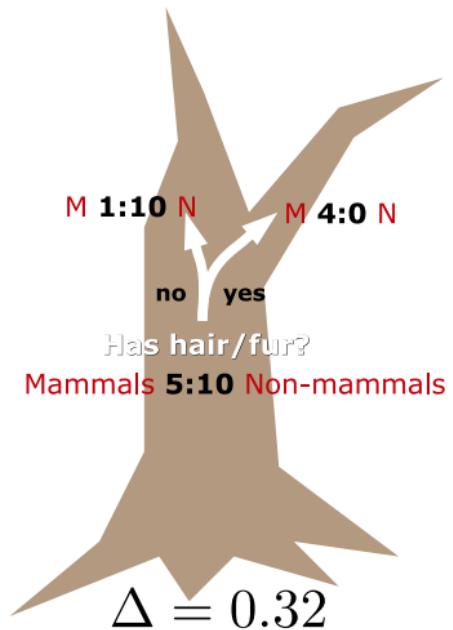
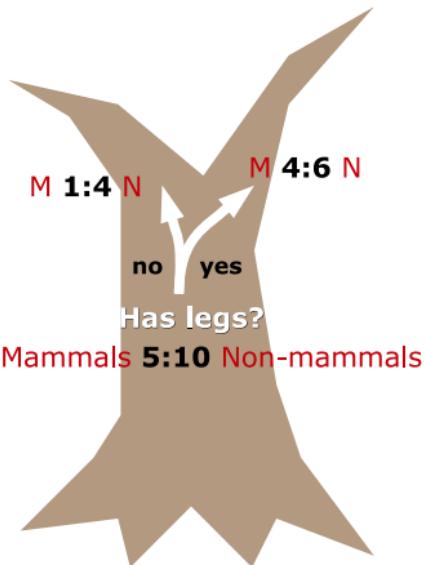
$$I(r) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2, I(v_1) = 1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2, I(v_2) = 1 - \left(\frac{4}{10}\right)^2 - \left(\frac{6}{10}\right)^2,$$

and finally

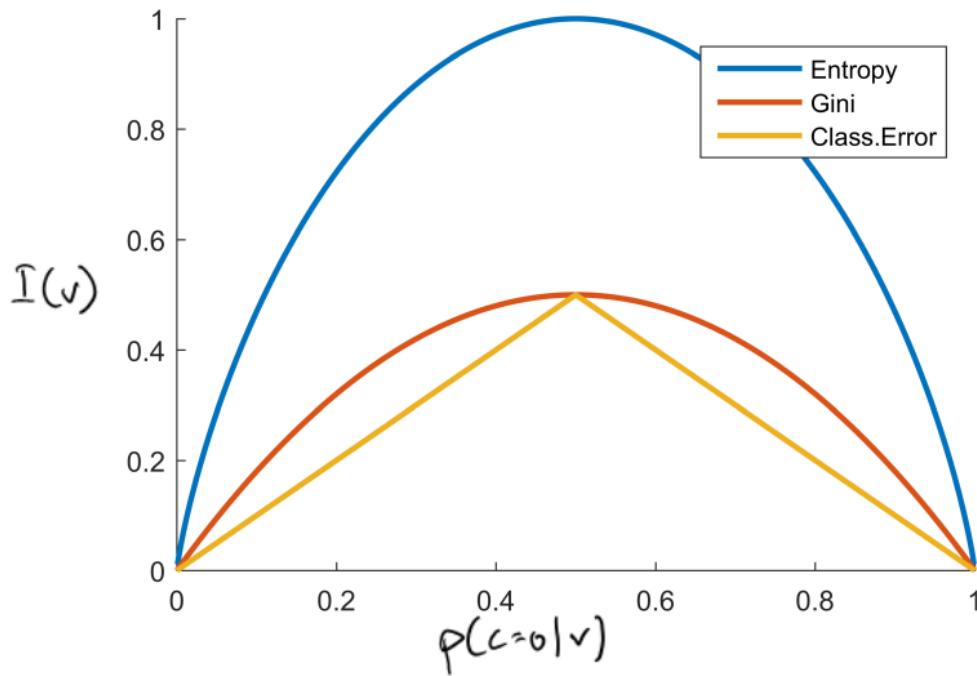
$$\Delta = I(r) - \frac{5}{15}I(v_1) - \frac{10}{15}I(v_2) \approx 0.0177$$

Selecting the best split

- Consider a large number of possible splits
- Compute a measure of impurity after the proposed split
 - For each new branch of the tree
 - Compute weighted average impurity
- Choose split that reduces impurity most



For a two class problem

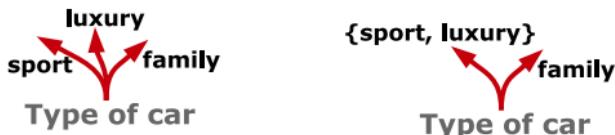


Which splits to consider

- Binary



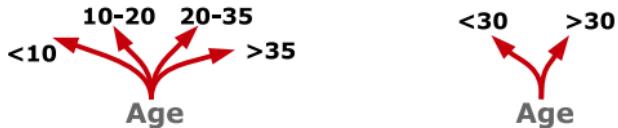
- Nominal



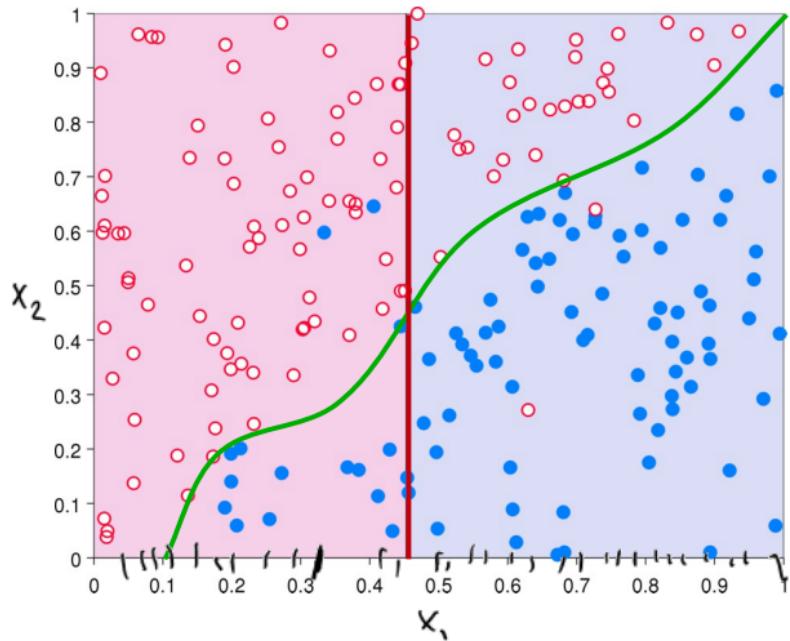
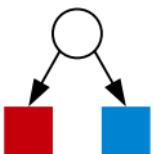
- Ordinal



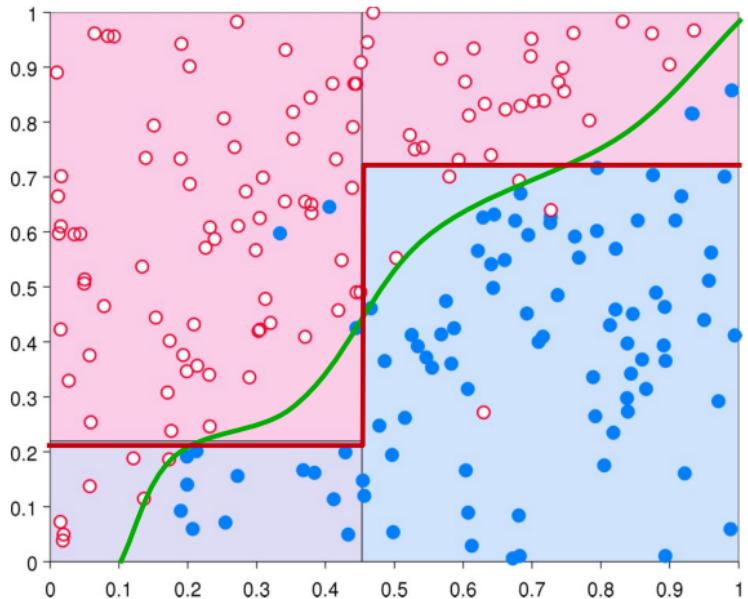
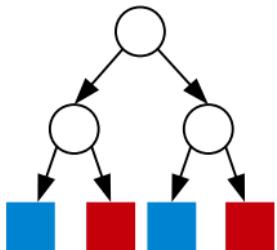
- Continuous



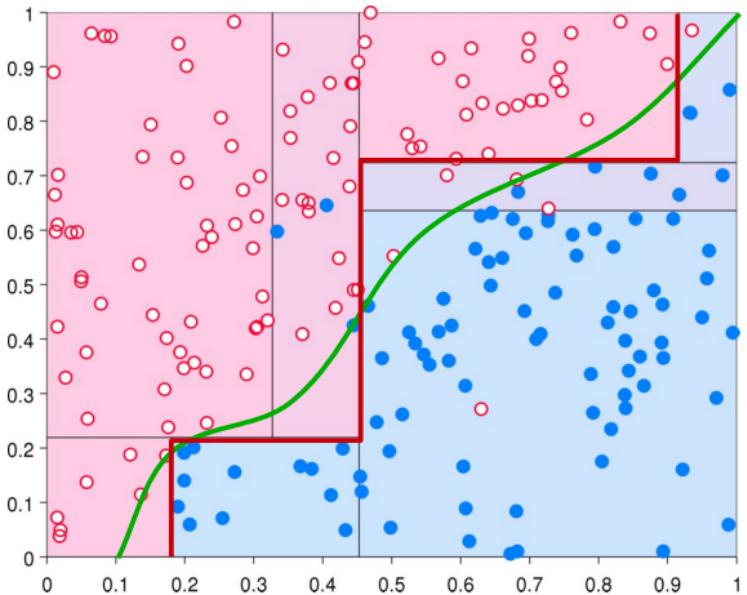
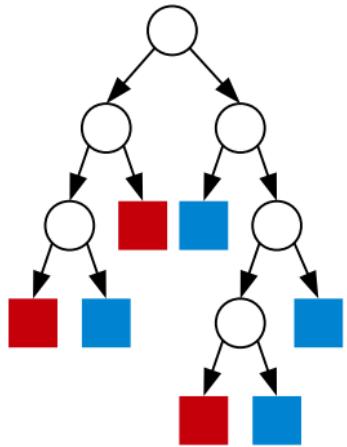
Classification Trees



Classification trees



Classification trees

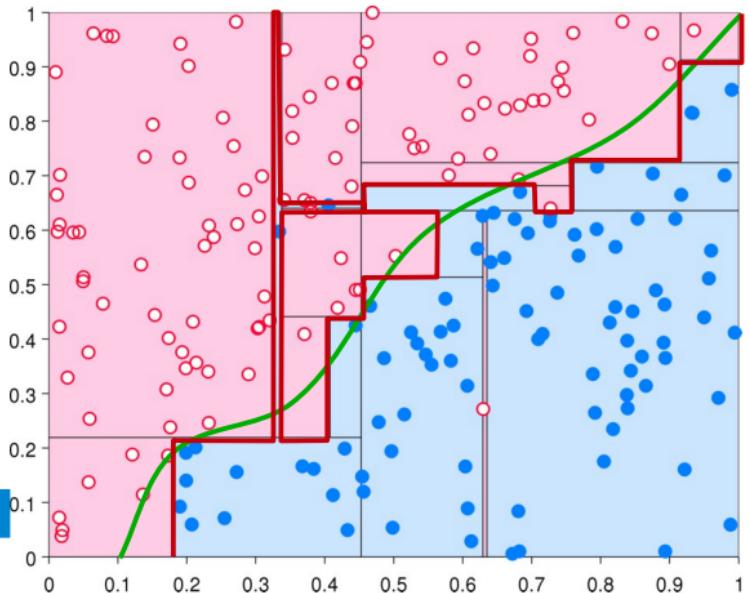
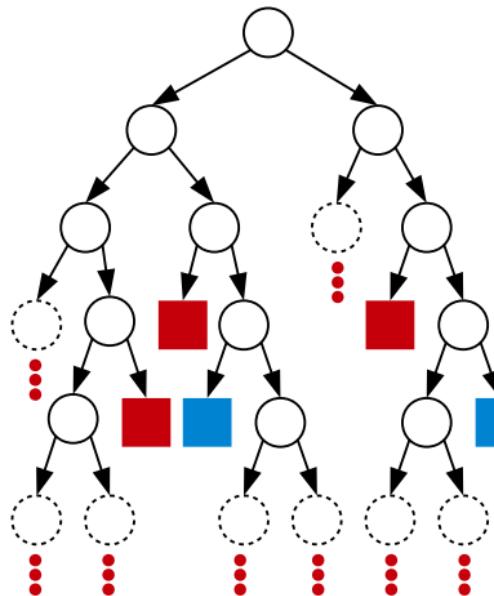


Classification trees

Common stopping criteria:

All records have the same class label

The number of observations have fallen below some minimum threshold



Evaluating a classifier

Confusion matrix

- Visualization of actual versus predicted class labels

- **Accuracy**

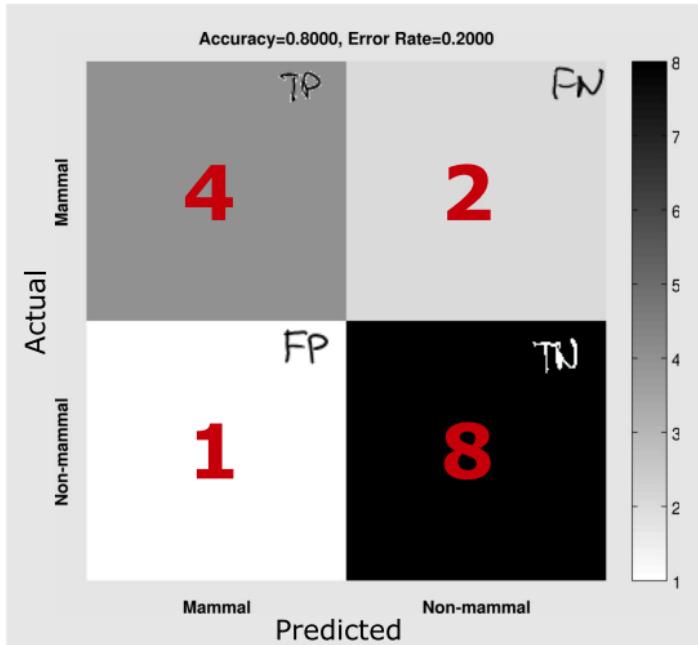
(Number of correctly predicted observations divided by the total number of observations)

$$\frac{4 + 8}{4 + 2 + 1 + 8} = 80\%$$

- **Error rate**

(Number of in-correctly predicted observations divided by the total number of observations)

$$\frac{2 + 1}{4 + 2 + 1 + 8} = 20\%$$



Example: Iris data

The iris data set

- **Three flowers**

- 50 instances of each class, 150 in total

- **Attributes**

- Sepal (outermost leaves)

- length in cm
- width in cm

- Petal (innermost leaves)

- length in cm
- width in cm

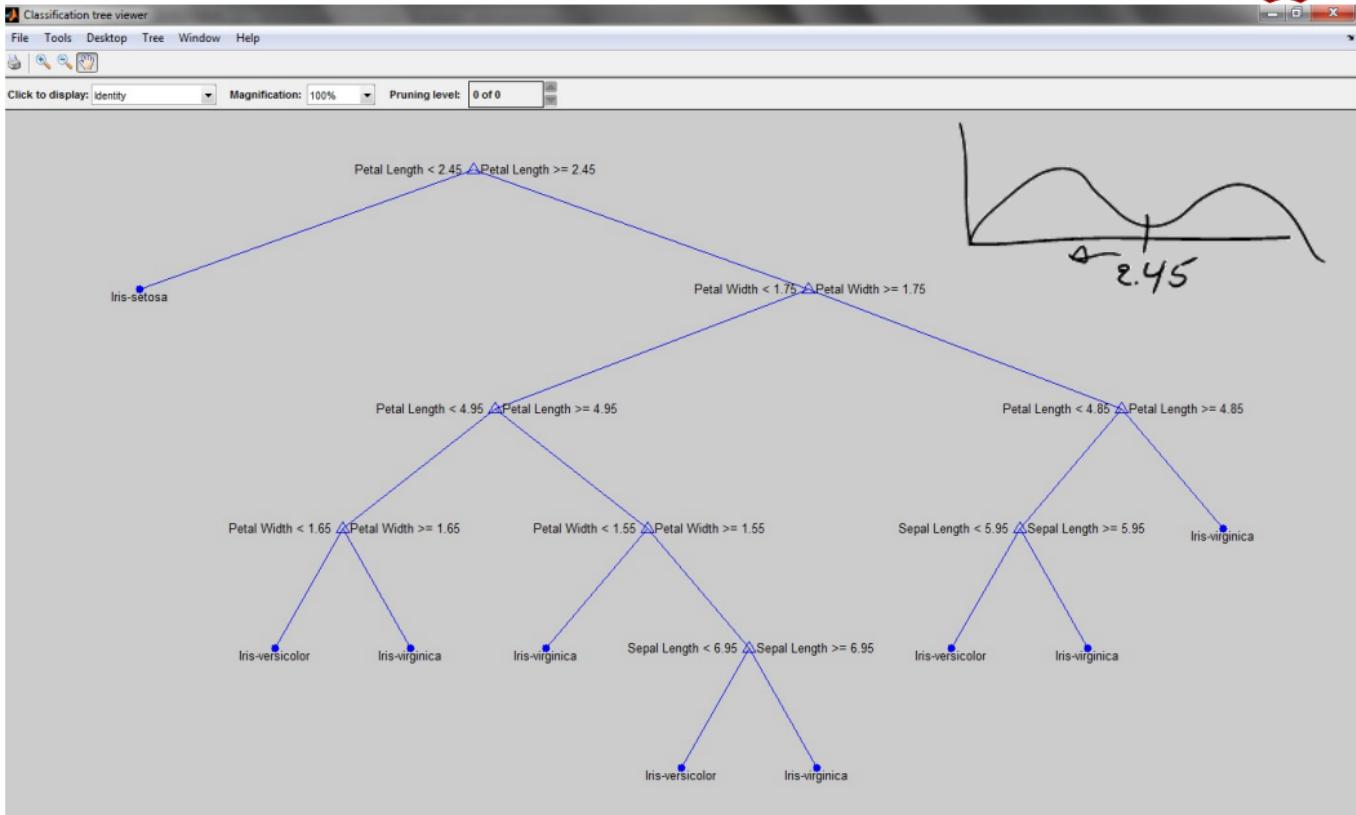
- Class of flower

- Iris Setosa
- Iris Versicolour
- Iris Virginica



| Flower ID | Attribute | | | |
|-----------|--------------|-------------|--------------|-------------|
| | Sepal Length | Sepal Width | Petal Length | Petal Width |
| 1 | 5.1 | 3.5 | 1.4 | 0.2 |
| 2 | 4.9 | 3.0 | 1.4 | 0.2 |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 |
| 4 | 4.6 | 3.1 | 1.5 | 0.2 |
| . | . | . | . | . |
| . | . | . | . | . |
| 150 | 5.9 | 3.0 | 5.1 | 1.8 |

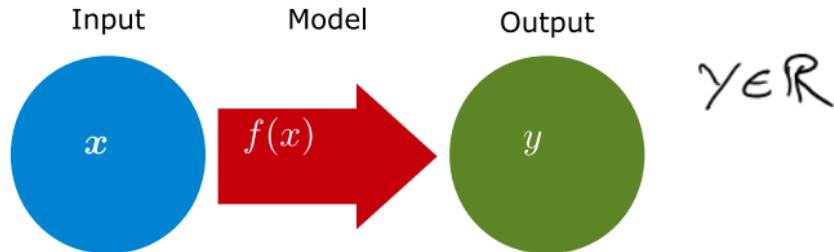
X ^{Observation x Attribute}



What would the following iris flower be classified as?

| Sepal Length | Sepal Width | Petal Length | Petal Width |
|--------------|-------------|--------------|-------------|
| 4.0 | 3.5 | 3.0 | 2.0 |

Supervised learning



- **Mapping between domains**

- Classification: Discrete (nominal) output
- Regression: Continuous output

Supervised learning

- **Data**

- Inputs and outputs

$$\{\boldsymbol{x}_n, y_n\}_{n=1}^N$$

- **Model**

- Function that maps inputs to outputs

$$f(\boldsymbol{x})$$

- **Cost function**

- Dissimilarity measure between data and model

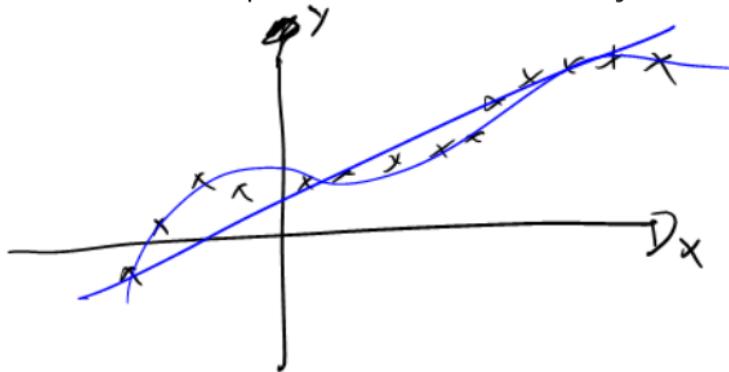
$$d(y, f(\boldsymbol{x}))$$

Regression

- **Definition:** Learning a function that maps a data object to a continuous-valued output

- **Why Regression?**

- Descriptive modeling
 - Explain / understand the relation between attributes and continuous-valued output
- Predictive modeling
 - Predict the output value of a new data object



Regression trees

Algorithm 4: Hunt's algorithm for regression trees

Require: Initial tree T only containing the root node

Require: D_r : Dataset associated with the current branch. Initially just the full dataset

if The **stop criterion** is met then

Add a leaf node to the tree which assigns every observation the mean value of the nodes in D_r :

$$y(r) = \frac{1}{N(r)} \sum_{i \in r} y_i$$

else

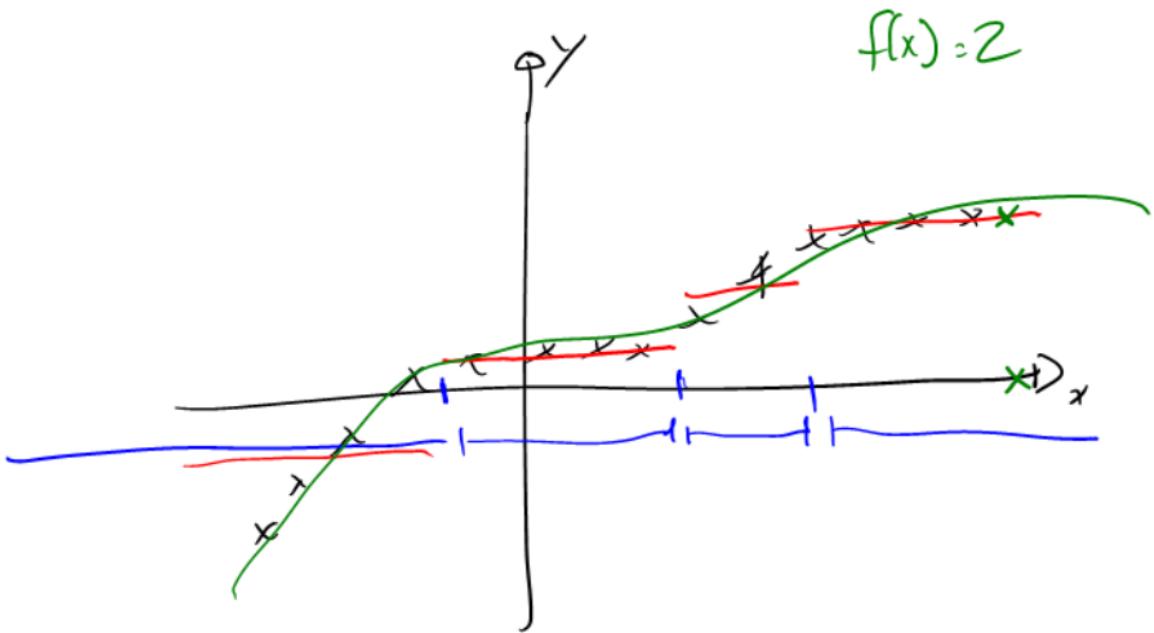
Try a number of different splits on D_r . For each split, compute the **purity gain** using the sum-of-squares impurity measure and select the split $D_r = \{D_{v_1}, \dots, D_{v_K}\}$ with the highest purity gain

Recursively call the method on D_{v_1}, \dots, D_{v_K}

end if

Use mean square error as purity gain

$$I(v) = \frac{1}{N(v)} \sum_{i \in v} (y_i - \hat{y}_v)^2, \quad \hat{y}_v = \frac{1}{N(v)} \sum_{i \in v} y_i$$



Evaluating a regression model

Compute average loss per observation:

$$E = \frac{1}{N} \sum_{i=1}^N L(y_i, f(\mathbf{x}_i))$$

Where we either use L_1 or L_2 (Euclidean) loss

$$L_1(y_i, \hat{y}_i) = |y_i - \hat{y}_i|, \quad L_2(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$$

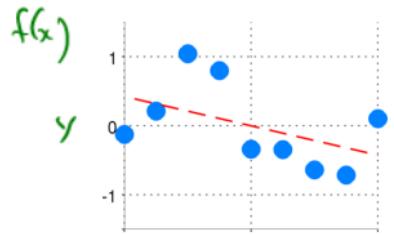
(Compare these to p -norms)

Linear regression

- 1-dimensional inputs

$$f(x) = w_0 + w_1 x$$

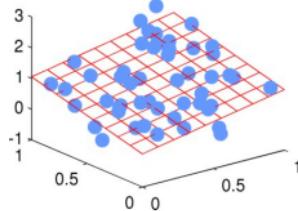
y



- 2-dimensional inputs

$$f(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2$$

x



- K-dimensional inputs

$$f(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_K x_K$$

Linear regression

- K-dimensional inputs

$$f(x) = w_0 + w_1x_1 + w_2x_2 + \cdots + w_Kx_K$$

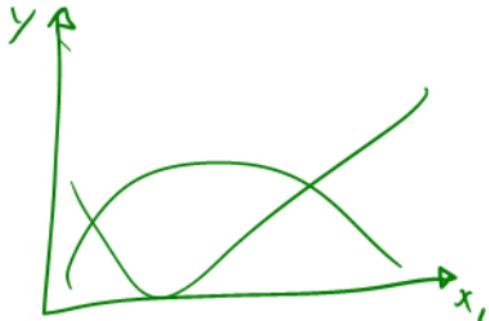
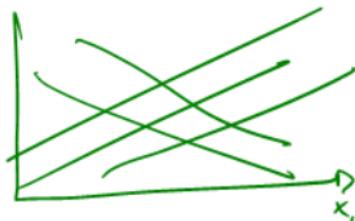
- Non-linearly transformed inputs

$$f(x) = w_0 + w_1x + w_2x^2 + \cdots + w_Kx^K$$

$$f(x) = w_0 + w_1\sin(x) + w_2\cos(x)$$

$$\mathbf{x} = \mathbf{x}_1$$

$$\mathbf{X} = [1 \ x \ \sin(x) \ \cos(x) \ x^2 \ x^3]$$



Linear regression

- K-dimensional inputs

$$f(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2 + \cdots + w_Kx_K$$

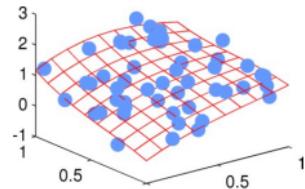
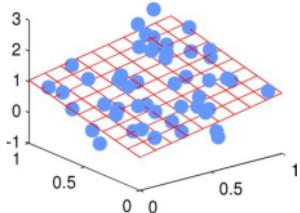
- Non-linearly transformed inputs

$$f(x) = w_0 + w_1x + w_2x^2 + \cdots + w_Kx^K$$

$$f(x) = w_0 + w_1\sin(x) + w_2\cos(x)$$

- Example

$$f(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2$$



$$\begin{aligned} f(\mathbf{x}) = & w_0 + w_1x_1 + w_2x_2 \\ & + w_3x_1^2 + w_4x_2^2 + w_5x_1x_2 \\ & + w_6x_1^3 + w_7x_1^2x_2 + w_8x_1x_2^2 + w_9x_2^3 \end{aligned}$$

Vector notation

- The linear model can be written compactly using vector notation

$$f(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_K x_K$$

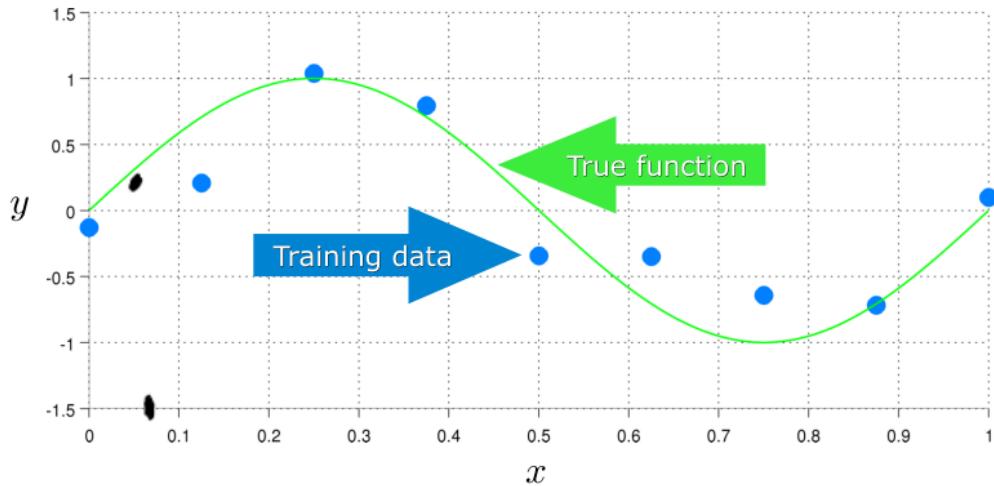
$$= \sum_{k=0}^K w_k x_k = \boxed{\mathbf{\tilde{x}}^\top \mathbf{w}}$$

- where $x_0 = 1$

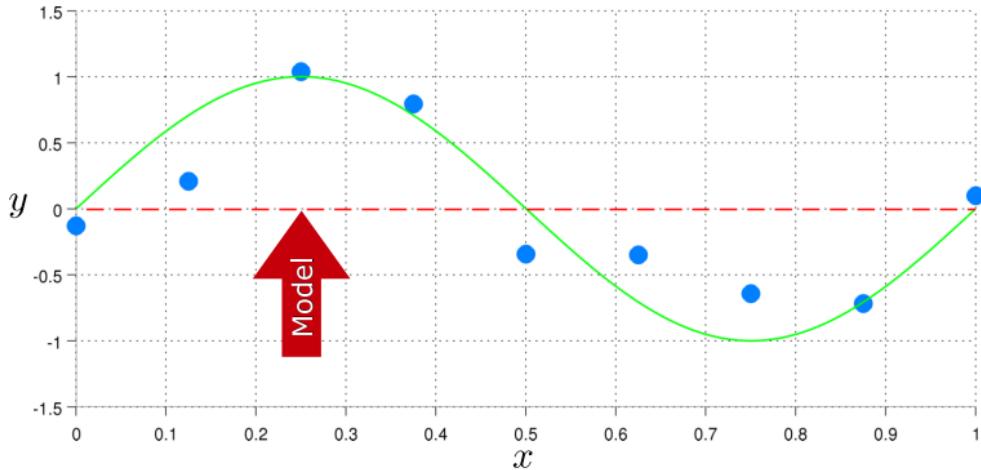
$$\mathbf{\tilde{x}} = [1 \ x_1 \ x_2 \ \dots \ x_n]$$

$$\mathbf{\tilde{x}} = [1 \ x_1 \ x_2 \ \dots \ x_n \ x_1^2 \ x_2^2 \ \sin(x_{100})]$$

Linear regression



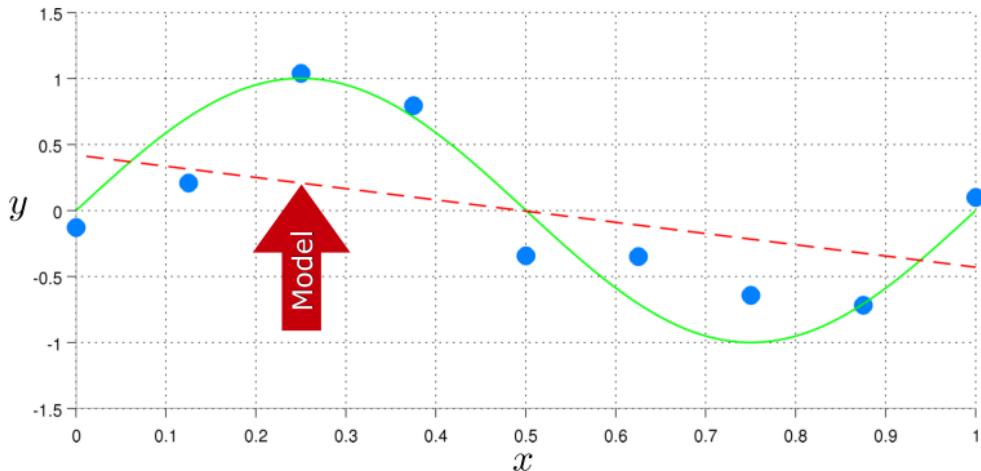
Linear regression



Model

$$f(x) = w_0$$

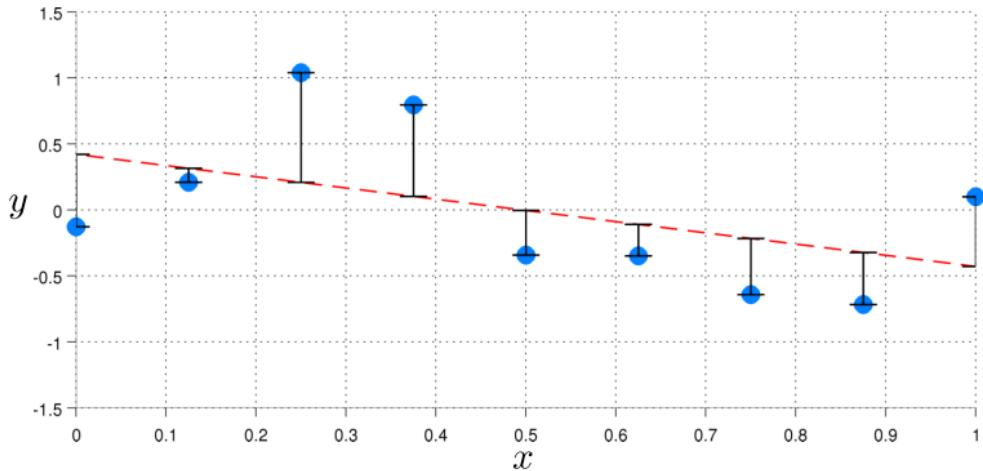
Linear regression



Model

$$f(x) = w_0 + w_1x$$

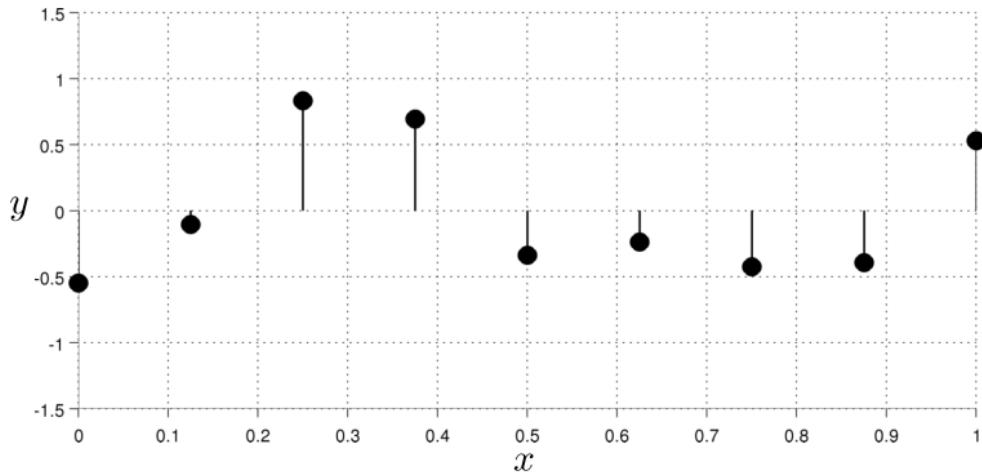
Residual error



Model

$$f(x) = w_0 + w_1x$$

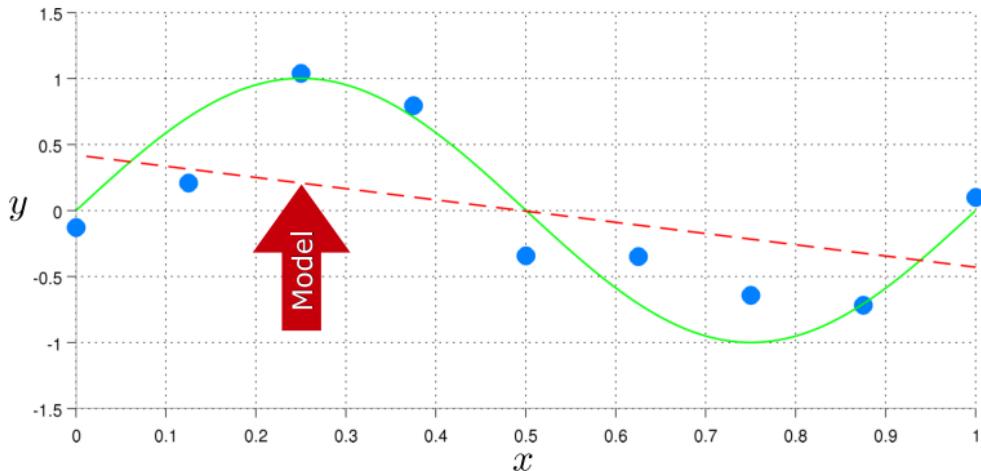
Residual error



Model

$$f(x) = w_0 + w_1 x$$

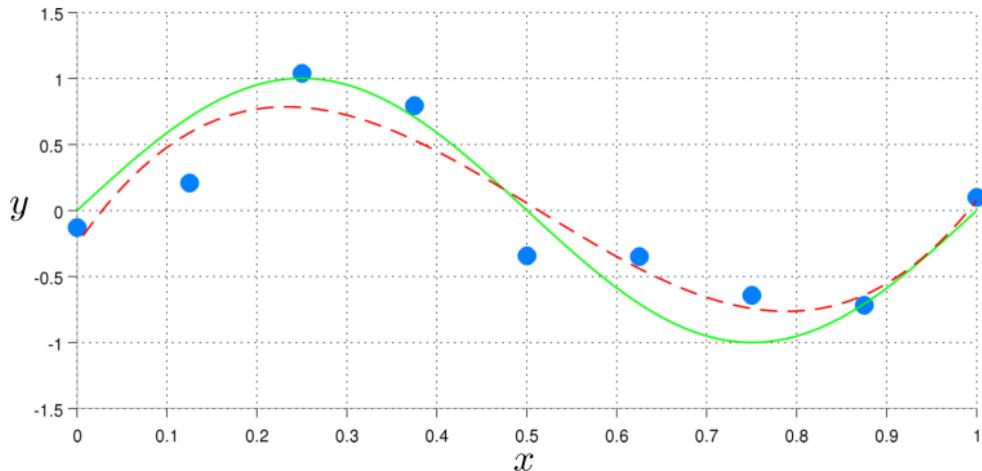
Linear regression



Model

$$f(x) = w_0 + w_1x$$

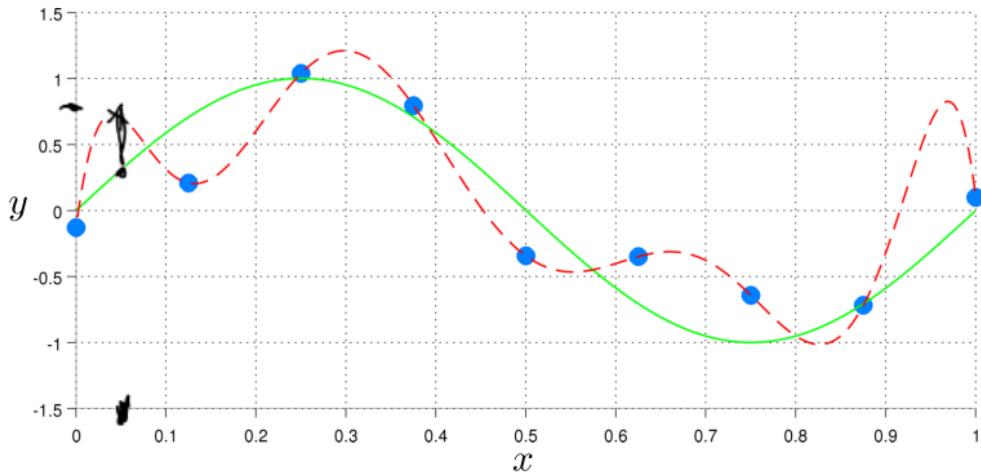
Linear regression



Model

$$f(x) = w_0 + w_1x + w_2x^2 + w_3x^3$$

Linear regression



Model

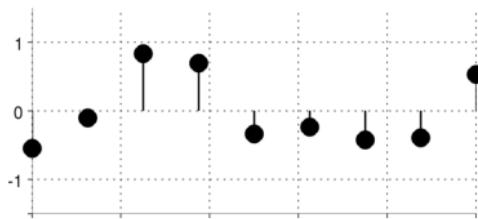
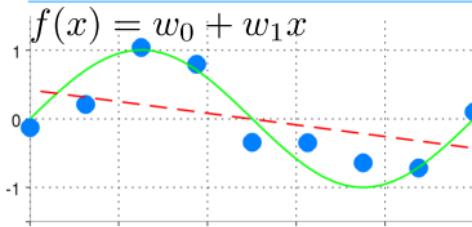
$$f(x) = w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4 + w_5x^5 + w_6x^6 + w_7x^7 + w_8x^8$$



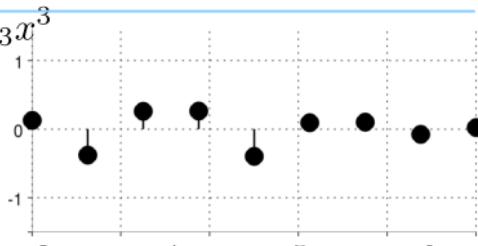
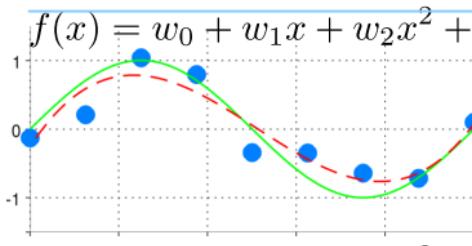
Model order

- Which model order
 - Gives the best fit
 - Do you think is most "correct"?

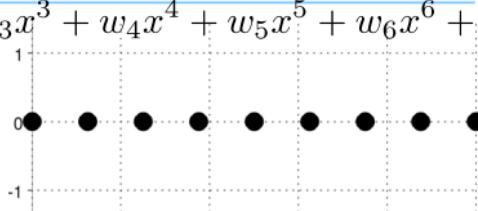
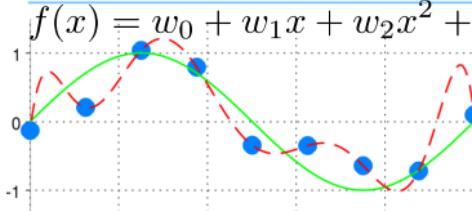
$$f(x) = w_0 + w_1x$$



$$f(x) = w_0 + w_1x + w_2x^2 + w_3x^3$$



$$f(x) = w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4 + w_5x^5 + w_6x^6 + w_7x^7 + w_8x^8$$

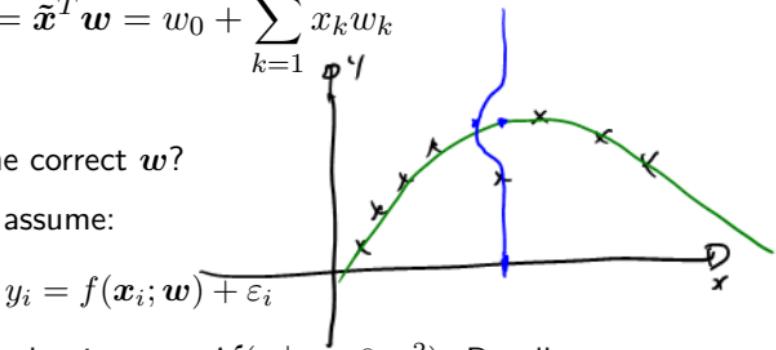


$$\tilde{\mathbf{X}} = \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix}$$

Setup: We got some data (\mathbf{y}, \mathbf{X}) . We have a way to define a function

$$f(\mathbf{x}; \mathbf{w}) = \tilde{\mathbf{x}}^T \mathbf{w} = w_0 + \sum_{k=1}^M x_k w_k$$

- **Question:** How do we learn the correct \mathbf{w} ?
- Answer: For each observation, assume:



where ε_i is a normally distributed noise term $\mathcal{N}(\varepsilon_i | \mu = 0, \sigma^2)$. Recall:

$$\mathcal{N}(\varepsilon_i | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\varepsilon_i - \mu)^2}{2\sigma^2}}$$

- This means that (since $\varepsilon_i = y_i - f(\tilde{\mathbf{x}}_i, \mathbf{w})$)

$$p(y_i | \mathbf{x}_i, \mathbf{w}) = p(\varepsilon_i | \tilde{\mathbf{x}}_i, \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{((y_i - f(\tilde{\mathbf{x}}_i, \mathbf{w})) - 0)^2}{2\sigma^2}} = \mathcal{N}(y_i | \mu = f(\tilde{\mathbf{x}}_i, \mathbf{w}), \sigma^2)$$

Recall from last time: Maximum A Posteriori (MAP) learning

- Consider some data $\mathbf{X} = \mathbf{x}_1, \dots, \mathbf{x}_N$ and $\mathbf{y} = y_1, \dots, y_N$
- Suppose \mathbf{x}_i relates to y_i by some parameters \mathbf{w}
- **Assume**

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^N p(y_i|\mathbf{x}_i, \mathbf{w}), \quad p(\mathbf{w}|\mathbf{X}) = p(\mathbf{w})$$

- Then

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\mathbf{X})}{p(\mathbf{y}|\mathbf{X})} = \frac{\prod_{i=1}^N p(y_i|\mathbf{w}, \mathbf{x}_i)p(\mathbf{w})}{p(\mathbf{X}|\mathbf{y})}$$

- And maximizing: $\mathbf{w}^* = \arg \max_{\mathbf{w}} p(\mathbf{w}|\mathbf{X}, \mathbf{y})$ is equivalent to

$$\text{Minimize: } \mathbf{w}^* = \arg \min_{\mathbf{w}} E(\mathbf{w})$$

$$E(\mathbf{w}) = \frac{1}{N} \left[- \sum_{i=1}^N \log p(y_i|\mathbf{x}_i, \mathbf{w}) - \log p(\mathbf{w}) \right]$$

Back to the linear model

$$p(y_i | \mathbf{x}_i, \mathbf{w}) = \mathcal{N}(y_i | f(\mathbf{x}_i, \mathbf{w}), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - f(\mathbf{x}_i; \mathbf{w}))^2}{2\sigma^2}}$$

Optimal $\mathbf{w}^* = \arg \max_{\mathbf{w}} p(\mathbf{w} | \mathbf{X}, \mathbf{y})$ found as $\mathbf{w}^* = \arg \min_{\mathbf{w}} E(\mathbf{w})$

$$E(\mathbf{w}) = \frac{1}{N} \left[- \sum_{i=1}^N \log p(y_i | \mathbf{x}_i, \mathbf{w}) - \log p(\mathbf{w}) \right]$$

By assuming a constant/flat prior we obtain (Maximum Likelihood learning)

$$E(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N \frac{(y_i - \tilde{\mathbf{x}}_i^\top \mathbf{w})^2}{2\sigma^2} = \frac{1}{N} \frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - \tilde{\mathbf{x}}_i^\top \mathbf{w})^2 \propto \|\mathbf{y} - \tilde{\mathbf{X}}\mathbf{w}\|^2$$

$$\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = 2\tilde{\mathbf{X}}^\top (\mathbf{y} - \tilde{\mathbf{X}}\mathbf{w}) = 0$$

$$\Rightarrow \mathbf{w}^* = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \mathbf{y}$$

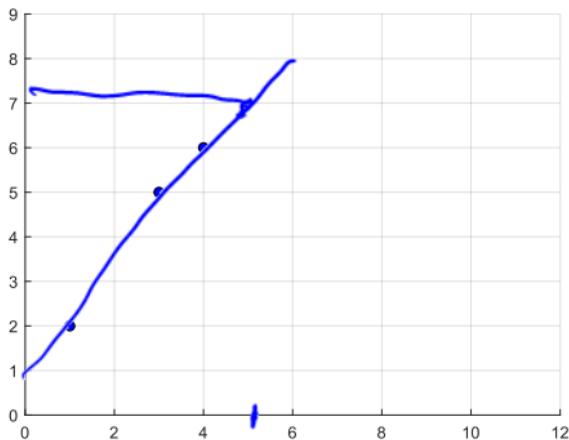
Quiz 2: The linear model

Suppose you observe three points:

$$(x, y) = \{(1, 2), (3, 5), (4, 6)\}$$

Knowing what you have learned so far, you first bring these points to the standard format:

$$\tilde{\mathbf{X}} = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 2 \\ 5 \\ 6 \end{bmatrix}$$



You wish to train a linear model of the form $y = ax + b$ on this dataset. What is $\mathbf{w} = \begin{bmatrix} b \\ a \end{bmatrix}$? Then, compute the prediction of the model at $x = 5$? (the prediction is given as: $y = \tilde{\mathbf{x}}^\top \mathbf{w}^*$)

- A. 6.5
- B. 7
- C. 7.5
- D. 8
- E. Don't know.

$$\tilde{\mathbf{X}} = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 2 \\ 5 \\ 6 \end{bmatrix}$$

$$f(x=5) = \mathbf{w}^\top \begin{bmatrix} 1 \\ 5 \end{bmatrix} = 7.5$$

Recall $\mathbf{w}^* = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \mathbf{y}$

The solution is found by first computing \mathbf{w} using the standard formula, and remembering to add a column of ones to \mathbf{X} to account for the offset. We

get:

$$\mathbf{w} \approx \begin{bmatrix} 0.7143 \\ 1.3571 \end{bmatrix}$$

Evaluating the model gives $f(5) = y = 7.5$.

Logistic regression

$$f(\mathbf{x}) = \mathbf{w}^\top \tilde{\mathbf{x}}$$

- Assume we are given (\mathbf{X}, \mathbf{y}) , but assume y is *binary*: $y_i = 0, 1$
- An idea is to use the Bernoulli distribution

$$p(y_i | \mathbf{x}_i, \mathbf{w}) = \text{Bernoulli}(y_i | \theta_i) = \theta_i^{y_i} (1 - \theta_i)^{1-y_i}$$

$\underbrace{\hspace{1cm}}$

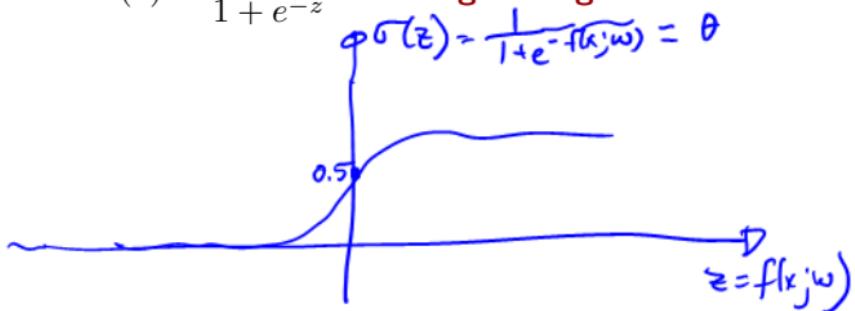
Where θ_i depends on \mathbf{w} and \mathbf{x}_i .

- Problem:** θ_i must belong to the unit interval, but $f(\mathbf{x}_i, \mathbf{w}) = \tilde{\mathbf{x}}_i^\top \mathbf{w}$ won't
- Solution:** Assume

$$\underline{\theta_i} = \sigma(f(\mathbf{x}, \mathbf{w})), \quad \text{where } \sigma(z) = \frac{1}{1 + e^{-z}} \text{ is the logistic sigmoid}$$

Then

$$-\log p(y_i | \mathbf{x}_i, \mathbf{w}) =$$



Logistic regression

- Assume we are given (\mathbf{X}, \mathbf{y}) , but assume y is *binary*: $y_i = 0, 1$
- An idea is to use the Bernoulli distribution

$$p(y_i | \mathbf{x}_i, \mathbf{w}) = \text{Bernoulli}(y_i | \theta_i) = \theta_i^{y_i} (1 - \theta_i)^{1-y_i}$$

Where θ_i depends on \mathbf{w} and \mathbf{x}_i .

- **Problem:** θ_i must belong to the unit interval, but $f(\mathbf{x}_i; \mathbf{w}) = \tilde{\mathbf{x}}_i^\top \mathbf{w}$ won't
- **Solution:** Assume

$$\theta_i = \sigma(f(\mathbf{x}, \mathbf{w})), \quad \text{where } \sigma(z) = \frac{1}{1 + e^{-z}} \text{ is the logistic sigmoid}$$

Then

$$\begin{aligned} -\log p(y_i | \mathbf{x}_i, \mathbf{w}) &= -\log [\text{Bern}(y_i | \theta_i = \sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w}))] = -\log [\theta_i^{y_i} (1 - \theta_i)^{1-y_i}] \\ &= -y_i \log(\theta_i) - (1 - y_i) \log(1 - \theta_i) \end{aligned}$$

Recall: Maximum A Posteriori (MAP) learning

- Consider some data $\mathbf{X} = \mathbf{x}_1, \dots, \mathbf{x}_N$ and $\mathbf{y} = y_1, \dots, y_N$

- **Assume**

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^N p(y_i|\mathbf{x}_i, \mathbf{w}), \quad p(\mathbf{w}|\mathbf{X}) = p(\mathbf{w})$$

- Then

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{\prod_{i=1}^N p(y_i|\mathbf{w}, \mathbf{x}_i)p(\mathbf{w})}{p(\mathbf{X}|\mathbf{y})}$$

- Maximizing: $\mathbf{w}^* = \arg \max_{\mathbf{w}} p(\mathbf{w}|\mathbf{X}, \mathbf{y})$ is equivalent to $\mathbf{w}^* = \arg \min_{\mathbf{w}} E(\mathbf{w})$

$$E(\mathbf{w}) = \frac{1}{N} \left[- \sum_{i=1}^N \log p(y_i|\mathbf{x}_i, \mathbf{w}) - \log p(\mathbf{w}) \right]$$

- By assuming a constant/flat prior we obtain (Maximum Likelihood learning)

$$E(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N [-y_i \log(\theta_i) - (1 - y_i) \log(1 - \theta_i)], \quad \theta_i = \sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w}) = \frac{1}{1 + e^{-\tilde{\mathbf{x}}_i^\top \mathbf{w}}}$$

Quiz 3: Logistic regression

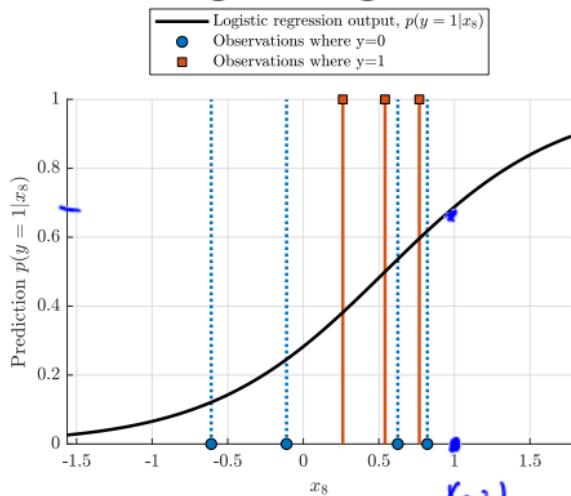


Figure 1: Output of a logistic regression classifier trained on 7 observations from the dataset.

Consider the Avila Bible dataset. We are particularly interested in predicting whether a bible copy was written by copyist 1, and we therefore wish to train a logistic regression classifier to distinguish between copyist one vs. copyist two and three.

To simplify the setup further, we select just 7

$$f(x; w) = w_0 + w_1 x_8$$

observations and train a logistic regression classifier using only the feature x_8 as input (as usual, we apply a simple feature transformation to the inputs to add a constant feature in the first coordinate to handle the intercept term). To be consistent with the lecture notes, we label the output as $y = 0$ (corresponding to copyist one) and $y = 1$ (corresponding to copyist two and three).

In Figure 1 is shown the predicted output probability an observation belongs to the positive class, $\underline{p(y=1|x)}$. What are the weights?

A. $\begin{bmatrix} -0.93 \\ 1.72 \end{bmatrix}$

$$p(y=1|w, x) = \theta^y (1-\theta)^{1-y}$$

$y=1:$

B. $\begin{bmatrix} -2.82 \\ 0.0 \end{bmatrix}$

$$p(y=1|w, x) = \theta$$

C. $\begin{bmatrix} 1.36 \\ 0.4 \end{bmatrix}$

$$\theta = G(w^T \tilde{x})$$

D. $\begin{bmatrix} -0.65 \\ 0.0 \end{bmatrix}$

$$= \frac{1}{1 + e^{-(w_0 + w_1 x_8)}}$$

E. Don't know.

$$y=1$$

$$\frac{1}{1 + e^{-(w_0 + w_1 x_8)}}$$

The solution is easily found by simply computing the predicted $\hat{y} = p(y = 1|x_8)$ -value for an appropriate choice of x_8 . Notice that

$$p(y = 1|x_8) = \sigma(\hat{\mathbf{x}}_8^T \mathbf{w})$$

If we select $x_8 = 1$ and select the weights as in option

A we find $p(y = 1|x_8) = 0.69$, in good agreement with the figure. On the other hand, for the weights in option C we obtain $\hat{y} = 0.85$, for D that $\hat{y} = 0.34$ and finally for B that $\hat{y} = 0.06$. We can therefore conclude that A is correct.

Generalized linear model

Linear regr.: $E(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N \|y_i - \tilde{\mathbf{x}}_i^\top \mathbf{w}\|^2$

Logistic regr.: $E(\mathbf{w}) = \frac{-1}{N} \sum_{i=1}^N [y_i \log(\theta_i) + (1-y_i) \log(1-\theta_i)], \quad \theta_i = \sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w})$

GLM $E(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N d(y_i, g(\tilde{\mathbf{x}}_i^\top \mathbf{w}))$

Generalized linear model

Linear regr.: $E(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N \|y_i - \tilde{\mathbf{x}}_i^\top \mathbf{w}\|^2$

Logistic regr.: $E(\mathbf{w}) = \frac{-1}{N} \sum_{i=1}^N [y_i \log(\theta_i) + (1-y_i) \log(1-\theta_i)], \quad \theta_i = \sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w})$

GLM $E(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N d(y_i, g(\tilde{\mathbf{x}}_i^\top \mathbf{w}))$

We call d the cost function and g the link function. In our examples:

Lin.reg. : $d(y, z) = \|y - z\|^2, \quad z = g(\tilde{\mathbf{x}}_i^\top \mathbf{w}) = \tilde{\mathbf{x}}_i^\top \mathbf{w}$

Log.reg. : $d(y, z) = -y \log z - (1-y) \log(1-z), \quad z = g(\tilde{\mathbf{x}}_i^\top \mathbf{w}) = \sigma(\tilde{\mathbf{x}}_i^\top \mathbf{w})$

Resources

<http://www2.imm.dtu.dk> Our interactive regression demo

(<http://www2.imm.dtu.dk/courses/02450/DemoComplexityRegression.html>)

\hat{w}



02450: Introduction to Machine Learning and Data Mining

Probability densities and data visualization

Bjørn Sand Jensen

DTU Compute, Technical University of Denmark (DTU)

$$f(x+\Delta x) = \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} f^{(i)}(x)$$

A collage of mathematical symbols and numbers including $\sqrt{17}$, Θ , Ω , δ , $e^{i\pi} = -1$, Σ , λ , ∞ , χ^2 , \gg , \approx , \approx , $\{2.7182818284\}$, and dtu compute .

DTU Compute

Department of Applied Mathematics and Computer Science

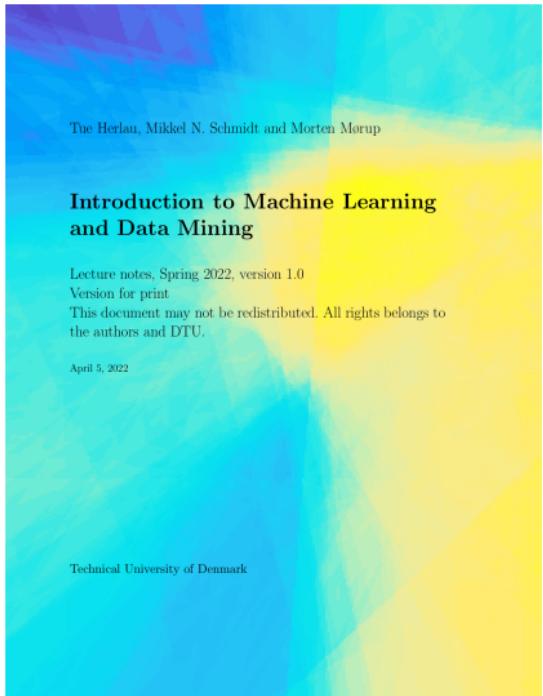
Today

Feedback Groups of the day:

Collister Chua, Rasmus Nørgaard Clausen, Emmie Cluzel, José Dinis Coelho Da Silva Ferreira, Roei Cohen, Arthur Compoint, Oscar Augusto Cordero Sosa, Alice Cruz Coimbra De Matos, Louis Cuendet, Martina Curcuruto, Florinda Da Luz, Ahmad Khaled Dahbour, Asmus Tang Dalsgaard, Rói Dam Dalsgaard, Mykhailo Datsenko, Larissa de Souza Dos Santos, Maria Teresa de Victoria Pereira Rebello, Sofie Bjerre Degen, Rodrigo Delgado Sapien, Athene Elizabeth Stuart Demuth, Yokesh Dhanabal, Dionysios Dimitreas, Denitsa Diyanova Dimitrova, Ella Kirsten Dinesen, Emile Hourman Ditlefsen, Pætur Djurhuus, Gergely Dombóvári, Zheng Dong, Tommaso Dordoni, Brandon Xintian Duan, Oussama Eddaoudi, Málfríður Anna Eiríksdóttir, Abdessamad El Kabid, Marius Mainz Elkjær, Anas Majed El-Youssef, Lukas Blander Enevoldsen, Daniel Engelberg, Mads Dan Eriksen, Rune Esbjerg, Alberte Heering Estad, Ejnar Billeskov Exsteen, Romisa Fakhari, Kaizhi Fan, Ivan Fan

Reading material:

Chapter 6, Chapter 7



Lecture Schedule

1 Introduction

30 January: C1

Data: Feature extraction, and visualization

2 Data, feature extraction and PCA

6 February: C2, C3

3 Measures of Similarity, summary statistics and probabilities

13 February: C4, C5

4 Probability densities and data visualization

20 February: C6, C7

Supervised learning: Classification and regression

5 Decision trees and linear regression

27 February: C8, C9 (Project 1 due 29 February at 17:00)

6 Overfitting, cross-validation and Nearest Neighbor

5 March: C10, C12

7 Performance evaluation, Bayes, and Naive Bayes

12 March: C11, C13

8 Artificial Neural Networks and Bias/Variance

19 March: C14, C15

9 AUC and ensemble methods

2 April: C16, C17

Unsupervised learning: Clustering and density estimation

10 K-means and hierarchical clustering

9 April: C18 (Project 2 due 11 April at 17:00)

11 Mixture models and density estimation

16 April: C19, C20

12 Association mining

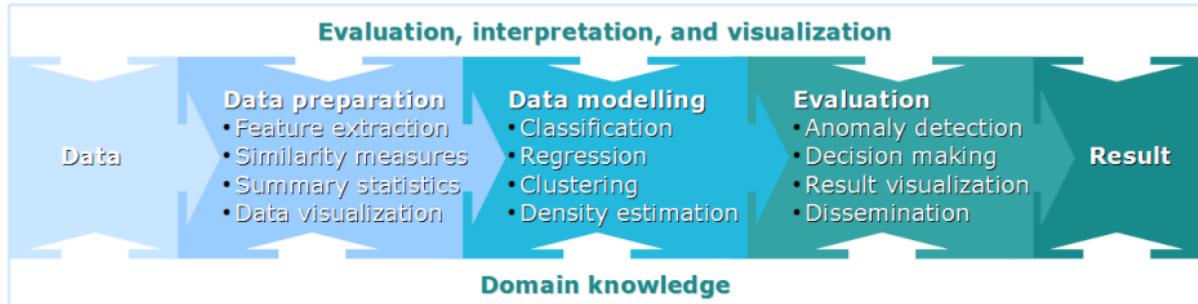
23 April: C21

Recap

13 Recap and discussion of the exam

30 April: C1-C21

Online 24/7 help: Discussion Forum/Piazza
Streaming: Zoom (see link on DTU Learn)
Recordings: <https://panopto.dtu.dk/>
Online exercises: MS Teams



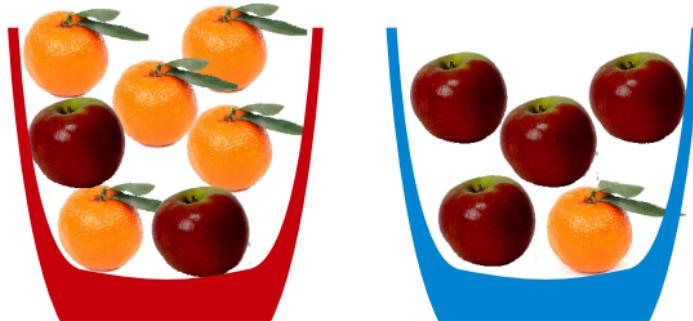
Learning Objectives

- Understand probability densities and related concepts
- Derive cost-functions from likelihood functions using Bayes' theorem
- Understand and apply a wide range of data visualization approaches
- Understand good practice in plotting including Tufte's guidelines

Probabilities recap

Example: Computing with probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



Apple taken from: https://upload.wikimedia.org/wikipedia/commons/3/32/Dark_apple.png
Orange (clementine) taken from: https://commons.wikimedia.org/wiki/File:Clementine_orange.jpg

Probabilities

- In more common notation we have

- Sum rule

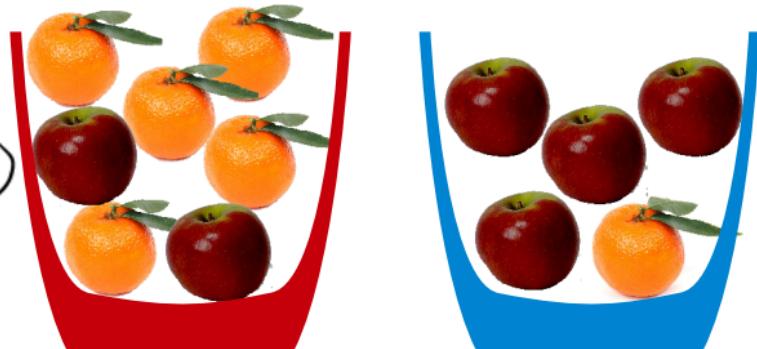
$$p(x) = p(x, y=0) + p(x, y=1)$$

- Product rule

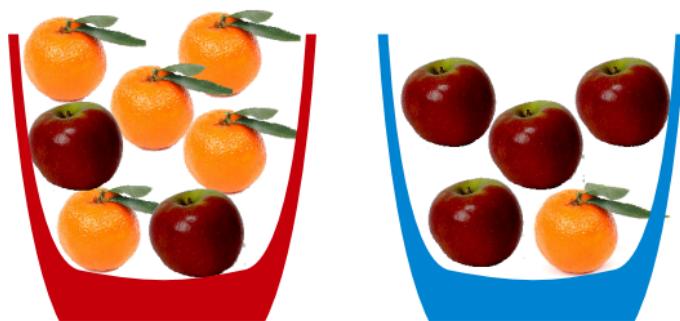
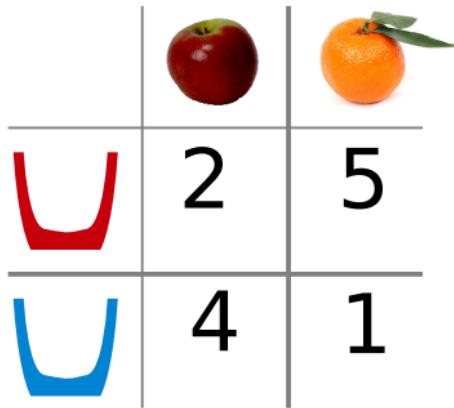
$$\begin{aligned} p(x, y) &= p(x|y)p(y) \\ &= p(y|x)p(x) \end{aligned}$$

- Bayes' rule

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$



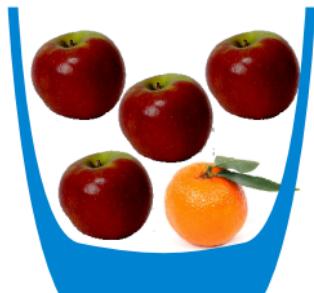
- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

| | | | |
|--|---|----|---|
|  | 2 | 5 | 7 |
|  | 4 | 1 | 5 |
| 6 | 6 | 12 | |

$p(\text{FruitType}, \text{Color})$



- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

| | | | |
|---|---|---|---|
| | |  |  |
| U | 2 | 5 | 7 |
| U | 4 | 1 | 5 |
| 6 | 6 | 12 | |

$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

| | | | |
|---|---|---|---|
| | |  |  |
| U | 2 | 5 | 7 |
| U | 4 | 1 | 5 |
| 6 | 6 | 12 | |

$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

$$p(r|o) = \frac{p(r,o)}{p(o)} = \frac{5/12}{6/12} = 5/6$$

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

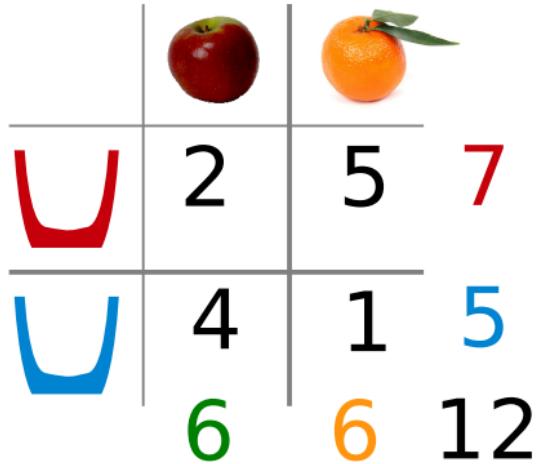
| | | | |
|---|---|---|---|
| | |  |  |
| U | 2 | 5 | 7 |
| U | 4 | 1 | 5 |
| 6 | 6 | 12 | |

$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

$$p(r|o) = \frac{p(r,o)}{p(o)} = \frac{5/12}{6/12} = 5/6$$

$$= \frac{p(o|r)p(r)}{p(o)}$$

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

$$p(r|o) = \frac{p(r,o)}{p(o)} = \frac{5/12}{6/12} = 5/6$$

$$= \frac{p(o|r)p(r)}{p(o)}$$

$$= \frac{5/7 \cdot 7/12}{6/12} = 5/6$$



Medical test

A medical test for a given disease

- Correctly identifies the disease 99% of the time (true positives), and
- Incorrectly turns out positive 2% of the time (false positives).

You know that

- 1% of the population suffers from the disease.

You go to the doctor to get tested, and the test turns out to be positive.

What is the probability you have the disease?

Hints:

- Identify from the text: ($x = \text{Positive}$,
 $y=0: \text{no disease}$, $y=1: \text{Disease}$)

$p(\text{Positive}|\text{Disease})$

$p(\text{Positive}|\text{No Disease})$

$p(\text{Disease})$

$p(\text{No Disease})$

- Use the basic rules of probability given to the right to find:

$p(\text{Disease}|\text{Positive})$

$$\begin{aligned} p(y) &= \sum_x p(y, x) \\ &= p(y|x)p(x) + p(y|\bar{x})p(\bar{x}) \end{aligned}$$

$$p(x, y) = p(x|y)p(y)$$

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

$$P(+|D) = \underline{0.99}$$

$$P(+|\bar{D}) = 0.02$$

$$P(D) = \underline{0.01}$$

$$P(-|D) = 0.01$$

$$P(-|\bar{D}) = 0.98$$

$$P(\bar{D}) = 0.99$$

$$P(D|+) = \frac{P(+, D)}{P(+)} \approx \frac{P(+|D)P(D)}{(P(+))}$$

$$= \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|\bar{D})P(\bar{D})}$$

$$= \frac{0.99 \times 0.01}{0.99 \times 0.01 + 0.02 \times 0.99} = \frac{1}{3}$$

Quiz 1, Probabilities (Spring 2014)

Consider a dataset which describe the consumption of delicatessen products in different cities. Each observation in the dataset is a customer, and we record the city the customer is from as well as their consumption of delicatessen. Suppose you are told:

- 17.5 % were from Lisbon, 10.7 % were from Oporto and 71.8 % from the Other region.
- 44.1 % of the costumers from Lisbon spent above the median consumption on delicatessen (DELI).
- 48.9 % of the costumers from Oporto spent above the median consumption on delicatessen (DELI).
- 51.6 % of the costumers from the Other region spent above the median consumption on delicatessen (DELI).

What is the probability based on the wholesale data that a costumer that spent above the median consumption on delicatessen (DELI) come from Lisbon?

- A. 7.7 %
 B. 15.4 %
 C. 44.1 %
 D. 59.6 %
 E. Don't know.

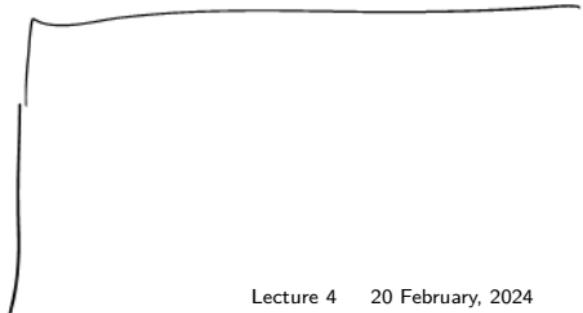
$$\underbrace{P(Lis|A)}_{?} = ?$$

$$P(Lis) = 0.175 \quad P(Op) = 0.107 \quad P(ot) = 0.718$$

$$P(A|Lis) = 0.441$$

$$P(A|Op) = \underline{\hspace{2cm}}$$

$$P(A|ot) = \underline{\hspace{2cm}}$$



$$\begin{aligned} p(Lis | A) &= \frac{p(Lis, A)}{p(A)} \\ &= \frac{p(A | Lis) p(Lis)}{\sum_{c \in \{Lis, op, ot\}} p(A | c) p(c)} \\ &= 0.154 \end{aligned}$$

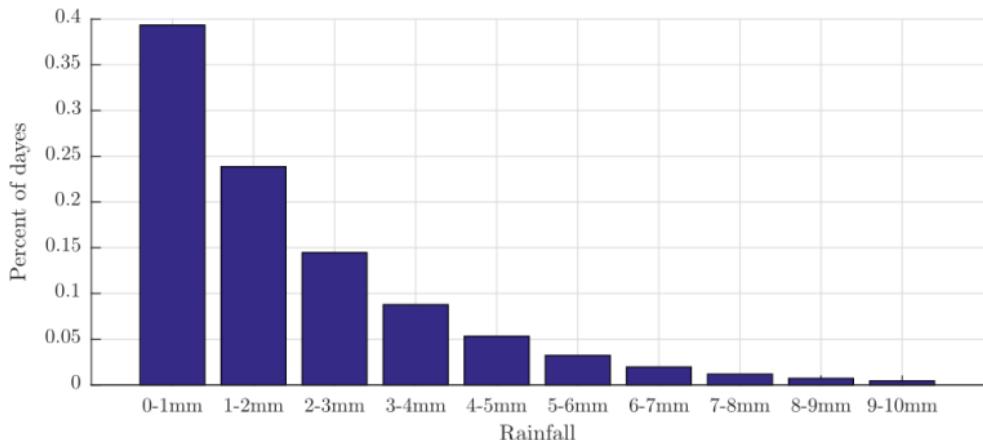
We will let *LISBON*, *OPORTO* and *OTHER* denote coming from Lisbon, Oporto and the other region respectively. $DELI_H$ will denote above the median value of delicatessen consumption. We now

find using Bayes theorem:

$$\begin{aligned} P(LISBON|DELI_H) &= \frac{P(DELI_H|LISBON)P(LISBON)}{P(DELI_H)} \\ &= \frac{P(DELI_H|LISBON)P(LISBON)}{P(DELI_H|LISBON)P(LISBON) + P(DELI_H|OPORTO)P(OPORTO) + P(DELI_H|OTHER)P(OTHER)} \\ &= \frac{0.441 \cdot 0.175}{0.441 \cdot 0.175 + 0.489 \cdot 0.107 + 0.516 \cdot 0.718} \\ &= 0.1544 \approx 15.4\% \end{aligned}$$

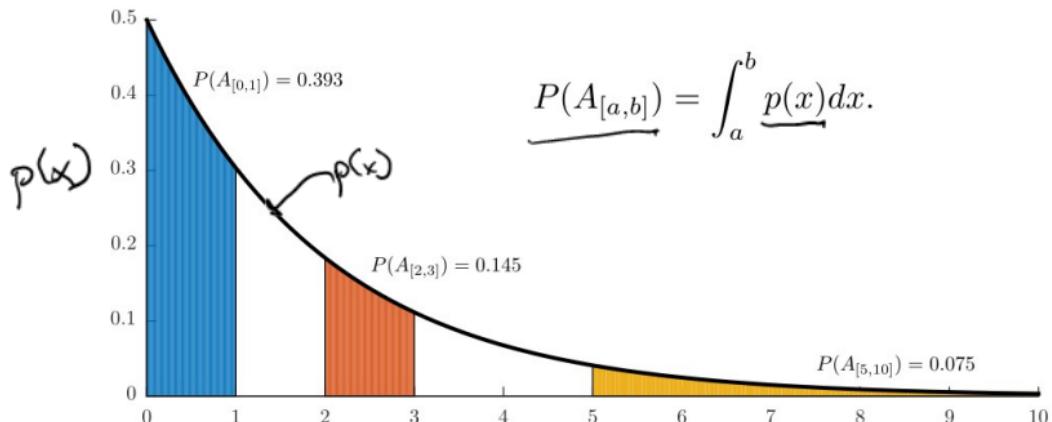
Probability vs. Density

- Suppose we consider the rainfall on an average day r
- **Can't** talk about the probability there will be **exactly** $r=2.3$ mm of rain, $P(r=2.3\text{mm})$
- **Can** talk about the probability there will be **between** 1 and 2 mm of rain



Probability vs. Density

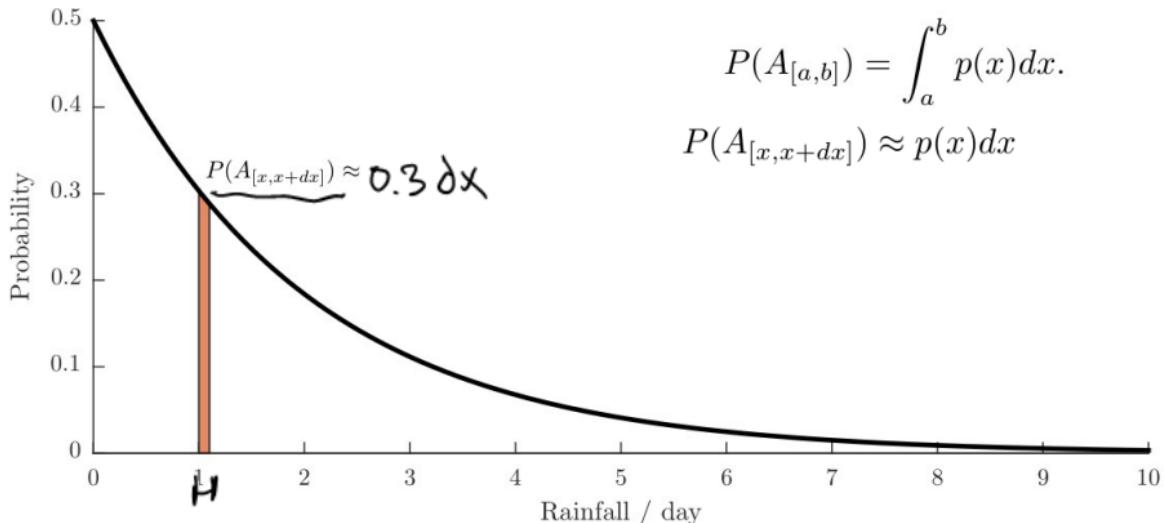
- These probabilities can be **represented** as integrals
- **Events** are **intervals**, the **probability** is the **integral**, the **curve** is the **density**



$A_{[a,b]}$: There will be between a and b mm of rain

Probability vs. Density

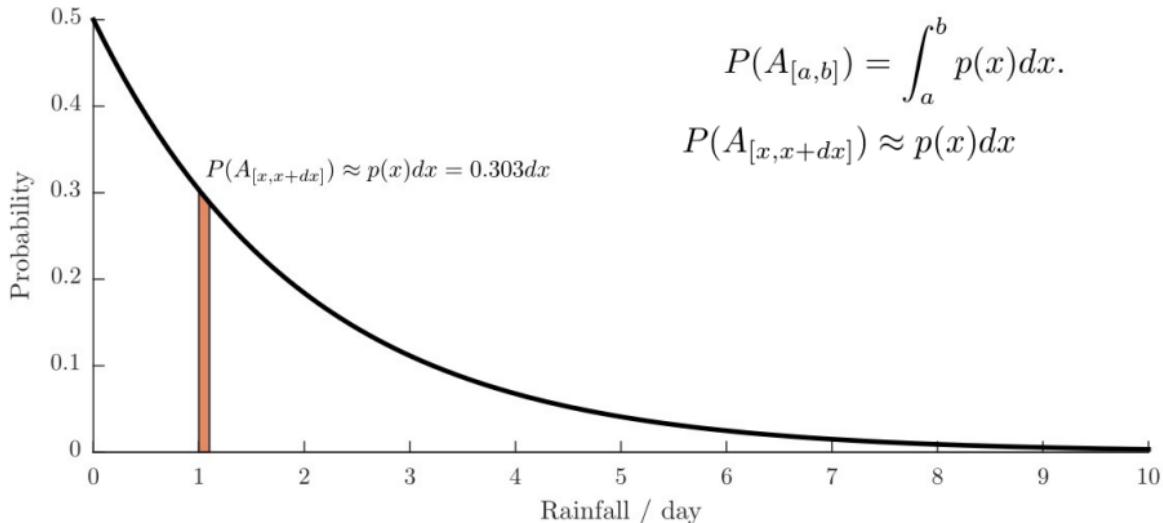
- These probabilities can be **represented** as integrals
- **Events** are **intervals**, the **probability** is the **integral**, the **curve** is the **density**
- What is the probability there will be between 1 and 1.1 mm of rain?



$A_{[a,b]}$: There will be between a and b mm of rain

Probability vs. Density

- These probabilities can be **represented** as integrals
- **Events** are **intervals**, the **probability** is the **integral**, the **curve** is the **density**
- What is the probability there will be between 1 and 1.1 mm of rain?



$A_{[a,b]}$: There will be between a and b mm of rain

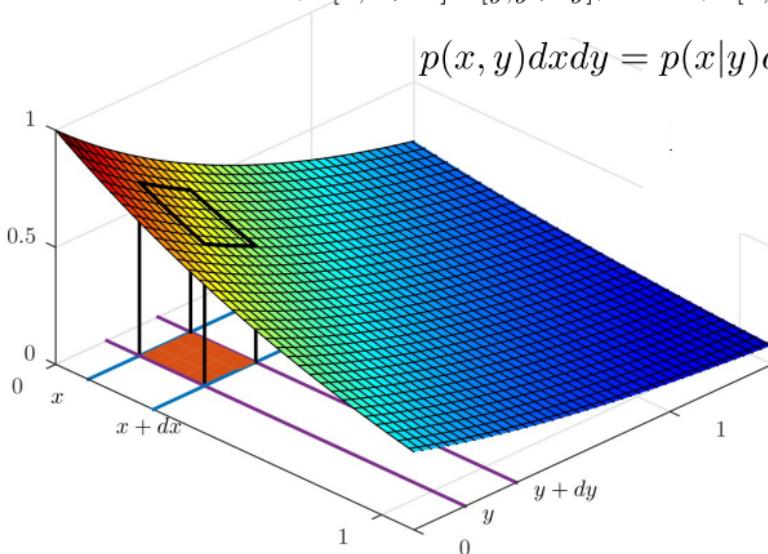
Probability vs. Density

- For two variables x and y , the **probability** is an integral over an **area**

$$P((x, y) \in D) = \int_{(x,y) \in D} p(x, y) dx dy$$

$$P(A_{[x, x+dx]} B_{[y, y+dy]}) = P(A_{[x, x+dx]} | B_{[y, y+dy]}) P(B_{[y, y+dy]})$$

$$p(x, y) dx dy = p(x|y) dx p(y) dy$$



This implies:

$$p(x, y) = p(y|x)p(x)$$

Probability vs. Density

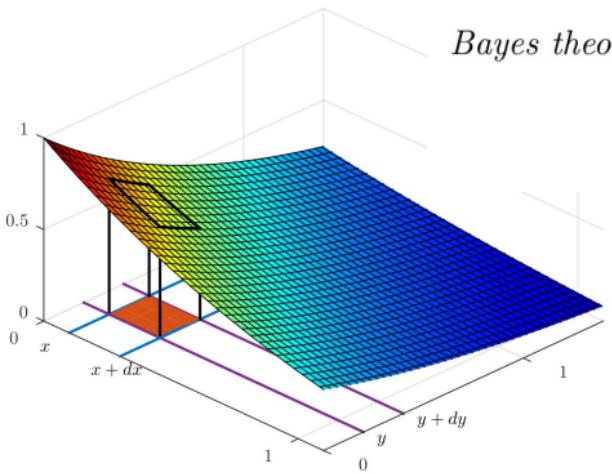
- Thus, we have shown the rules of probability theory also holds for densities

The sum rule:

$$\int dx \ p(x|z) = 1$$

The product rule:

$$p(x, y|z) = p(y|x, z)p(x|z)$$



$$\begin{aligned} p(x|y, z) &= \frac{p(y|x, z)p(x|z)}{p(y|z)} \\ &= \frac{p(y|z)p(x|y, z)}{\int p(y|x', z)p(x'|z)dx'}. \end{aligned}$$

Collecting all of this we obtain:

- Rules of probability for densities

Marginalization:

$$\int p(x, y|z)dx = p(y|z)$$

The product rule:

$$p(x, y|z) = p(y|x, z)p(x|z)$$

Bayes theorem:

$$p(x|y, z) = \frac{p(y|x, z)p(x|z)}{\int p(y|x', z)p(x'|z)dx'}.$$

- Rules of probability for discrete variables

Marginalization:

$$\sum_c p(x = c, y|z) = p(y|z)$$

The product rule:

$$p(x, y|z) = p(y|x, z)p(x|z)$$

Bayes theorem:

$$p(x|y, z) = \frac{p(y|x, z)p(x|z)}{\sum_c p(y|x = c, z)p(x = c|z)}.$$

Expected values

- Discrete random variable

$$\overbrace{\mathbb{E}[g]}^{\text{Discrete}} = \sum_i g(x_i)P(x_i)$$

$$\hat{x} = \frac{1}{N} \sum_i x_i$$

- Continuous random variable

$$\overbrace{\mathbb{E}[g]}^{\text{Continuous}} = \int_{-\infty}^{\infty} g(x)p(x)dx$$

Statistics

- **Mean**

$$\bar{x} = \mathbb{E}[x] = \int x p(x) dx$$

- **Covariance**

$$\text{cov}(x, y) = \mathbb{E}[(x - \bar{x})(y - \bar{y})]$$

- **Variance**

$$\text{var}(x) = \text{cov}(x, x) = \mathbb{E}[(x - \bar{x})^2]$$

- **Standard deviation**

$$\text{std}(x) = \sqrt{\text{var}(x)}$$

Break 14:00

Densities and models

- In machine learning, we want to learn a parameter from data
- Models of the data which use parameters are how we do that
- We build models out of simpler building blocks (distributions (discrete variables see chapter 5) and densities (continuous variables see chapter 6)). In this course we will use four:

Bernoulli distribution

The Categorical distribution

The Beta density

The Multivariate normal density

The Mahalanobis distance

How far are x_1 and x_2 apart?

- mahalanobis($\mathbf{0}_1, \mathbf{0}_2$) = 4.2
- $d_{\text{Euclidean}}(\mathbf{0}_1, \mathbf{0}_2)^2$ = 8.0

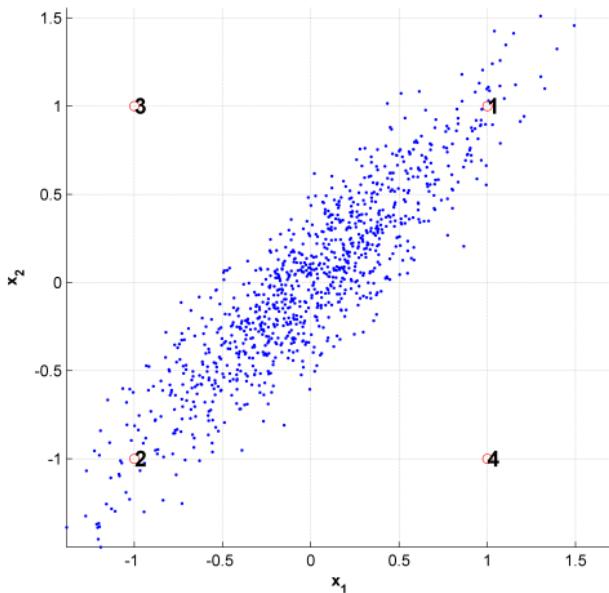
How far are x_3 and x_4 apart?

- mahalanobis($\mathbf{0}_3, \mathbf{0}_4$) = 80
- $d_{\text{Euclidean}}(\mathbf{0}_3, \mathbf{0}_4)^2$ = 8.0

$$\Sigma = \begin{bmatrix} \text{Cov}(x_1, x_1) & \text{Cov}(x_1, x_2) \\ \text{Cov}(x_2, x_1) & \text{Cov}(x_2, x_2) \end{bmatrix}$$

$$\text{mahalanobis}(x, y)^2 = (x - y)^\top \Sigma^{-1} (x - y)$$

$$d_{\text{euclidian}}(x, y)^2 = (x - y)^\top I^{-1} (x - y)$$

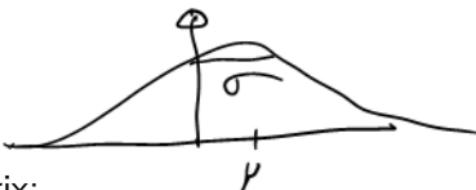


The multivariate normal distribution

A distribution for M -dimensional vectors \boldsymbol{x} :

$$\varphi(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{M}{2}} \sqrt{|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}$$

$$M = 1 : \quad \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



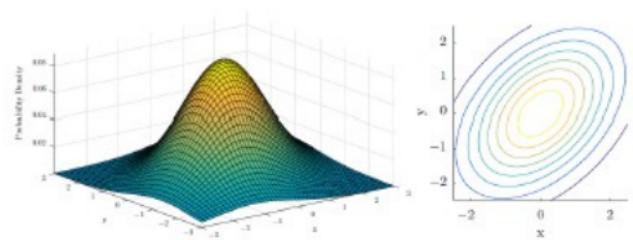
$\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix:

$$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{x}], \quad \Sigma_{ij} = \text{cov}[x_i, x_j]$$

- Example: 2-dimensional Normal distribution

$$\boldsymbol{\mu} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \text{var}(x_1) & \text{cov}(x_1, x_2) \\ \text{cov}(x_2, x_1) & \text{var}(x_2) \end{bmatrix}$$



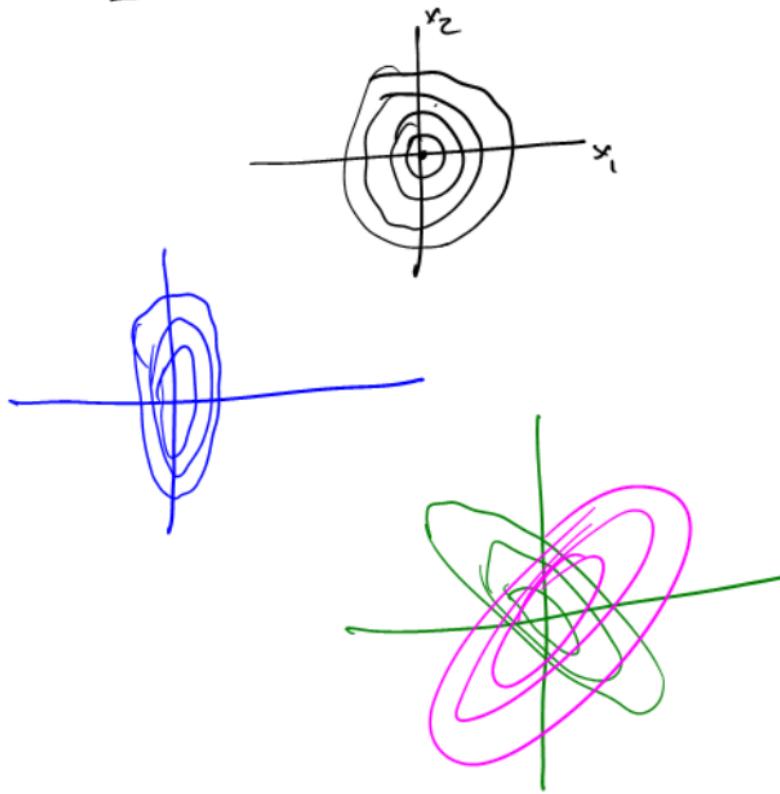
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} a & c \\ c & b \end{bmatrix}$$

$$\Sigma: a \approx b \\ c \approx 0$$

$$\Sigma: a \ll b \\ c \approx 0$$

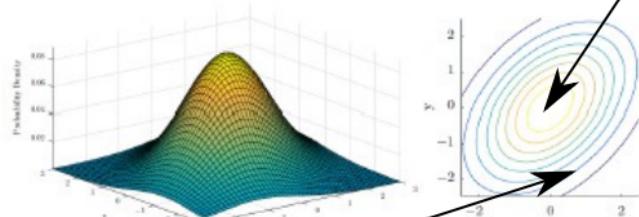
$$\Sigma: a \approx b \\ c < 0$$

$$\Sigma: c > b$$



Quiz 2, Covariance

- Match the covariances to the contour plots



$$\Sigma = \begin{bmatrix} \text{var}(x_1) & \text{cov}(x_1, x_2) \\ \text{cov}(x_2, x_1) & \text{var}(x_2) \end{bmatrix}$$

$$\mu = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix}$$

A. Covariance of A is $\Sigma_A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, $\Sigma_C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

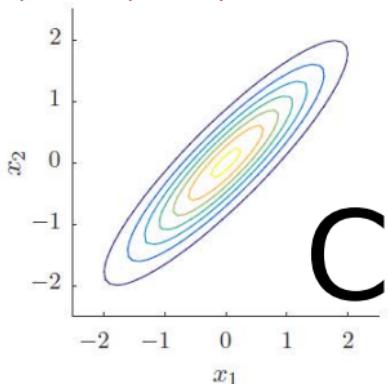
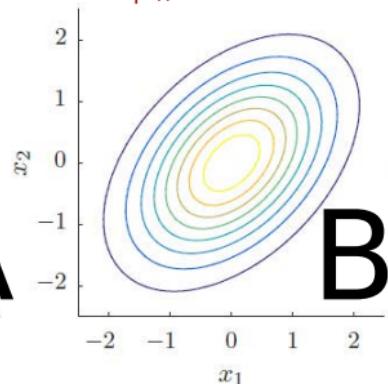
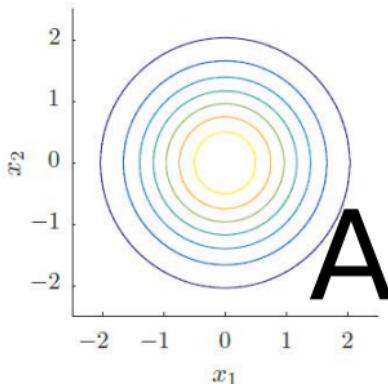
B. $\Sigma_B = \begin{bmatrix} 1 & 0.45 \\ 0.45 & 1 \end{bmatrix}$, $\Sigma_C = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$

C. $\Sigma_B = \begin{bmatrix} 10 & 4.5 \\ 4.5 & 10 \end{bmatrix}$, $\Sigma_C = \begin{bmatrix} 10 & 9 \\ 9 & 10 \end{bmatrix}$

D. $\Sigma_A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\Sigma_C = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$

E. Don't know.

Check out the online demo <http://www2.imm.dtu.dk/courses/02450/DemoNormal.html>



The right answer is *D*. The covariance has to be positive (because x_1 and x_2 are positively correlated), and the variance is 1 in all cases. Furthermore, since A is axis-aligned, the covariance terms are zero. All

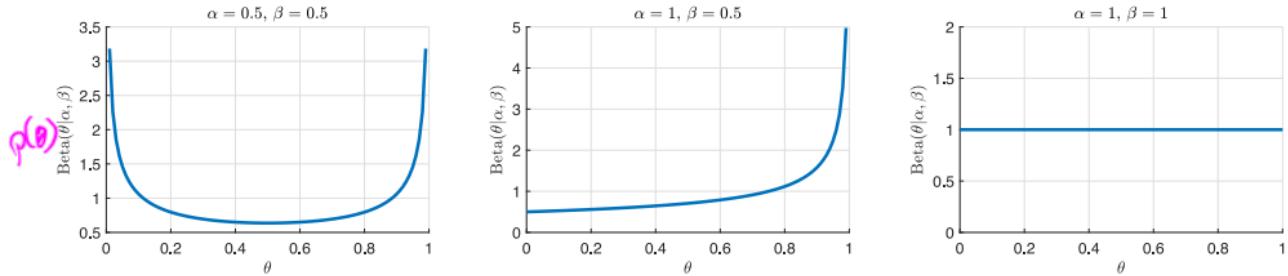
in all

$$\Sigma_A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \Sigma_B = \begin{bmatrix} 1 & 0.45 \\ 0.45 & 1 \end{bmatrix}, \quad \Sigma_C = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$$

Beta distribution

Suppose θ is defined on the unit interval $[0, 1]$

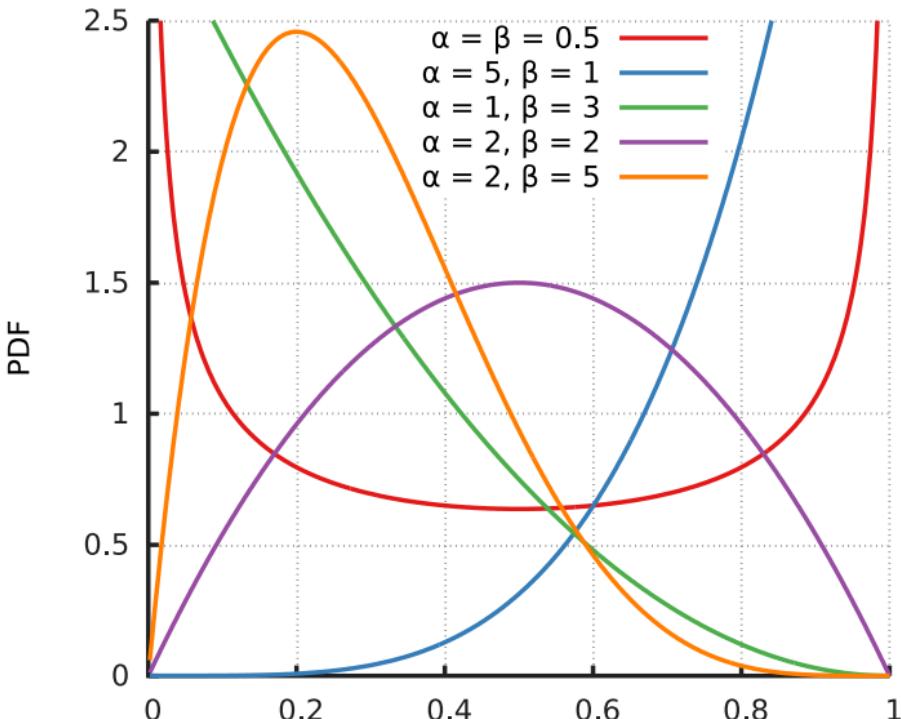
Beta density: $p(\theta|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$



$\alpha, \beta > 0$ are related to the variance and mean

$$\mathbb{E}_{p(\theta|\alpha, \beta)}[\theta] = \frac{\alpha}{\alpha + \beta}, \quad \text{Var}_{p(\theta|\alpha, \beta)}[\theta] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

Beta distribution



Probabilities and learning

- Your friend just got a dog.
- The dog can either be in the doghouse or outside
- The first four times you come by the dog is in the doghouse
- What is the chance the dog is in the doghouse tomorrow?



- Your friend buys a coin.
- The coin can either come up heads or tails
- The first four times you flip the coin it comes up tails
- What is the chance the coin comes up tails in the next flip?



Intuition tells us the answers are different, but the situation seems similar...

Recall from lecture 3: The Bernoulli distribution

- Suppose a coin come up tails with probability θ
 - Suppose $b = 0$ is the event the coin land heads
 - $b = 1$ is the event the coin land tails
- The density is given by the **Bernoulli distribution**

$$p(b|\theta) = \theta^b(1-\theta)^{1-b}$$

- For a sequence of N flips b_1, b_2, \dots, b_N

Independence

$$\begin{aligned} p(b_1, \dots, b_N | \theta) &= \prod_{i=1}^N p(b_i | \theta) \\ &= \theta^{\sum_{i=1}^N b_i} (1-\theta)^{N - \sum_{i=1}^N b_i} \\ &= \theta^m (1-\theta)^{N-m}, \quad m = b_1 + b_2 + \dots + b_N \end{aligned}$$

• What is θ ? $\underset{\theta}{\operatorname{argmax}} \theta^m (1-\theta)^{N-m} = \frac{\theta^4}{\overbrace{?}^4}$

The Bernoulli distribution

- A magic coin is a coin that comes up tails with probability θ
 - Suppose $b = 0$ is the event the coin land heads
 - $b = 1$ is the event the coin land tails
- For a sequence of N flips b_1, b_2, \dots, b_N

$$p(b_1, \dots, b_N) = \theta^m(1 - \theta)^{N-m}, \quad m = b_1 + b_2 + \dots + b_N$$

- **What is θ ?** Answer: **Use Bayes' Theorem!**

$$p(\theta|\mathbf{b}) = \frac{p(\mathbf{b}|\theta)p(\theta)}{p(\mathbf{b})} = \frac{p(\mathbf{b}|\theta)p(\theta)}{\int_0^1 p(\mathbf{b}|\theta')p(\theta')d\theta'}$$

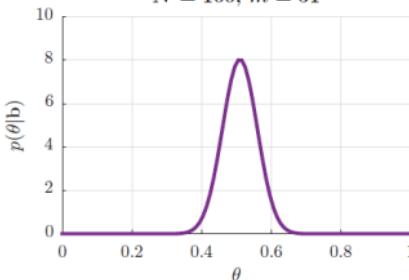
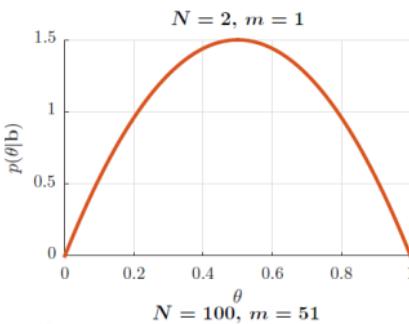
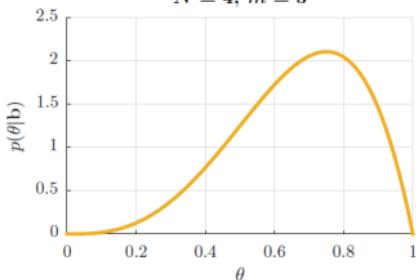
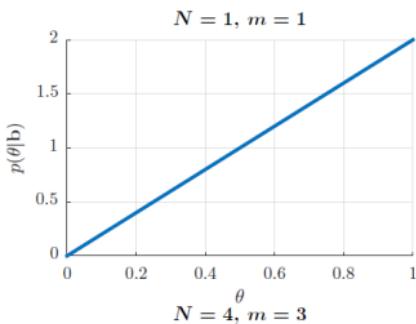
- Assume $p(\theta) = p(\theta|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$

$$\begin{aligned} p(\theta|\mathbf{b}, \alpha, \beta) &= \frac{\theta^m(1-\theta)^{N-m} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int \theta'^m(1-\theta')^{N-m} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta'^{\alpha-1}(1-\theta')^{\beta-1}d\theta'} \\ &= \frac{\Gamma(\alpha + \beta + N)}{\Gamma(\alpha + m)\Gamma(\beta + N - m)}\theta^{\alpha+m-1}(1-\theta)^{\beta+N-m-1} \end{aligned}$$

Example: $\alpha = \beta = 1$



$$\begin{aligned} p(\theta | \mathbf{b}, \alpha, \beta) &= \frac{\Gamma(\alpha + \beta + N)}{\Gamma(\alpha + m)\Gamma(\beta + N - m)} \theta^{\alpha+m-1} (1-\theta)^{\beta+N-m-1} \\ &= \frac{(N+1)!}{m! (N-m)!} \theta^m (1-\theta)^{N-m} \end{aligned}$$



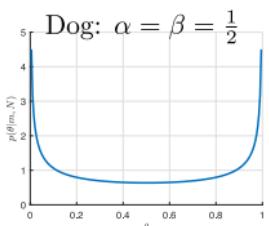
Dogs and coins



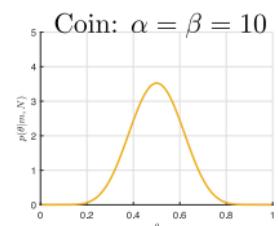
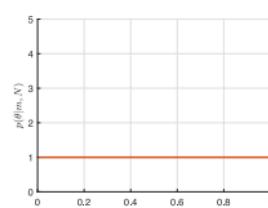
- Your friend just got a dog.
- The dog can either be in the doghouse or outside
- The first four times you come by the dog is in the doghouse
- What is the chance the dog is in the doghouse tomorrow?

Prior

$$p(\theta|\alpha, \beta) =$$



- Your friend buys a coin.
- The coin can either come up heads or tails
- The first four times you flip the coin it comes up tails
- What is the chance the coin comes up tails in the next flip?



Likelihood

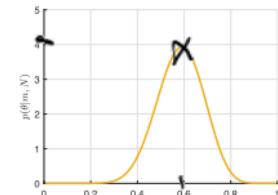
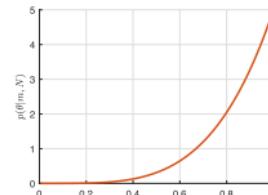
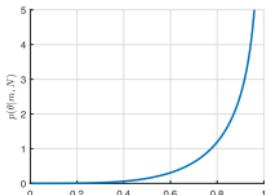
$$p(m=4, N=4|\theta) = \theta^m(1-\theta)^{N-m} = \theta^4$$

The difference between the two cases is that we have prior knowledge which tell us most coins are fair, and this affects our conclusions.

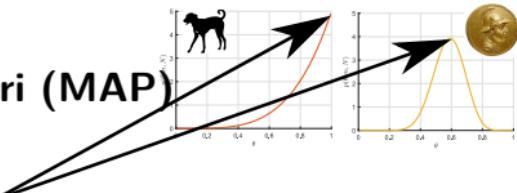
In most practical situations, we should assume as little as possible and choose $\alpha = \beta = \frac{1}{2}$

Posterior

$$p(\theta|m, N) = \frac{p(m, N|\theta)p(\theta|\alpha, \beta)}{p(m, N)} =$$



Learning principle: Maximum a posteriori (MAP)



- Another idea: Select θ which is "most probably"

$$\theta^* = \arg \max_{\theta} p(\theta | M, N) = \arg \max_{\theta} \left[\frac{p(m, N | \theta) p(\theta | \alpha, \beta)}{p(M, N)} \right]$$

- Use that $\arg \max_x f(x) = \arg \min_x [-\log f(x)]$ if $f(x) > 0$:

$$\theta^* = \arg \min_{\theta} \left[-\log \frac{p(m, N | \theta) p(\theta | \alpha, \beta)}{p(m, N)} \right]$$

(likelihood)

$$p(m, N | \theta) = \theta^m (1 - \theta)^{N-m}$$

(prior)

$$p(\theta | \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}$$

- All in all:

$$\theta^* = \arg \min E(\theta), \quad E(\theta) = -\log p(m, N | \theta) - \log p(\theta | \alpha, \beta)$$

A learning principle: Maximum a posteriori (MAP)

- Another idea: Select θ which is "most probably"

$$\theta^* = \arg \max_{\theta} p(\theta|M, N) = \arg \max_{\theta} \left[\frac{p(m, N|\theta)p(\theta|\alpha, \beta)}{p(M, N)} \right]$$

- Use that $\arg \max_x f(x) = \arg \min_x [-\log f(x)]$ if $f(x) > 0$:

$$\begin{aligned}\theta^* &= \arg \min_{\theta} \left[-\log \frac{p(m, N|\theta)p(\theta|\alpha, \beta)}{p(m, N)} \right] \\ &= \arg \min_{\theta} [-\log p(m, N|\theta) - \log p(\theta|\alpha, \beta) + \log p(m, N)] \\ &= \arg \min_{\theta} [-\log p(m, N|\theta) - \log p(\theta|\alpha, \beta)]\end{aligned}$$

- All in all:

$$\theta^* = \arg \min E(\theta), \quad E(\theta) = -\log p(m, N|\theta) - \log p(\theta|\alpha, \beta)$$

Maximum a posteriori (MAP) learning

- Consider some data $\mathbf{X} = \mathbf{x}_1, \dots, \mathbf{x}_N$ and $\mathbf{y} = y_1, \dots, y_N$
- Suppose we think x_i relates to y_i by some parameters \mathbf{w}
- **Assume**

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^N p(y_i|\mathbf{x}_i, \mathbf{w}), \quad p(\mathbf{w}|\mathbf{X}) = p(\mathbf{w})$$

Observations are not informative about each other when we know parameters

Without \mathbf{y} , we cannot learn the parameters

- Then

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\mathbf{X})}{p(\mathbf{y}|\mathbf{X})}$$

- The following are equivalent:

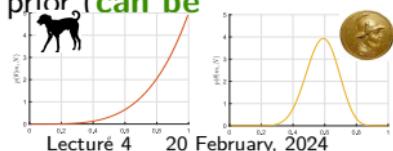
$$\text{Maximize: } \mathbf{w}^* = \arg \max_{\mathbf{w}} p(\mathbf{w}|\mathbf{X}, \mathbf{y})$$

$$\text{Minimize: } \mathbf{w}^* = \arg \min_{\mathbf{w}} E(\mathbf{w}), \quad E(\mathbf{w}) = \left[\frac{1}{N} \sum_{i=1}^N -\log p(y_i|\mathbf{x}_i, \mathbf{w}) \right] - \frac{1}{N} \log p(\mathbf{w})$$

- All we need is a likelihood (**usually pretty simple**) and a prior (**can be omitted**) and we have a machine-learning method.

- **Pro:** Easy, conceptually simple, efficient

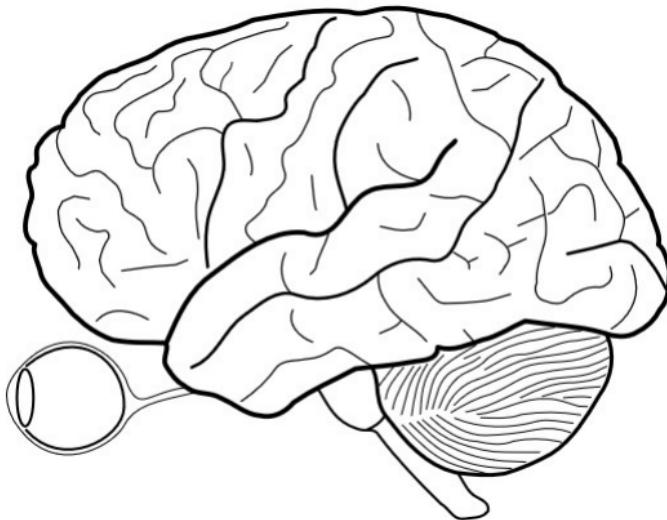
- **Con:** Can sometimes give spurious results (overfit)



The drawing shows me at one glance what might be spread over ten pages in a book."
- Ivan S. Turgenev's novel Fathers and Sons, 1862.
Use a picture. It's worth a thousand words."
- Arthur Brisbane to the Syracuse Advertising Men's Club, in March 1911

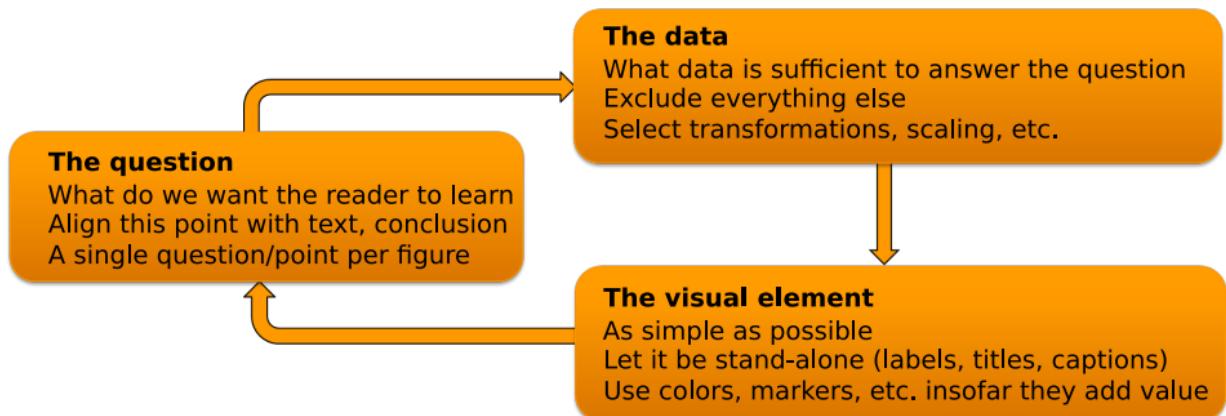
Visualization

- A main function of the brain is to process visual information
- We can exploit this capacity using visualization of the data in order to:
 - Detect new patterns, i.e. exploratory data analysis)
 - **Dissiminate results, i.e. visualizations/plots in written work (today)**
- We should take into account how the brains visual system works



Illustrations as technical writing

- The purpose of the text is to communicate an idea (*vs. plots has a purpose*)
- Be grammatically correct (*vs. elementary "rules" of good plotting*)
- Ensure the text is readable (*vs. labels, legends or lines nobody can read*)
- Avoid long/complicated paragraph (*vs. plots that are overly complicated*)
- Dont lie or exaggerate. (*vs. distort data in a plot*)

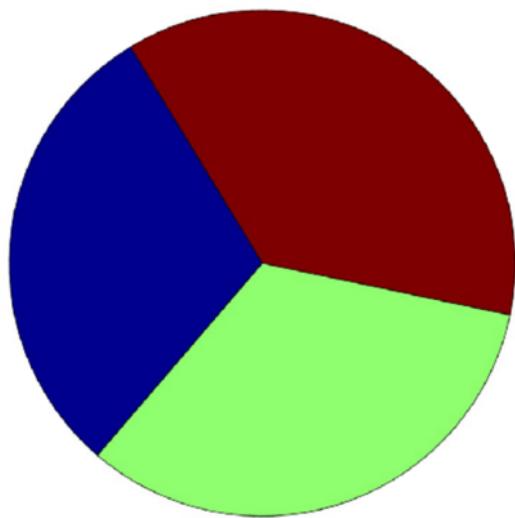


Important choices for visualizations

- **Representation:** How will you map objects, attributes, and relations to visual elements?
 - Positions, lengths, colors, areas, orientation
- **Arrangement:** How will you display the visual elements?
 - Viewpoint, transparency, separation, grouping
- **Selection:** How will you handle a large number of attributes and data objects?
 - Display a subset, focus on a region of interest, show summaries

Representation

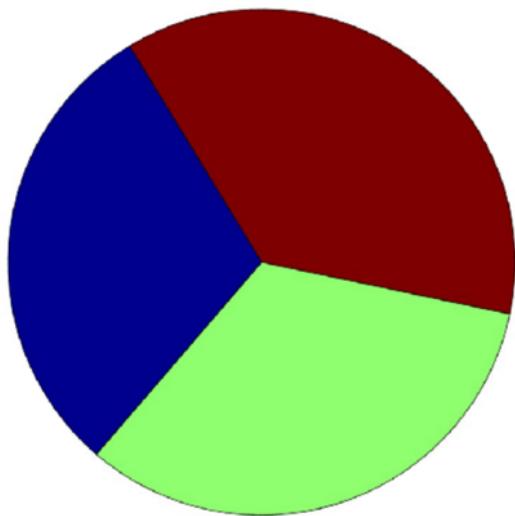
- **Area represents proportion**
 - Which is smallest, middle, and largest?
 - What are the proportions approximately?



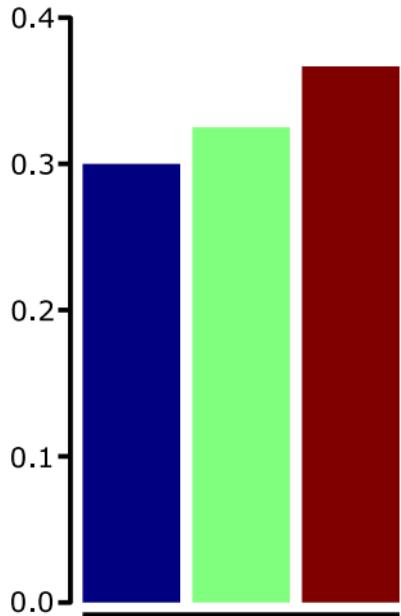
Representation

- **Area represents proportion**

- Which is smallest, middle, and largest?
- What are the proportions approximately?



- **Height represents proportion**



Selection

- Elimination or de-emphasis of certain objects or attributes
- A subset of **attributes**
 - **Why?** A graph can only show so many attributes – focus on the relevant
 - **How?**
 - Dimensionality reduction
 - Plot pairs of attributes
- A subset of **objects**
 - **Why?** A graph can only show so many objects – focus on the relevant
 - **How?**
 - Random sampling
 - Display of region of interest
 - Use density estimation

Types of plots

- **Distribution of a single attribute**

- Histogram
- Empirical cumulative distribution
- Percentile plots
- Box plot

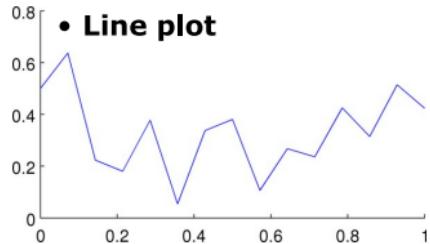
- **Relation between attributes**

- 2D histogram
- Heat maps and contour plots
- Scatter plots

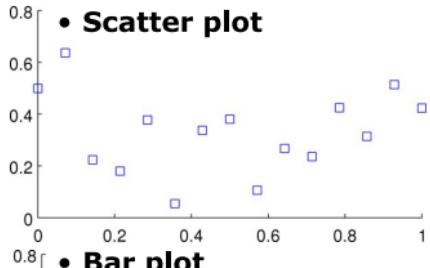
- **Visualization of high-dimensional objects**

- Matrix plots
- Parallel coordinates
- Star plots

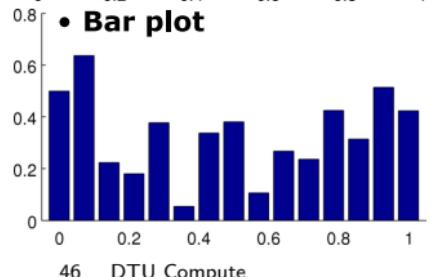
Basic plots



```
plot(x,y);
```



```
plot(x,y, 's');  
scatter(x,y, 's')
```



```
bar(x,y);
```

The iris data set

- **Three flowers**

- 50 instances of each class, 150 in total

- **Attributes**

- Sepal (outermost leaves)

- length in cm
- width in cm

- Petal (innermost leaves)

- length in cm
- width in cm

- Class of flower

- Iris Setosa
- Iris Versicolour
- Iris Virginica

| Flower ID | Attribute | | | |
|-----------|--------------|-------------|--------------|-------------|
| | Sepal Length | Sepal Width | Petal Length | Petal Width |
| 1 | 5.1 | 3.5 | 1.4 | 0.2 |
| 2 | 4.9 | 3.0 | 1.4 | 0.2 |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 |
| 4 | 4.6 | 3.1 | 1.5 | 0.2 |
| ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... |
| 150 | 5.9 | 3.0 | 5.1 | 1.8 |

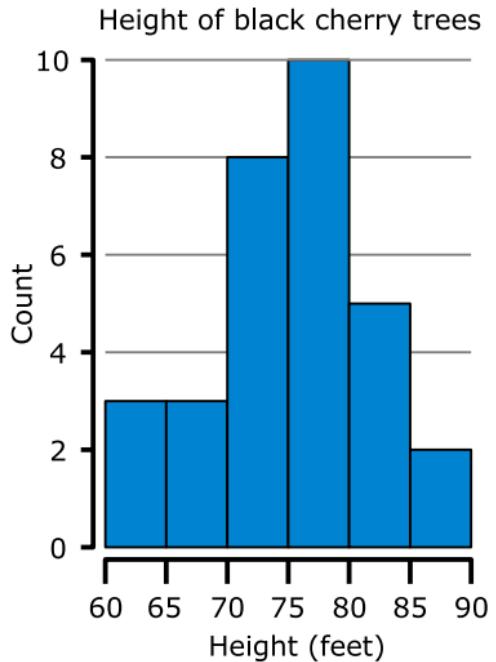
$$X^{\text{Observation} \times \text{Attribute}}$$

Distribution of a single attribute

Histograms

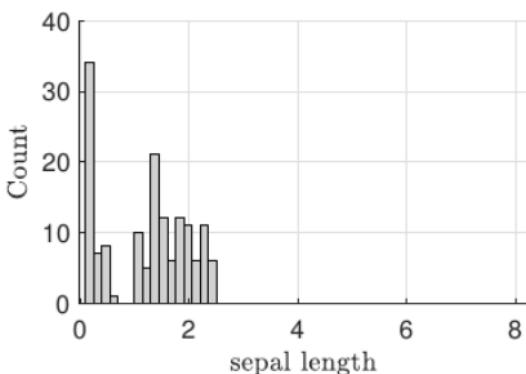
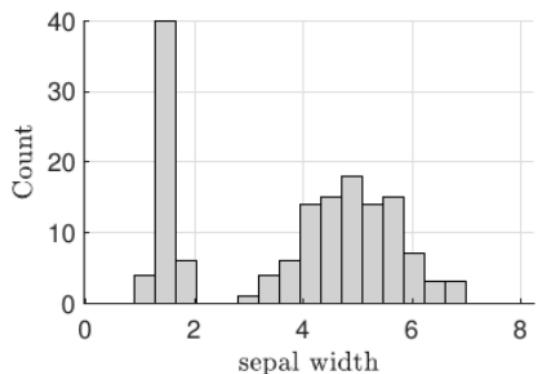
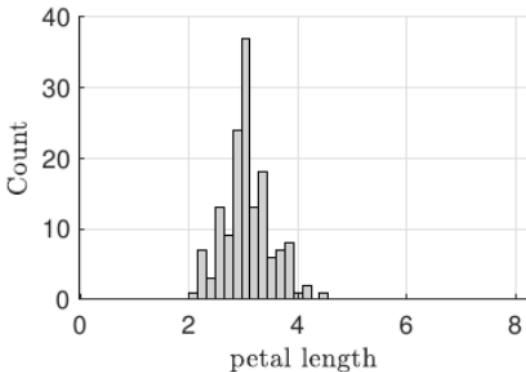
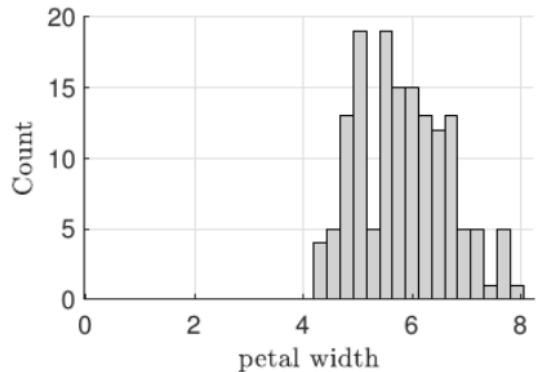
- Shows distribution of a single variable

- Divide the values into bins
- Bar plot of the number of values in bin
- Height indicates count of values
- Shape determined by
 - Distribution of data
 - Number of bins / bin width

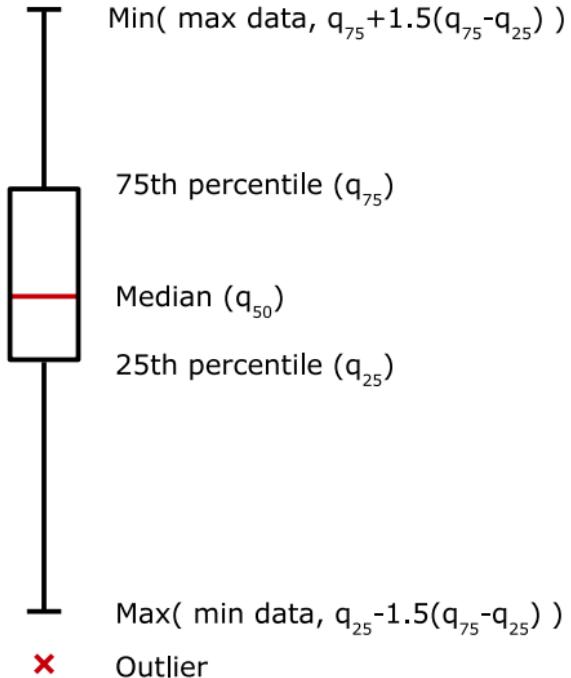


$$H = \{60, 64, 64, 66, 67, 69, 71, 72, 72, 72, 72, 73, 74, 74, 75, 75, 76, 76, 77, 77, 78, 78, 79, 80, 80, 81, 82, 84, 85, 85, 89\}$$

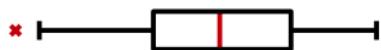
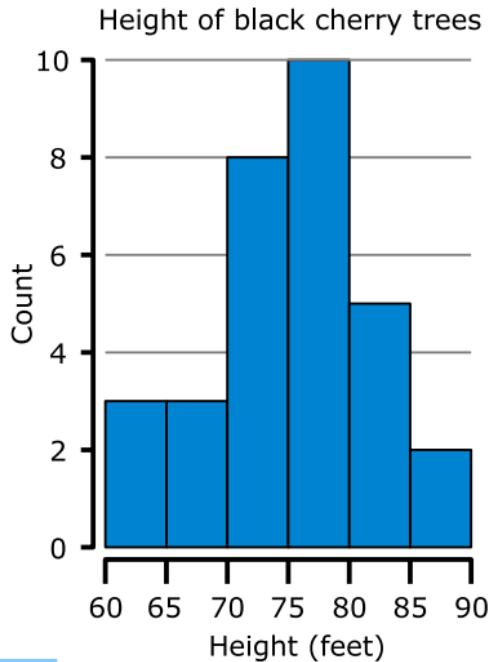
Histograms of the Iris data attributes



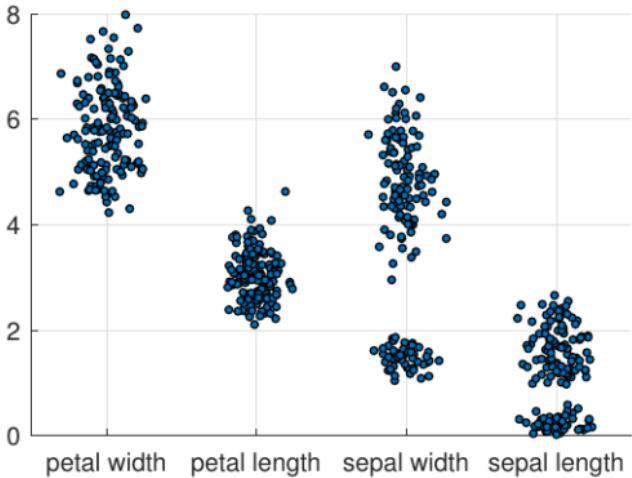
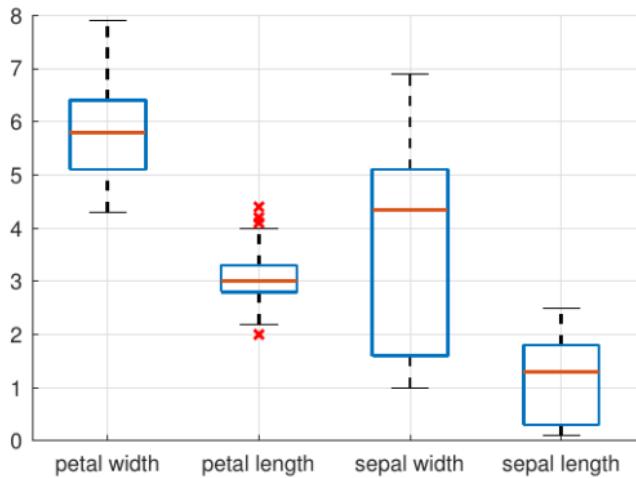
Box plots



The plotted whisker extends to the adjacent value, which is the most extreme data value that is not an outlier.



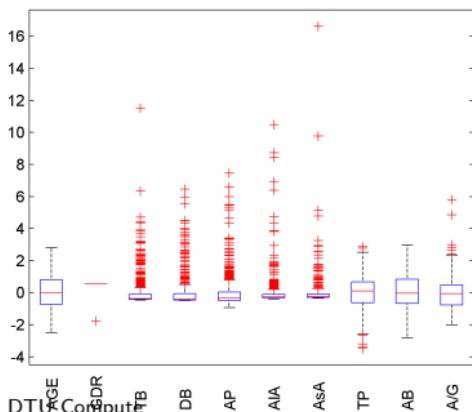
Box plots



Quiz 3, Boxplots (Fall 2012)

| No. | Attribute description | Abbrev. |
|----------|-------------------------------------|---------|
| x_1 | Age (in years) | AGE |
| x_2 | Gender (Female=0, Male=1) | GDR |
| x_3 | Total Bilirubin | TB |
| x_4 | Direct Bilirubin | DB |
| x_5 | Alkaline Phosphotase | AP |
| x_6 | Alamine Aminotransferase | AIA |
| x_7 | Aspartate Aminotransferase | AsA |
| x_8 | Total Proteins | TP |
| x_9 | Albumin | AB |
| x_{10} | Albumin to Globulin ratio | A/G |
| y | 0=No liver disease, 1=Liver disease | LD |

Table 1: Liver disease dataset.



The attributes $x_1 \dots x_{10}$ are standardized (i.e., the mean has been subtracted each attribute and the attributes divided by their standard deviations). The figure shows a boxplot for the standardized data. Which of the following statements is *correct*?

- A. The value of the 50th and 75th percentiles of the attribute DB coincides.
- B. Even though the distribution of AIA and AsA may have a similar shape this does not imply that the two attributes are correlated.
- C. The attribute TB is likely to be normal distributed.
- D. The attribute GDR has a clear outlier that should be removed.
- E. Don't know.

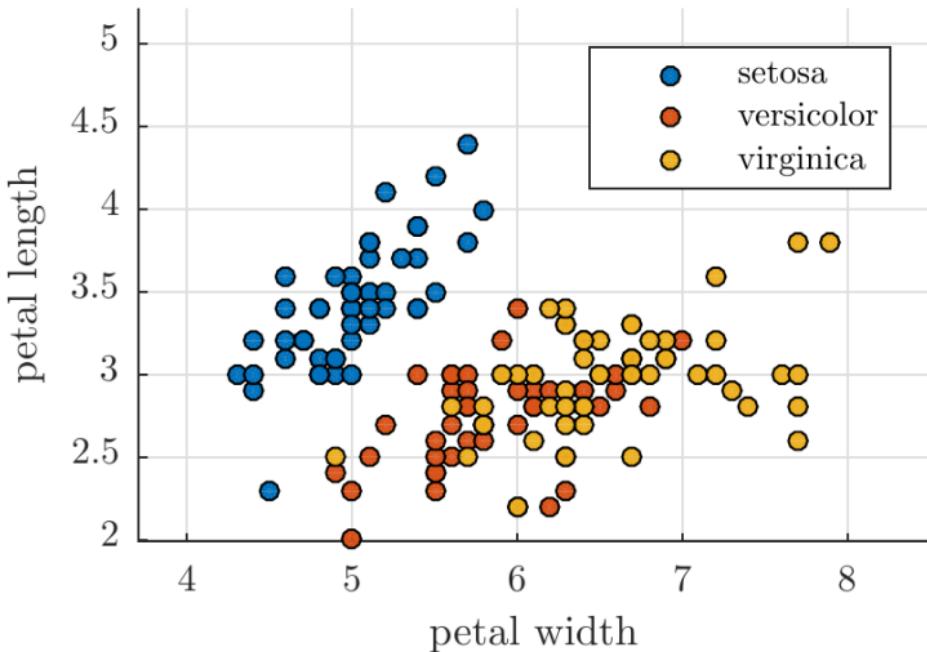
The 25th and 50th percentile but not the 50th and 75th percentiles of the attribute DB coincides. A1A and AsA will not necessarily be highly correlated even though their distributions may have a similar shape (hence, this is correct). For attributes to be correlated it is important they take on high or low values systematically, however, this can not be inspected in

a boxplot. TB is not likely to be normal distribution as this attribute does not have a symmetric but highly right skewed distribution. The attribute GDR does not have a clear outlier, in fact the outlier corresponds to the females in the dataset and all we can deduce from the plot is that more than 75 % of the observations are males.

Relation between attributes

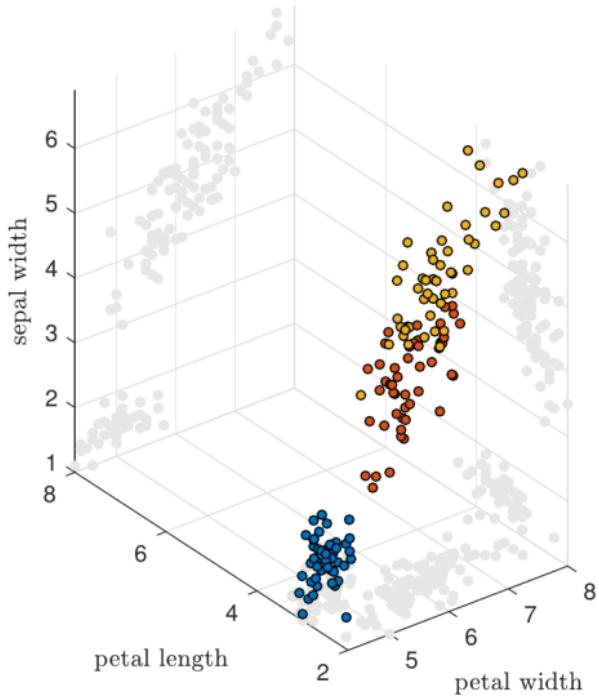
Scatter plots

- Shows **relation** between attributes
 - Assess dependence between attributes
 - Used with classes to assess separability



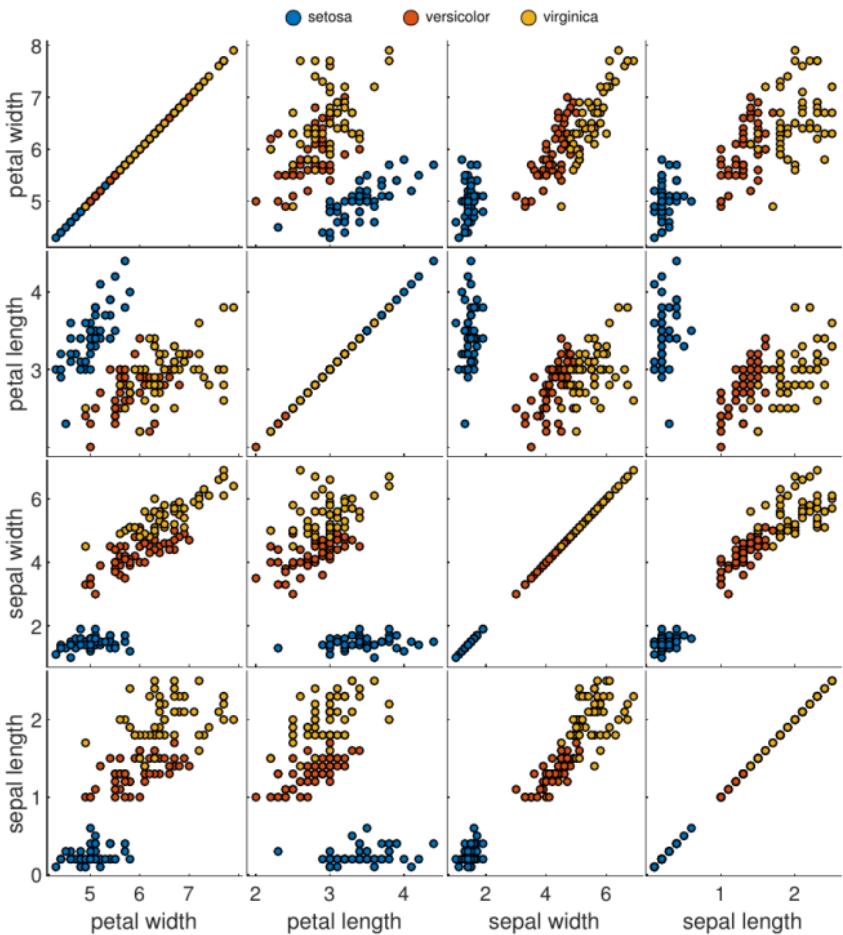
Scatter plots

- Shows **relation** between attributes
 - Assess dependence between attributes
 - Used with classes to assess separability
 - 3D plots are often confusing;
avoid if possible



Scatter plots

- Scatter plot matrix
 - All pairs of attributes



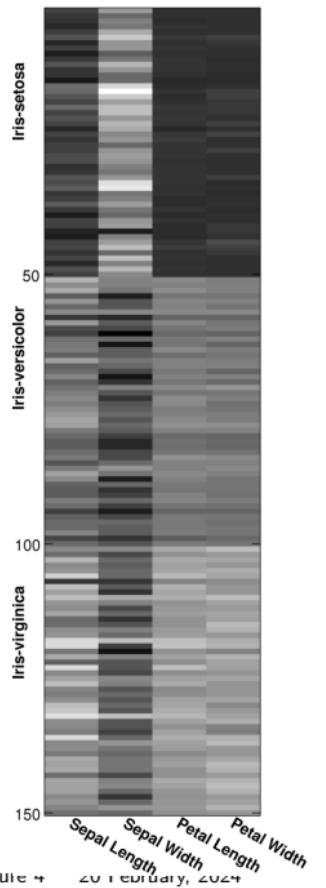
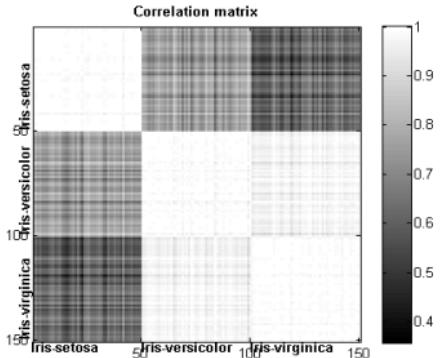
Matrix plots

- Plot of raw data matrix

- Useful when objects are sorted according to class
 - Typically, attributes are normalized

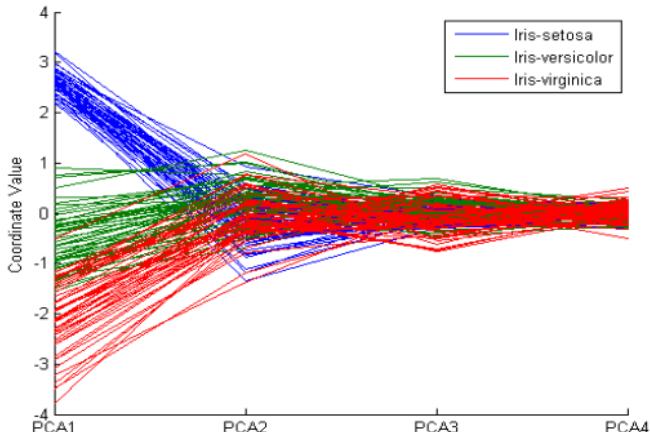
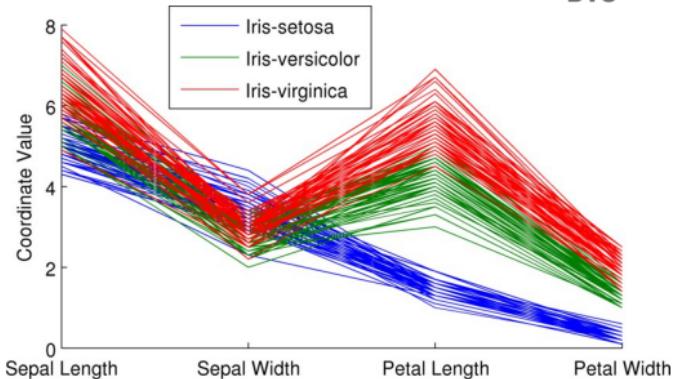
- Plots of similarity matrices

- Useful for visualizing the relation between objects



Parallel coordinates

- Plot high-dimensional data
- Instead of perpendicular axes
 - Use parallel axes
- Attribute values are plotted as a point
 - and the points are connected by a line
- Each object is represented as a line
- Lines representing a group of objects
 - Are similar in some sense
 - Ordering of attributes is important in seeing such groupings



ACCENT

- **Apprehension**

- Is it easy to see what is important in the graph?

- **Clarity**

- Are the most important elements visually most prominent?

- **Consistency**

- Have you used the same colors, shapes, etc. as in other graphs?

- **Efficiency**

- Does it convey its information in the most simple and efficient way?

- **Necessity**

- Are all elements of the graph necessary to represent data?

- **Truthfulness**

- Does the graph represent the data correctly?

Tufte's guidelines

- **Graphical excellence**

- Well-designed presentation of interesting data – a matter of
 - substance, statistics, and design
- Complex ideas communicated with
 - clarity, precision, and efficiency
- Gives the viewer
 - the greatest number of ideas
 - in the shortest time
 - with the least ink
 - in the smallest place.
- Nearly always multivariate
- Requires telling the truth about the data
- Maximise Data-ink ratio:

$$\text{Data-ink ratio} = \frac{\text{Data-ink}}{\text{Total ink used}}$$



Edward Tufte

https://commons.wikimedia.org/wiki/File:Edward_Tufte_-_cropped.jpg
Made available by Keegan Peterzell

Making good data visualizations is an art

For some interesting data visualizations see also

http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization.html

<http://www.informationisbeautiful.net/>

<http://www.junkcharts.typepad.com/>

Resources

<https://www.khanacademy.org> An excellent introduction to probability theory which we recommend as a go-to resource

(<https://www.khanacademy.org/math/statistics-probability/probability-library>)

<https://junkcharts.typepad.com> Excellent resource on creating good visualizations (https://junkcharts.typepad.com/junk_charts/)

<http://www2.imm.dtu.dk> Our demo of the multivariate normal distribution which illustrates the effect of the covariance matrix

(<http://www2.imm.dtu.dk/courses/02450/DemoNormal.html>)



02450: Introduction to Machine Learning and Data Mining

Measures of Similarity, summary statistics and probabilities

Bjørn Sand Jensen

DTU Compute, Technical University of Denmark (DTU)

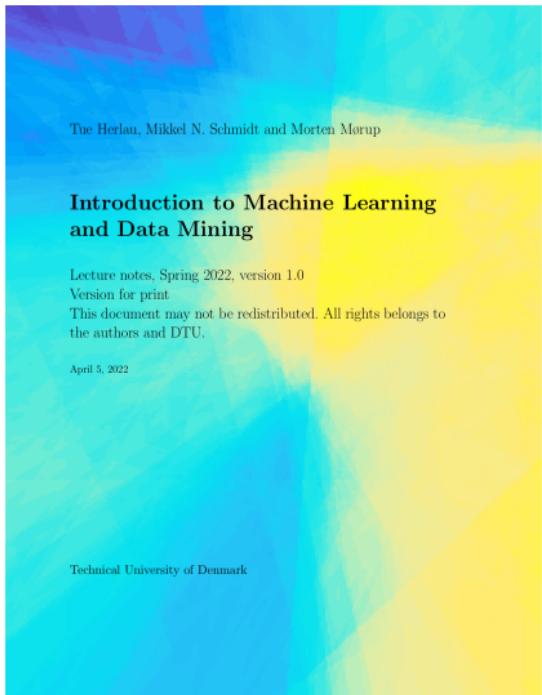
Today

Feedback Groups of the day:

Tobias Rodrigues Bjerre, Nana Bjerring-Kjærsgaard, Jonas Bloch, Louise Bober, Emma Sofie Bock, Emma Pihl Bødker, Albina Boran, Nicholas Borch, Carl Borg, Monika Karolina Borkowska, Niels Moll Bown, Filip Sebastian Boye, Robin Pearson Braagaard, Francisca Miriam Breyer, Mikkel Nielsen Broch-Lips, Julie Bruun Brockhoff, Clara Louise Brodt, Jon Brygge, Samer Ernesto Bujana Escalona, Jan Bures, Vaneeza Fatima Butt, Jeppe Urup Byberg, Mario Caliò, Gabriel Alejandro Camps Farrujia, Valentin Carboniero, William Allerup Carlsen, Sofus Alexander Kjelgaard Carstens, Diogo Miguel Lopes Carvalho, Kaj Richard Caspersen, Sergio Catalá Rodríguez, Benjamin CHAMBAUDET, Pauline Charpentier, Olivia Wenyu Chen, Yuechen Chen, Xi Chen, Khadija Chgoura, Jiwon Choi, Mobashshira Chowdhury, Freja Damgaard Christensen, Mattis Bo Clade Christensen, Johanne Birk Christensen, Lise Bjerre Christiansen, Tobias Alexander Christiansen, Xiaozhuo Chu

Reading material:

Chapter 4, Chapter 5



Lecture Schedule

1 Introduction

30 January: C1

Data: Feature extraction, and visualization

2 Data, feature extraction and PCA

6 February: C2, C3

3 Measures of Similarity, summary statistics and probabilities

13 February: C4, C5

4 Probability densities and data visualization

20 February: C6, C7

Supervised learning: Classification and regression

5 Decision trees and linear regression

27 February: C8, C9 (Project 1 due 29 February at 17:00)

6 Overfitting, cross-validation and Nearest Neighbor

5 March: C10, C12

7 Performance evaluation, Bayes, and Naive Bayes

12 March: C11, C13

8 Artificial Neural Networks and Bias/Variance

19 March: C14, C15

9 AUC and ensemble methods

2 April: C16, C17

Unsupervised learning: Clustering and density estimation

10 K-means and hierarchical clustering

9 April: C18 (Project 2 due 11 April at 17:00)

11 Mixture models and density estimation

16 April: C19, C20

12 Association mining

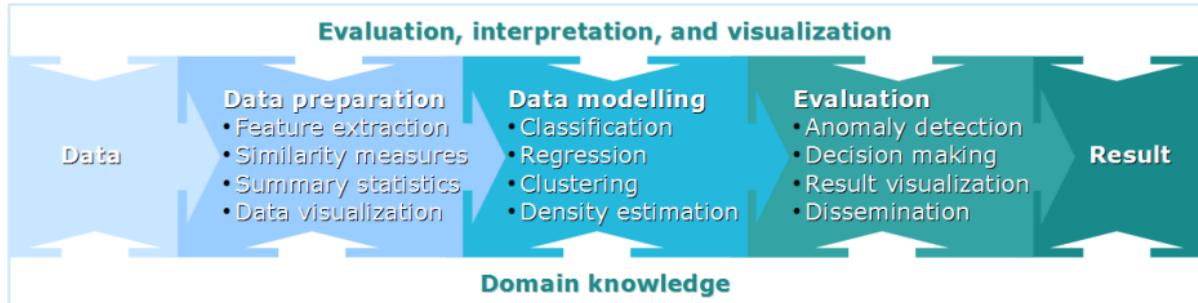
23 April: C21

Recap

13 Recap and discussion of the exam

30 April: C1-C21

Online 24/7 help: Discussion Forum/Piazza
Streaming: Zoom (see link on DTU Learn)
Recordings: <https://panopto.dtu.dk/>
Online exercises: MS Teams



Learning Objectives

- Compute measures of similarity/dissimilarity (L_p distance, cosine distance, etc.)
- Understand basics of probability theory and stochastic variables in the discrete setting
- Understand probabilistic concepts such as expectations, independence and the Bernoulli distribution
- Understand the maximum likelihood principle for repeated binary events

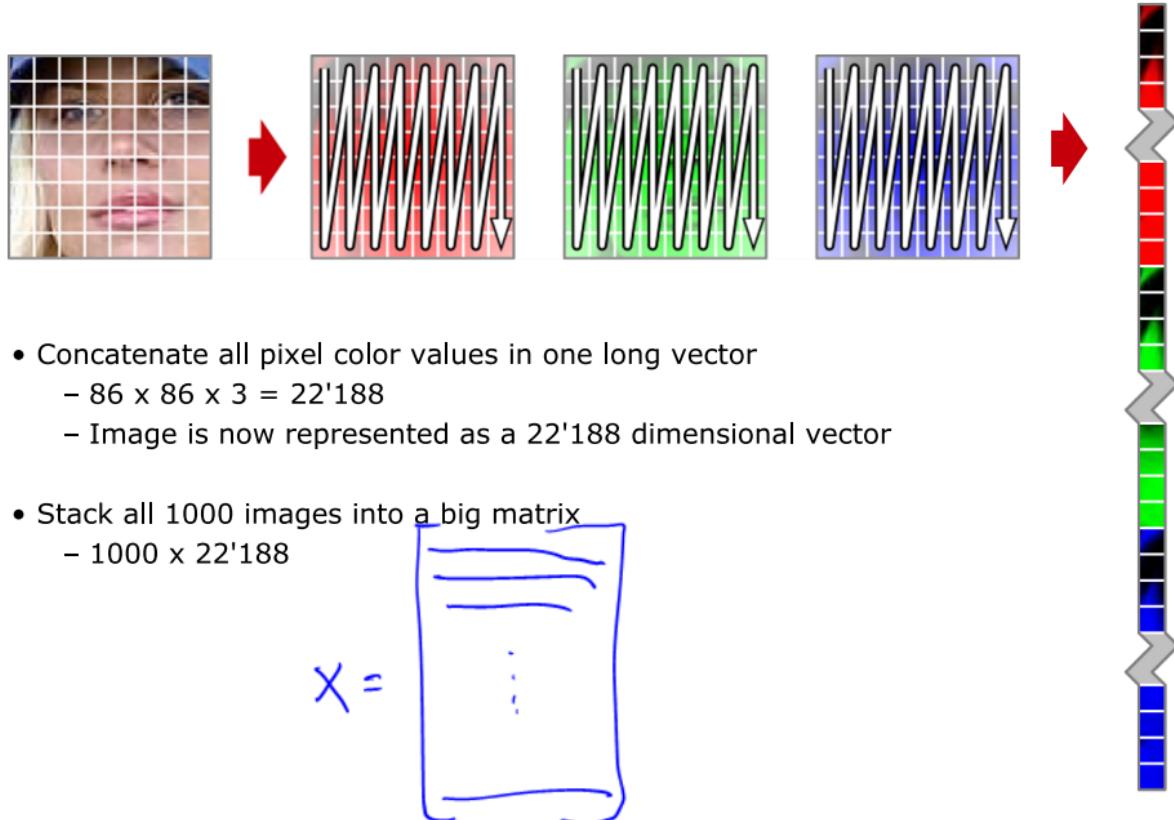
PCA recap: Principal component analysis on images



- 1000 images, 86 x 86 pixels, 3 RGB intensities

Tamara Berg "Faces in the wild"

Pre-processing



Principal component analysis (PCA)

$$\tilde{X}^T \tilde{X} v = \lambda v \quad \text{s.t. } v^T v = 1$$

1. Subtract the mean $\Rightarrow D \tilde{X}$

- Consider dividing with variance; use 1-out-of-K coding for nominal attributes

2. Compute the singular value decomposition (SVD)

- Orthogonal linear transformation
- Transforms data to a new coordinate system
 - Greatest variance along the first axis (first column of V)
 - Second greatest variance along the second axis

$$\tilde{X} = U \Sigma V^T$$

Dimensions:
 \tilde{X} : $N \times M$ U : $N \times N$ (Orthonormal) Σ : $N \times M$ (Diagonal) V^T : $M \times M$ (Orthonormal)

$$\frac{\sigma_1^2}{\sum_j \sigma_j^2} = \text{exp. var by } V_1$$

$$V = \begin{bmatrix} | & | & \dots & | \\ v_1 & v_2 & \dots & v_m \\ | & | & \dots & | \end{bmatrix}$$

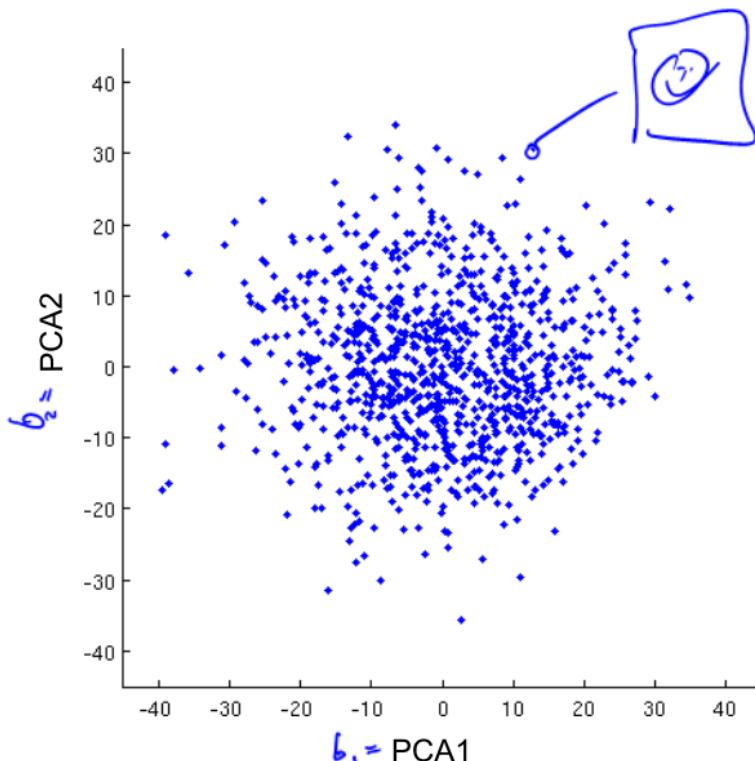
$$\Sigma = \begin{bmatrix} \sigma_1 & & & & 0 \\ & \sigma_2 & & & 0 \\ & & \ddots & & 0 \\ 0 & & & \ddots & 0 \\ 0 & & & & \sigma_m \end{bmatrix}$$

• Plot data in the transformed coordinate system

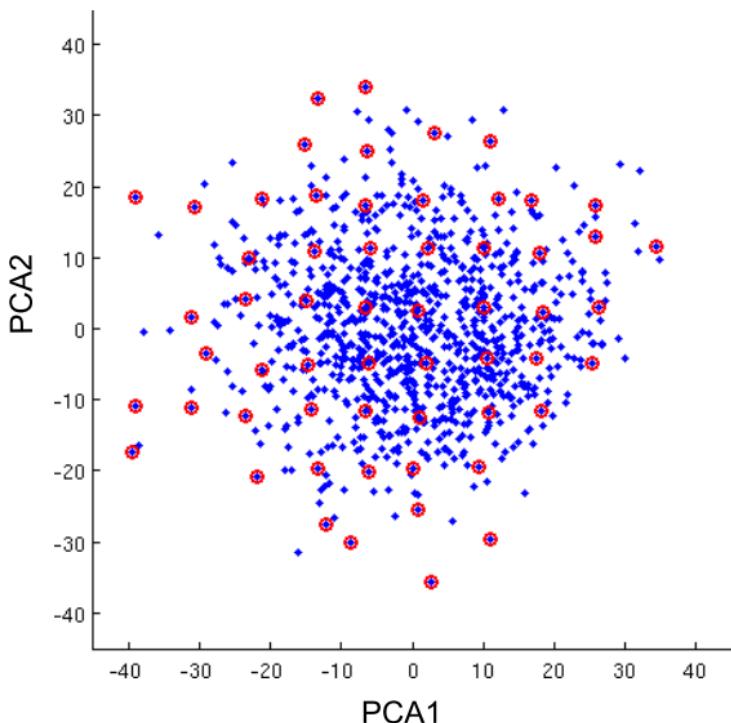
- Corresponds to looking at data from an angle where it is most spread out

PCA on face images

$$V_k = \begin{bmatrix} v_1, v_2 \end{bmatrix}$$
$$b^T = V_k^T \tilde{x}$$
$$= [b_1, b_2]$$



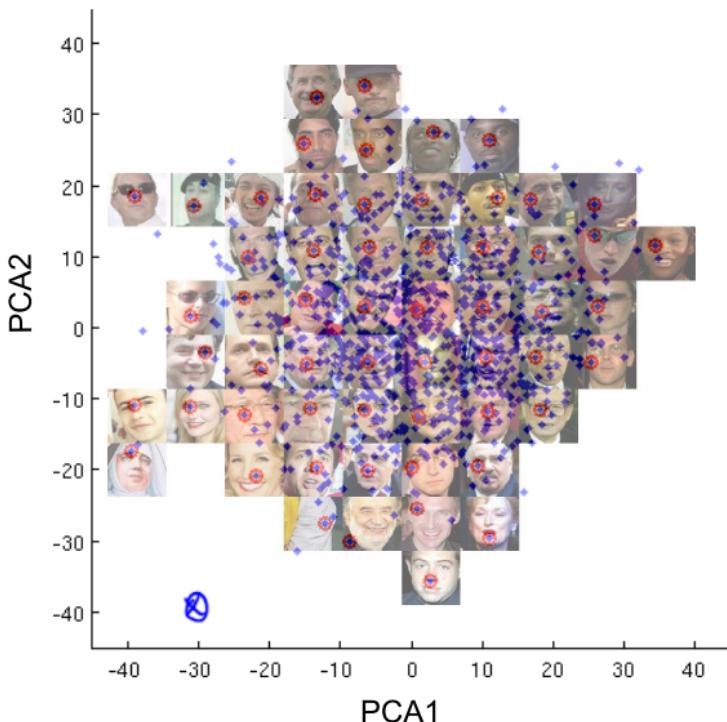
PCA on face images



PCA on face images

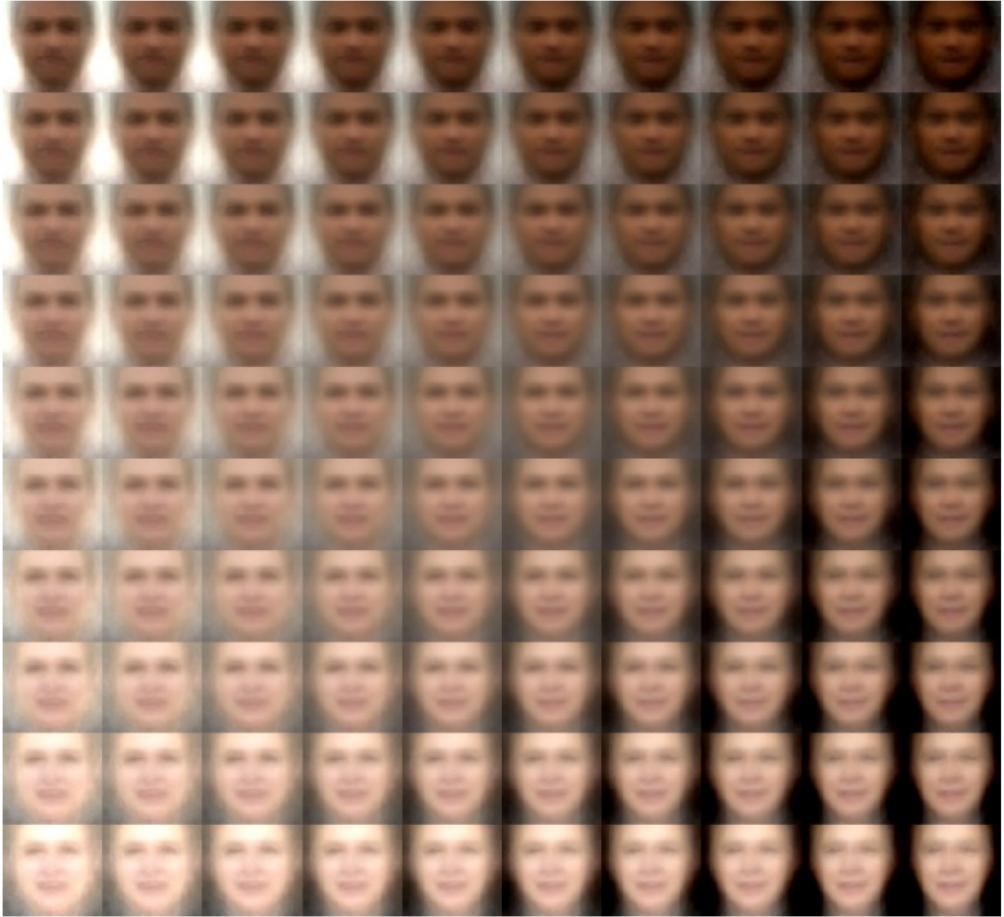
$$\underline{b} = \begin{bmatrix} -30 \\ -40 \end{bmatrix}$$

$$\underline{x}' = \underline{V}_k^T \underline{b} + \underline{\mu}$$



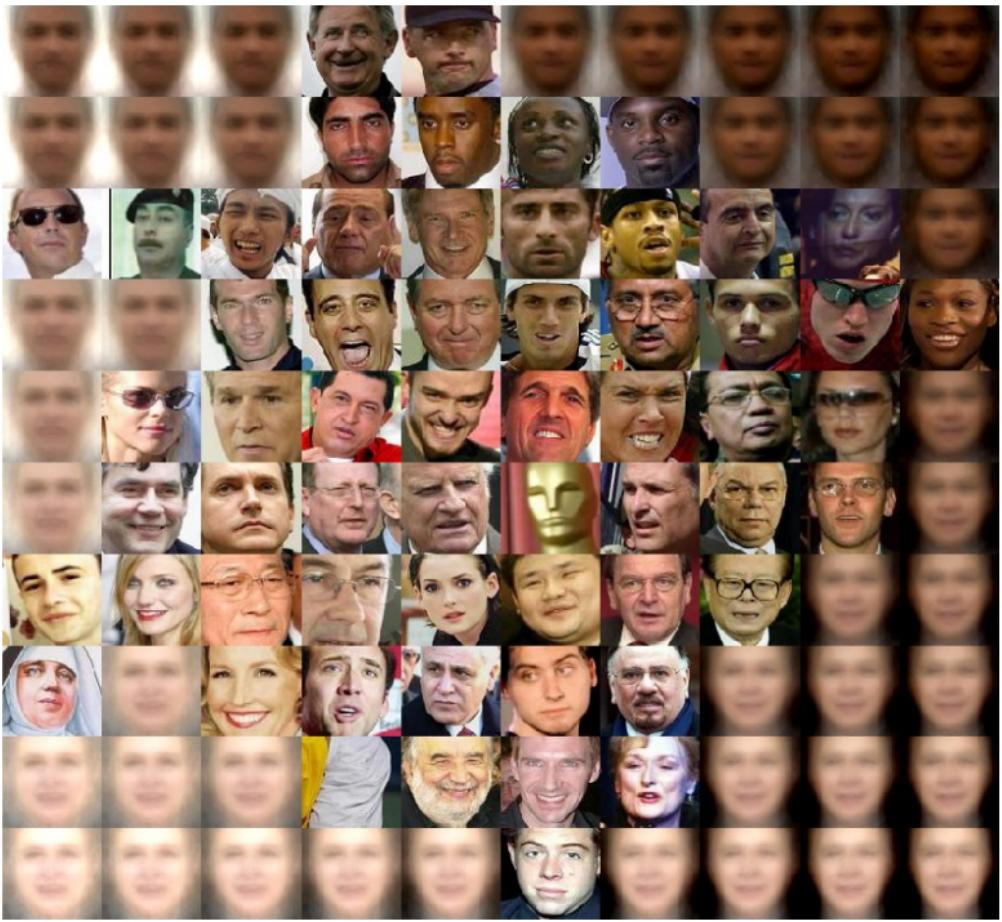


- What information do the two principal axes capture?

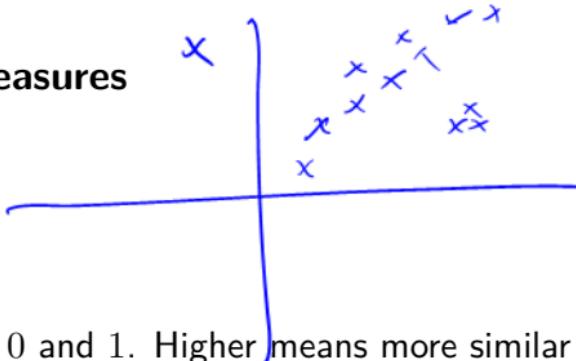




What information do the two principal axes capture?



Similarity / Dissimilarity measures



Similarity $s(x, y)$ Often between 0 and 1. Higher means more similar

Dissimilarity $d(x, y)$ Greater than 0. Lower means more similar.

Classification Classify a document as having the same topic y as the document it is **most similar/least dissimilar** to.

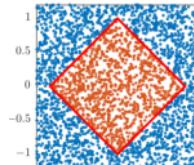
Outlier detection The observation most **dissimilar** to all other observations is an outlier



Dissimilarity measures

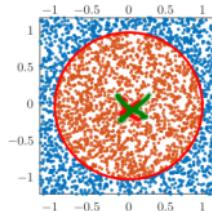
- General Minkowsky distance (p -distance) $d_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{j=1}^M |x_j - y_j|^p \right)^{\frac{1}{p}}$
- One-norm ($p = 1$)

$$d_1(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^M |x_j - y_j|$$



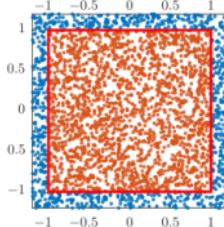
- Euclidean ($p = 2$)

$$d_2(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^M (x_j - y_j)^2}$$



- Max-norm distance ($p = \infty$)

$$d_\infty(\mathbf{x}, \mathbf{y}) = \max \{|x_1 - y_1|, |x_2 - y_2|, \dots, |x_M - y_M|\}$$



Usage: Regularization and alternative optimization targets. For instance, $p = \infty$ is very affected by outliers, $p = 1$ much less so.

Similarity measures

$$x = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

K : Total number of attributes

f_{00} : Number of attributes where $x_k=y_k=0$

f_{11} : Number of attributes where $x_k=y_k=1$

Simple Matching Coefficient (SMC)

$$\begin{array}{ll} f_{01} & x_k=0, y_k=1 \\ f_{10} & x_k=1, y_k=0 \end{array}$$

$$SMC(x, y) = \frac{f_{00} + f_{11}}{K}$$

$$K = f_{00} + f_{01} + f_{10} + f_{11}$$

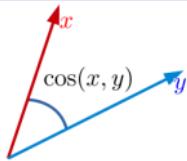
+ Symmetric

+ Positive matches

Jaccard Coefficient

$$J(x, y) = \frac{f_{11}}{K - f_{00}}$$

Cosine similarity



$$\cos(x, y) = \frac{x^\top y}{\|x\|_2 \|y\|_2}$$

+ Positive matches
+ Document length

Extended Jaccard coefficient

$$EJ(x, y) = \frac{x^\top y}{\|x\|^2 + \|y\|^2 - x^\top y}$$

Also defined for continuous data

Quiz 1, similarity measures

Calculate the simple matching coefficient, Jaccard, cosine and extended jaccard similarity between customer 1 and customer 2 in the market basket data below.

| ID | Bread | Soda | Milk | Beer | Diaper |
|----|-------|------|------|------|--------|
| 1 | 1 | 1 | 1 | 0 | 0 |
| 2 | 0 | 1 | 1 | 0 | 1 |

K : Total number of attributes

f_{00} : Number of attributes where $x_k = y_k = 0$

f_{11} : Number of attributes where $x_k = y_k = 1$

Which of the following statements are true?

- A. $\text{SMC}(o_1, o_2) = \frac{3}{5}$, $J(o_1, o_2) = \frac{1}{2}$, $\cos(o_1, o_2) = \frac{2}{3}$,
- B. $\text{SMC}(o_1, o_2) = \frac{3}{5}$, $J(o_1, o_2) = \frac{3}{4}$, $\cos(o_1, o_2) = \sqrt{\frac{2}{3}}$,
- C. $\text{SMC}(o_1, o_2) = \frac{2}{5}$, $J(o_1, o_2) = \frac{1}{3}$, $\cos(o_1, o_2) = \frac{2}{3}$,
- D. $\text{SMC}(o_1, o_2) = \frac{2}{5}$, $J(o_1, o_2) = \frac{1}{3}$, $\cos(o_1, o_2) = \sqrt{\frac{2}{3}}$,
- E. Don't know.

$$\text{SMC}(x, y) = \frac{f_{00} + f_{11}}{K}$$

$$J(x, y) = \frac{f_{11}}{K - f_{00}}$$

$$\cos(x, y) = \frac{x^\top y}{\|x\|_2 \|y\|_2}$$

$$\text{EJ}(x, y) = \frac{x^\top y}{\|x\|_2^2 + \|y\|_2^2 - x^\top y}$$

$$SMC(x,y) = \frac{1+2}{5} = \frac{3}{5}$$

$$J(x,y) = \frac{2}{5-1} = \frac{2}{4} = \frac{1}{2}$$

$$\cos(x,y) = \frac{2}{\sqrt{3}\sqrt{3}} = \frac{2}{3}$$

$$\begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\sum_{i=1}^5 o_i = f_1$$

$$\|x\|_2 = \sqrt{\sum_{i=1}^5 x_i^2} = \sqrt{3}$$

$$\|y\|_2 = \sqrt{3}$$

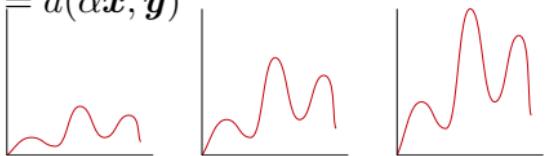
The problem is easily solved by using the inserted formula. We obtain: $SMC(o_1, o_2) = \frac{3}{5}$, $J(o_1, o_2) = \frac{1}{2}$, $\cos(o_1, o_2) = \frac{2}{3}$ and therefore the A is true. Since the

data is binary, the extended Jaccard and the jaccard coefficient agree.

Invariance

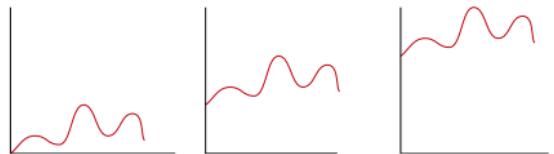
Scale invariance

$$d(\mathbf{x}, \mathbf{y}) = d(\alpha \mathbf{x}, \mathbf{y})$$



Translation invariance

$$d(\mathbf{x}, \mathbf{y}) = d(\alpha + \mathbf{x}, \mathbf{y})$$



General invariances

$$d(\boxed{6}, \boxed{2}) = d(\boxed{6}, \boxed{2})$$

| # | age | \$ |
|---|-----|-----|
| 1 | 40 | 50k |
| 1 | 30 | 50k |
| 1 | 40 | 49k |

$$\delta(x_1, x_2) = 10$$

$$\delta(x_1, x_3) = 1000$$

Transformations

Standardization: Ensure a single attribute will not dominate:

$$\tilde{x}_{ik} = \frac{x_{ik} - \hat{\mu}_k}{\hat{\sigma}_k}$$

Example:

- **Number of children** ~ 0-5
- **Age** ~ 0-100 years
- **Annual income** ~ 0-50.000 €

Combining heterogeneous attributes Transform measures and combine

$$s_{\text{Edu.}} = \text{SMC}(x_{\text{Edu.}}, y_{\text{Edu.}})$$

$$s_{\text{Age.}} = a (a + d_1(x_{\text{Age.}}, y_{\text{Age.}}))^{-1}, \quad a = 1$$

$$s(x, y) = \frac{1}{2} (s_{\text{Edu.}} + s_{\text{Age.}})$$

Example:

- **Age:** Continuous
- **Education:** Binary
 - Primary (yes/no)
 - Secondary (yes/no)
 - Tertiary (yes/no)

Weighting Attributes have different importance

$$s(x, y) = 0.99s_{\text{Edu.}} + 0.01s_{\text{Age.}}$$

Empirical statistics

Break: 14:10

Given two samples $x_1, x_2, \dots, x_N \in \mathbb{R}$ and $y_1, y_2, \dots, y_N \in \mathbb{R}$:

- Empirical mean

$$\hat{\mu} = \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

- Empirical variance

$$\hat{s} = \text{vár}[x] = \frac{1}{N-1} \sum_{i=1}^N (x_i - \hat{\mu})^2$$

- Empirical covariance

$$\text{côv}[x, y] = \frac{1}{N-1} \sum_{i=1}^N (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)$$

- Empirical standard deviation

$$\hat{\sigma} = \text{std}[x] = \sqrt{\hat{s}}$$

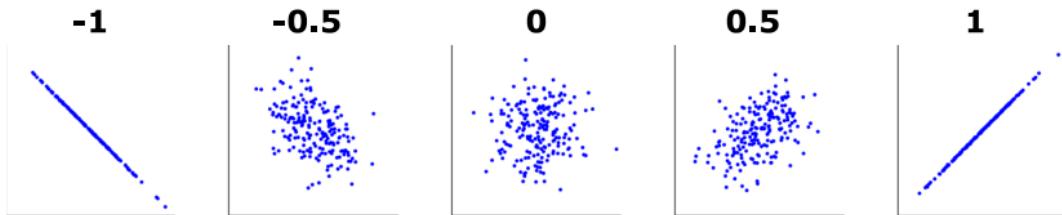
Correlation

- Measure of degree of linear relationship

$$\text{corr}[x, y] = \frac{\text{cov}[x, y]}{\hat{\sigma}_x \hat{\sigma}_y}$$

- A correlation of **1** or **-1** means there is a perfect linear relation

$$x_k = ay_k + b$$



Quantiles

Given N observations of an attribute $x_1, x_2, \dots, x_N \in \mathbb{R}$.

Quantiles describe the *points* that divide the underlying distribution into intervals that are equally probable:

- The one 2-quantile (**median**) divides the distribution in two intervals.
- The three 4-quantiles (**quartiles**) divides the distribution in four intervals.
- The 99 100-quantiles (**percentiles**) divides the distribution in 100 intervals.

The **median** is the same as the 2nd quartile or the 50th percentile.

E.g., we can (approximately) find the **median** by

- Sort the observations in ascending order $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(N)}$

$$\text{median}[x] = \begin{cases} x_{\left(\frac{N+1}{2}\right)} & \text{if } N \text{ is odd} \\ \frac{1}{2} \left(x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)} \right) & \text{if } N \text{ is even.} \end{cases}$$

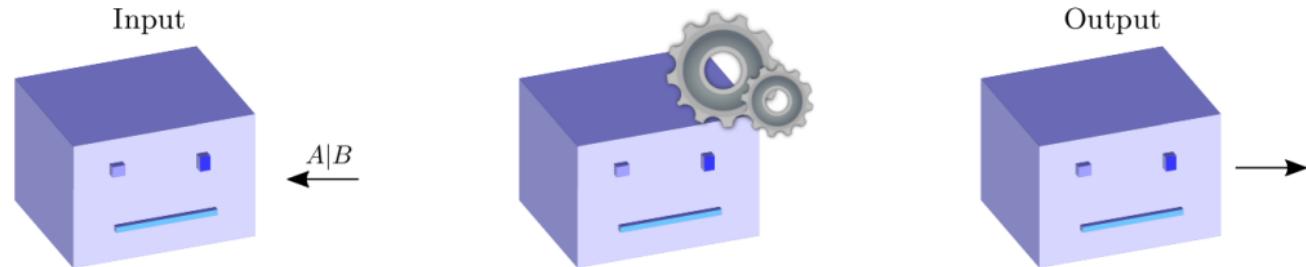
Probabilities

Pragmatically: A big part of AI is dealing with **uncertainty** and **incomplete information**. Probabilities is the formal framework for doing so.

Algorithmically: **If** an image belongs to a particular category is a discrete event. The **probability** it belongs to a category is continuous. Algorithmically, easier to optimize continuous quantities.

Convenience: There are boiler-plate ideas for transforming a **probabilistic** assumption into an algorithm (maximum likelihood).

Probabilities



Assuming B is true, how plausible is it A is true?

Assuming B , the plausibility of A is
(low / medium / high / certain)

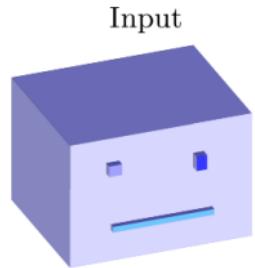
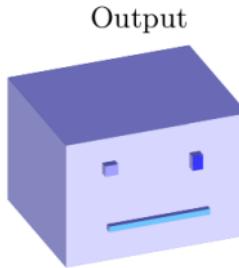
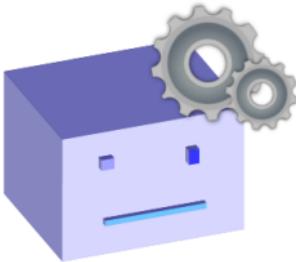
We reason about a proposition A in light of evidence B :

$$P(A|B) = x$$

The degree-of-belief that A is true given B is accepted as true is at a level x

- A number between 0 and 1
- A and B are always binary (true/false) propositions
- Represents a *state of knowledge*

Probabilities: Trial example

 $A|B$ 

Assuming B is true, how plausible is it A is true?

Assuming B , the plausibility of A is
(low / medium / high / certain)

G : *The accused is guilty*

E_1 : *A car similar to his was seen at the crime scene.*

E_2 : *A large sum of money was found in his posession*

E_3 : *His fingerprints was found at the door of the bank.*

Probabilities express states-of-knowledge

$$E \equiv E_1 \text{ and } E_2 \text{ and } E_3$$

$$P(G|E) > P(G|E_2)$$

Binary propositions

A binary proposition is a statement which is either true or false (we might not know, but someone with complete knowledge would)

A : *In 49 BCE, Caesar crossed the Rubicon*

B : *Acceleration sensor 39 measures more than 0.85*

C : *Patient 901 has high cholesterol*

Propositions can be combined with **and**, **or** and **not**:

$AB \equiv$ True if A and B are both true

$A + B \equiv$ True if either A or B are true

$\bar{A} \equiv$ True if A is false

We define two special propositions which is always **true/false**:

1 : *A proposition which is always true*

0 : *A proposition which is always false*

...and the following identities: $A1 = A$, $A + \bar{A} = 1$, $\bar{\bar{A}} = A$ and

$$A(B_1 + B_2 + \cdots + B_n) = AB_1 + AB_2 + \cdots + AB_n$$

Quiz 2, Probabilities

Assume we define the following 4 boolean variables.

R_1 : Handed in report 1

R_2 : Handed in report 2

R_3 : Handed in report 3

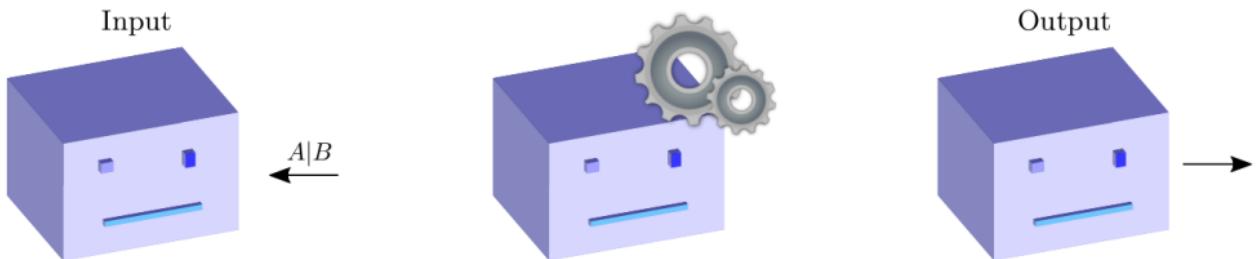
F : Student failed 02450



How would you express the probability of the statement:

If a student hand in report 1, 2 and 3, the chance of passing 02450 is greater than 90%?

- A. $P(R_1 R_2 R_3 | F) > 0.9$
- B. $P(\bar{F} | R_1 + R_2 + R_3) > 0.9$
- C. $P(\bar{F} | R_1 R_2 R_3) > 0.9$
- D. $P(R_1 + R_2 + R_3 | F) > 0.9$
- E. Don't know.



Assuming B is true, how plausible is it A is true?

Assuming B , the plausibility of A is (low / medium / high / certain)

Passing can be defined as not failing. Therefore, express the statement as:

$$P(\overline{F}|R_1R_2R_3) > 0.9$$

namely, if it is true report 1, 2 and 3 are all handed in, what is the chance of passing?

Rules of probability

The sum rule: $P(A|C) + P(\bar{A}|C) = 1$

The product rule: $P(AB|C) = P(B|AC)P(A|C)$

$$= p(A|Bc)p(B|C)$$



Interpretation:

$P(A|B) = 0$ (interpretation: given B is true, A is certainly false)

$P(A|B) = 1$ (interpretation: given B is true, A is certainly true)

We also use the shorthand:

$$P(A|1) = P(A)$$

$$p(A) + P(\bar{A}) = 1$$

$$p(AB) = P(A|B)P(B)$$

Remarkably, this is the mathematical basis for this course

Marginalization and Bayes' theorem

Sum rule $P(A|C) + P(\bar{A}|C) = 1$

Product rule $P(AB|C) = P(B|AC)P(A|C)$

$$= P(A|BC)P(B|C)$$

$$\begin{aligned} P(B|C) &= P(B|C) \left[P(A|BC) + P(\bar{A}|BC) \right] = P(AB|C) + P(\bar{A}B|C) \\ &= P(B|AC)P(A|C) + P(B|\bar{A}C)P(\bar{A}|C). \end{aligned}$$

$$\begin{aligned} P(A|BC) &= \frac{P(B|AC)P(A|C)}{P(B|C)} \\ &= \frac{P(B|AC)P(A|C)}{P(B|AC)P(A|C) + P(B|\bar{A}C)P(\bar{A}|C)}. \end{aligned}$$

DNA



Bayes theorem

$$P(A|BC) = \frac{P(B|AC)P(A|C)}{P(B|AC)P(A|C) + P(B|\bar{A}C)P(\bar{A}|C)}$$

Crimes may be solved by matching crime-scene DNA to DNA in a database

- If the two samples are from the same person, a DNA test will always give a positive match $P(D|G) = 1$

- If the DNA are from different persons, DNA will incorrectly give a positive match one time out of a million $P(D|\bar{G}) = 1e-6$

A crime is committed in Racoon City by an unidentified male. Assume all 8000 possible perpetrators undergo a DNA test, and suppose the DNA test gives a positive result for George. What is the chance George is guilty?



G : George is guilty, D : There was a positive DNA match

$$P(G|D) = \frac{P(D|G)p(G)}{P(D|G)p(G) + P(D|\bar{G})p(\bar{G})} = \frac{1 \times \frac{1}{8000}}{1 \times \frac{1}{8000} + 1e-6(1 - \frac{1}{8000})} \approx 99\%$$

Solution:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$\begin{aligned} P(G|D) &= \frac{P(D|G)P(G)}{P(D|G)P(G) + P(D|\bar{G})P(\bar{G})} \\ &= \frac{1 \times \frac{1}{8000}}{1 \times \frac{1}{8000} + 10^{-6} \times \left(1 - \frac{1}{8000}\right)} \\ &= 1 - \frac{1}{126} \approx 99\% \end{aligned}$$

Exclusive and exhaustive events

A_1 : The side \square face up.

A_2 : The side \square face up.

A_3 : The side \bullet face up.

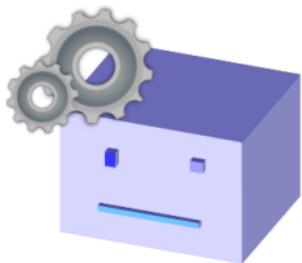
A_4 : The side \bullet face up.

A_5 : The side \circ face up.

A_6 : The side \circ face up.

- When no two propositions can be true at the same time, they are said to be **mutually exclusive**: $A_i A_j = 0$ for $i \neq j$
- Consider any two events A and B

$$P(A + B) = P(A) + P(B) - P(AB)$$



- In general, for n mutually exclusive events

$$P(A_1 + A_2 + \cdots + A_n) = \sum_{i=1}^n P(A_i)$$

- A set of events is **exhaustive** if one has to be true: $A_1 + \cdots + A_n = 1$. Then:

$$\sum_{i=1}^n P(A_i) = P(A_1 + A_2 + \cdots + A_n) = 1$$

Exclusive and exhaustive events

A_1 : The side \square face up.

A_2 : The side \circlearrowleft face up.

A_3 : The side \circlearrowright face up.

A_4 : The side \blacksquare face up.

A_5 : The side $\blacksquare\blacksquare$ face up.

A_6 : The side $\blacksquare\blacksquare\blacksquare$ face up.

- When no two propositions can be true at the same time, they are said to be **mutually exclusive**: $A_i A_j = 0$ for $i \neq j$
- Consider any two events A and B

$$\begin{aligned} P(A + B) &= 1 - P(\overline{A} \ \overline{B}) &= 1 - P(\overline{A}|\overline{B})P(\overline{B}) \\ &= 1 - [1 - P(A|\overline{B})] P(\overline{B}) &= P(B) + P(A\overline{B}) \\ &= P(B) + P(\overline{B}|A)P(A) &= P(B) + [1 - P(B|A)] P(A) \\ &= P(A) + P(B) - P(AB). \end{aligned}$$

- In general, for n mutually exclusive events $P(A_1 + A_2 + \dots + A_n) = \sum_{i=1}^n P(A_i)$
- A set of events is **exhaustive** if one has to be true: $A_1 + \dots + A_n = 1$. Then:

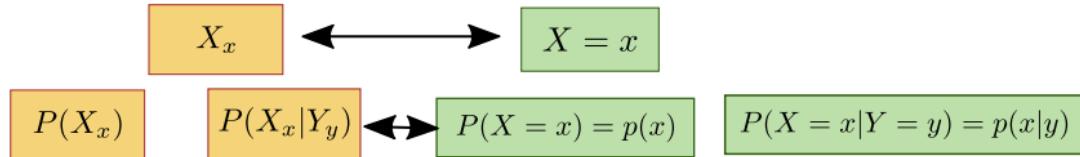
$$\sum_{i=1}^n P(A_i) = P(A_1 + A_2 + \dots + A_n) = 1$$

Stochastic variables

- Often, we will measure numerical quantities (number of children, age of a patient, etc.)
- Suppose a quantity X (number of children) takes a value $x = 3$. We can write this as the binary event X_3 and in general:

$X_x : \{\text{The binary event that } X \text{ is equal to the number } x\}$

- Stochastic variable simplify this notation by the definition:



Sum rule $P(A|C) + P(\bar{A}|C) = 1$

Product rule $P(AB|C) = P(B|AC)P(A|C)$

Marginalization

$$P(A|C) = P(A|BC)P(B|C) + P(A|\bar{B}C)P(\bar{B}|C)$$

Bayes theorem

$$P(A|BC) = \frac{P(B|AC)P(A|C)}{P(B|AC)P(A|C) + P(B|\bar{A}C)P(\bar{A}|C)}$$

Sum rule $\sum_i P(x_i|z_k) = 1$

Product rule $p(x_i, y_j|z_k) = p(x_i|y_j, z_k)p(y_j|z_k)$

Marginalization $p(x_i|z_k) = \sum_j p(x_i|y_j, z_k)p(y_j|z_k)$

Bayes theorem $p(y_j|x_i, z_k) = \frac{p(x_i|y_j, z_k)p(y_j|z_k)}{\sum_j p(x_i|y_j, z_k)p(y_j|z_k)}$

Quiz 3, Avila bible (Fall 2018)

| $p(\tilde{x}_2, \tilde{x}_{10} y)$ | $y = 1$ | $y = 2$ | $y = 3$ |
|---------------------------------------|---------|---------|---------|
| $\tilde{x}_2 = 0, \tilde{x}_{10} = 0$ | 0.19 | 0.3 | 0.19 |
| $\tilde{x}_2 = 0, \tilde{x}_{10} = 1$ | 0.22 | 0.3 | 0.26 |
| $\tilde{x}_2 = 1, \tilde{x}_{10} = 0$ | 0.25 | 0.2 | 0.35 |
| $\tilde{x}_2 = 1, \tilde{x}_{10} = 1$ | 0.34 | 0.2 | 0.2 |

Table 1: Probability of observing particular values of \tilde{x}_2 and \tilde{x}_{10} conditional on y .

We will consider a dataset based on the Avila bible. We wish to predict the copyist ($y = 1, 2, 3$) of a bible based on the two typographic attributes *upperm* and *mr/is*. We suppose the attributes have been binarized such that *upperm* corresponds to $\tilde{x}_2 = 0, 1$ and *mr/is* to $\tilde{x}_{10} = 0, 1$. Suppose the probability for each of the configurations of \tilde{x}_2 and \tilde{x}_{10} conditional on the copyist y are as given in Table 1. and the prior probability of

the copyists is

$$p(y=1) = 0.316, p(y=2) = 0.356, p(y=3) = 0.328$$

Using this, what is then the probability an observation was authored by copyist 1 given that $\tilde{x}_2 = 1$ and $\tilde{x}_{10} = 0$?

- A. $p(y=1|\tilde{x}_2=1, \tilde{x}_{10}=0) = 0.25$
- B. $p(y=1|\tilde{x}_2=1, \tilde{x}_{10}=0) = 0.313$
- C. $p(y=1|\tilde{x}_2=1, \tilde{x}_{10}=0) = 0.262$
- D. $p(y=1|\tilde{x}_2=1, \tilde{x}_{10}=0) = 0.298$
- E. Don't know.

3

Sum rule $\sum p(x_i|z_k) = 1$

Product rule $p(x_i, y_j | z_k) = p(x_i | y_j, z_k) p(y_j | z_k)$

Marginalization $p(x_i | z_k) = \sum_j p(x_i | y_j, z_k) p(y_j | z_k)$

Bayes theorem $p(y_j | x_i, z_k) = \frac{p(x_i | y_j, z_k) p(y_j | z_k)}{\sum_j p(x_i | y_j, z_k) p(y_j | z_k)}$

$$\begin{aligned} p(y=1 | \tilde{x}_2=1, \tilde{x}_{10}=0) &= \\ \frac{p(\tilde{x}_2=1, \tilde{x}_{10}=0 | y=1) p(y=1)}{\sum_{y' \in \{1, 2, 3\}} p(\tilde{x}_2=1, \tilde{x}_{10}=0 | y') p(y')} &= \end{aligned}$$

$$\begin{aligned}
 &= \frac{p(x_2=1, x_{10}=0 | y=1) p(y=1)}{p(x_2=1, x_{10}=0 | y=1) p(y=1) + p(x_2=1, x_{10}=0 | y=2) p(y=2) + p(x_2=1, x_{10}=0 | y=3) p(y=3)} \\
 &\quad \text{for } y=1 \qquad \qquad \qquad \text{for } y=2 \qquad \qquad \qquad \text{for } y=3 \\
 &= \frac{0.25 \times 0.316}{0.25 \times 0.316 + 0.2 \times 0.356 + 0.35 \times 0.328} \approx 0.296
 \end{aligned}$$

The problem is solved by a simple application of Bayes' theorem:

$$\begin{aligned}
 &p(y=1 | \tilde{x}_2=1, \tilde{x}_{10}=0) \\
 &= \frac{p(\tilde{x}_2=1, \tilde{x}_{10}=0 | y=1)p(y=1)}{\sum_{k=1}^3 p(\tilde{x}_2=1, \tilde{x}_{10}=0 | y=k)p(y=k)}
 \end{aligned}$$

The values of $p(y)$ are given in the problem text and the values of $p(\tilde{x}_2=1, \tilde{x}_{10}=0 | y)$ in ???. Inserting the values we see option D is correct.

Independence

Independent: $p(x_i, y_j) = p(x_i)p(y_j)$

Conditionally independent given z_k : $p(x_i, y_j | z_k) = p(x_i | z_k)p(y_j | z_k)$

Expectations

Expectation: $\mathbb{E}[f] = \sum_{i=1}^N f(x_i)p(x_i).$ (2)

mean: $\mathbb{E}[x] = \sum_{i=1}^N x_i p(x_i),$ Variance: $\text{Var}[x] = \sum_{i=1}^N (x_i - \mathbb{E}[x])^2 p(x_i).$ (3)

Example: Uniform probability

$$p(x_i) = \frac{1}{N}$$

$$\mathbb{E}[f] = \frac{1}{N} \sum_{i=1}^N f(x_i)$$

$$\mathbb{E}[x] = \frac{\sum_{i=1}^N x_i}{N}$$

$$\text{Var}[x] = \frac{1}{N} \sum_{i=1}^N (x_i - \mathbb{E}[x])^2$$

Densities and models

- In machine learning, we want to learn a parameter from data
- Models of the data which use parameters are how we do that
- We build models out of simpler building blocks (distributions (discrete variables see chapter 5) and densities (continuous variables see chapter 6)). In this course we will use four:

Bernoulli distribution

The Categorical distribution

The Beta density

The Multivariate normal density

The Bernoulli distribution

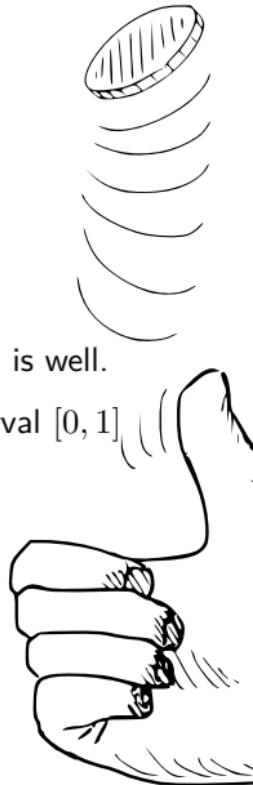
- Let $b = 0, 1$ denote a binary event.
- For instance,
 - $b = 0$ corresponds to heads, and $b = 1$ to tails, or
 - $b = 0$ corresponds to a person being ill, and $b = 1$ that a person is well.
- The probability of b is expressed using a parameter θ in the unit interval $[0, 1]$

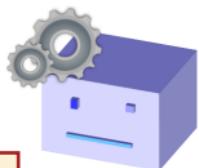
Bernoulli distribution: $p(b|\theta) = \theta^b(1-\theta)^{1-b}$.

$$p(b=1|\theta) = \theta^1(1-\theta)^{1-1} = \theta$$

$$p(b=0|\theta) = \theta^0(1-\theta)^{1-0} = 1-\theta$$

DETØ, B





The Bernoulli distribution, repeated events

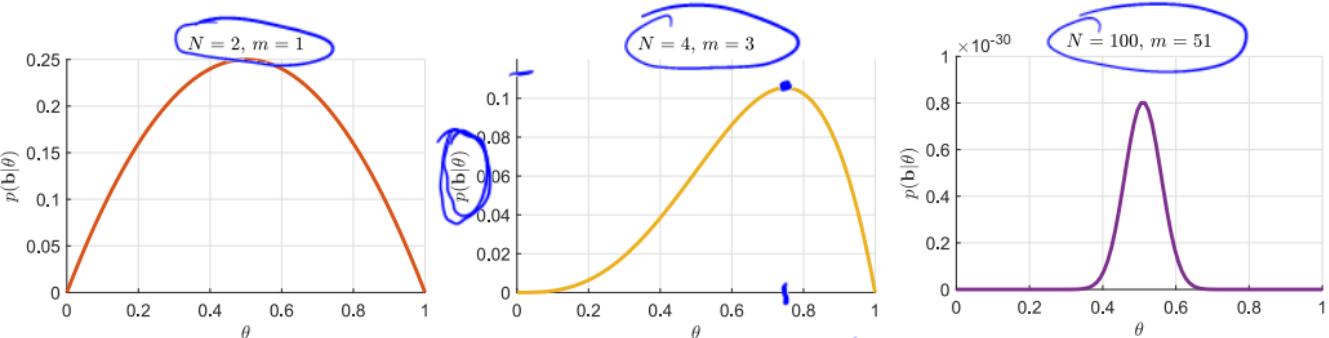
Conditional independence $p(x_i, x_j | z_k) = p(x_i | z_k)p(x_j | z_k)$

- Suppose we observe a sequence b_1, \dots, b_N of Bernoulli (binary) events.
- For instance, for N patients we record whether person 1 is ill or well ($b_1 = 0$ or $b_1 = 1$) and up to whether patient N is ill or well ($b_N = 0$ or $b_N = 1$)
- When we **know** θ (the chance a person is well or ill), the events are **independent**

Bernoulli distribution: $p(b|\theta) = \theta^b(1 - \theta)^{1-b}$.

$$\underbrace{p(b_1, \dots, b_N | \theta)}_{p(b_1, \dots, b_N | \theta)} = \prod_{i=1}^N p(b_i | \theta) = \prod_{i=1}^N \theta^{b_i} (1 - \theta)^{1-b_i} = \theta^{\sum_{i=1}^N b_i} (1 - \theta)^{N - \sum_{i=1}^N b_i}$$
$$= \theta^m (1 - \theta)^{N-m}, \quad m = b_1 + b_2 + \dots + b_N$$

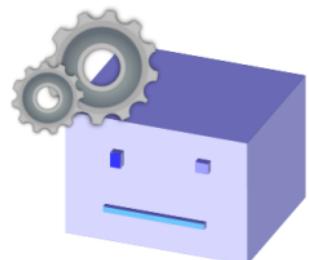
The Bernoulli distribution, maximum likelihood



$$\underbrace{p(b_1, \dots, b_N | \theta)}_{\text{Maximum likelihood}} = \theta^m (1 - \theta)^{N-m}$$

An idea for selecting θ^* is **Maximum likelihood**

$$\theta^* = \arg \max_{\theta} \underbrace{p(b_1, \dots, b_N | \theta)}_{\text{Maximum likelihood}}$$



The value of θ according to which the data is most plausible

Resources

<https://bayes.wustl.edu> Classical textbook which treats probabilities as states-of-knowledge and discuss many practical and philosophical issues (this book converted me to ML!)

(<https://bayes.wustl.edu/etj/prob/book.pdf>)

<https://02402.compute.dtu.d> A more in-depth description of summary statistics (see chapter 1) (<https://02402.compute.dtu.dk>)

<https://www.khanacademy.org> An excellent introduction to probability theory which we recommend as a go-to resource

(<https://www.khanacademy.org/math/statistics-probability/probability-library>)

<http://citeseerx.ist.psu.edu> A more in-depth discussion of Bayes in the court room (<http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=EF0328140036A4C668BB5B9FC76C9BE?doi=10.1.1.599.8675&rep=rep1&type=pdf>)

02450: Introduction to Machine Learning and Data Mining

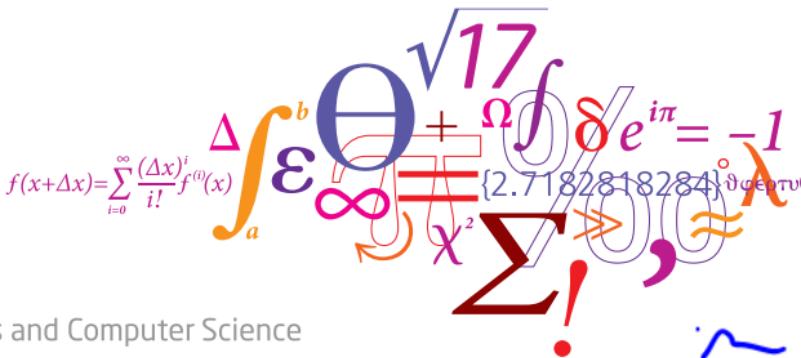
Data, feature extraction and PCA

Bjørn Sand Jensen

DTU Compute, Technical University of Denmark (DTU)

DTU Compute

Department of Applied Mathematics and Computer Science



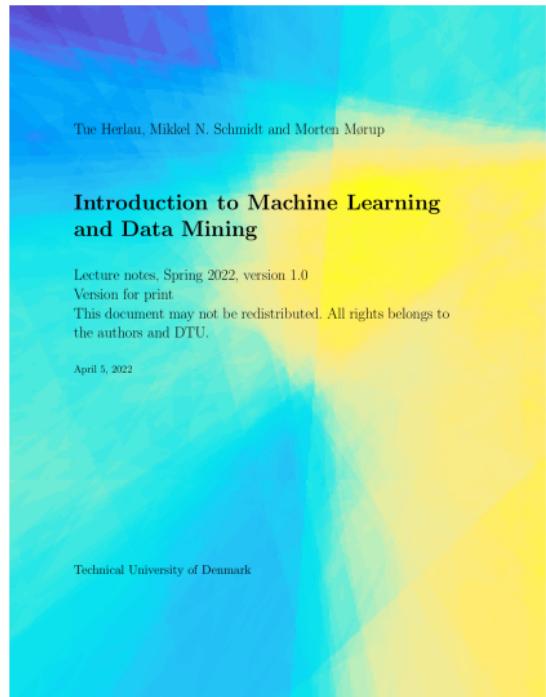
Today

Feedback Groups of the day:

Aref Abbasi, Yanis Zakaria Abdoun, Zahraa Abdulrihman, Samson Alfred Adsersen, Bunia Ingrid Adsersen, Noor Afifi, Gauri Agarwal, Teis Moldrup Sahi Aggerholm, Eunsun Ahn, Saleh Al-Kabani, Anna-mai Allikmäe, Harel Aloni, Saeed Mohamud Amin, Pernille Dyrlund Ammentorp, Pelle Hvidbjørn Hartvig Andersen, Ronnie In Soo Andersen, Michael Andersen, Ida Grum-Schwensen Andersen, Emma Qingjie Andersen, Casper Andresen, Andreas Bjarnastein Antoft, Vár Antoniussen, Prathamesh Arvind Apte, Pedro Aragon Fernandez, Jorge Arias Cuesta, Rasmus Grønnegaard Arnmark, Ramakrishnan Arul Babu, Jonas Babendererde, Oskar Gotthardt Bak, Martynas Baltusis, Fie Søndergaard Bang, Aske Peter Banke, Benjamin Banks, Inés Maria Barroso Müller, Muhammad Numan Bashir, Alexander Baumkirchner, Søren Baunsgaard, Mathilde Wismann Bechgaard, Abderrahim Bendahia, Eleni Ghislaine Marie Berard, Anna-Beatrice Berciu, Marius Arleth Bergstein, Casper Langthjem Bertelsen, Thorfinn Brummer Birkelund, Laura Munch Bjerg

Reading material:

Chapter 2, Chapter 3



Lecture Schedule

1 Introduction

30 January: C1

Data: Feature extraction, and visualization

2 Data, feature extraction and PCA

6 February: C2, C3

3 Measures of Similarity, summary statistics and probabilities

13 February: C4, C5

4 Probability densities and data visualization

20 February: C6, C7

Supervised learning: Classification and regression

5 Decision trees and linear regression

27 February: C8, C9 (Project 1 due 29 February at 17:00)

6 Overfitting, cross-validation and Nearest Neighbor

5 March: C10, C12

7 Performance evaluation, Bayes, and Naive Bayes

12 March: C11, C13

8 Artificial Neural Networks and Bias/Variance

19 March: C14, C15

9 AUC and ensemble methods

2 April: C16, C17

Unsupervised learning: Clustering and density estimation

10 K-means and hierarchical clustering

9 April: C18 (Project 2 due 11 April at 17:00)

11 Mixture models and density estimation

16 April: C19, C20

12 Association mining

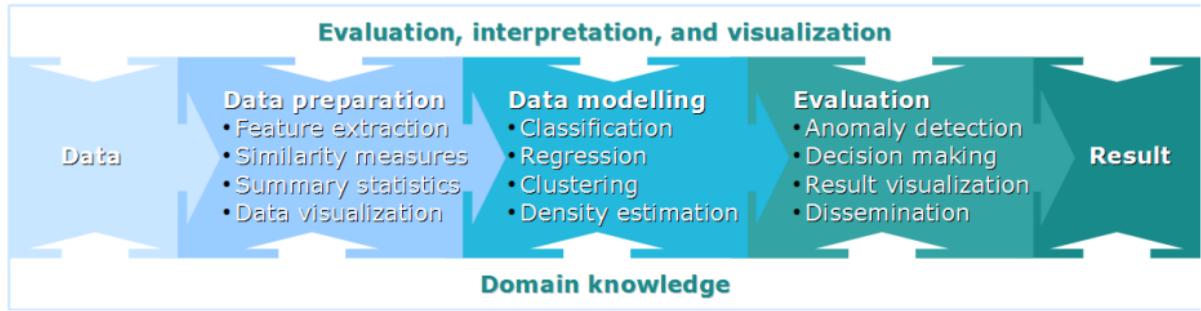
23 April: C21

Recap

13 Recap and discussion of the exam

30 April: C1-C21

Online 24/7 help: Discussion Forum/Piazza
Streaming: Zoom (see link on DTU Learn)
Recordings: <https://panopto.dtu.dk/>
Online exercises: MS Teams



Learning Objectives

- Understand the types of data, their attributes and data issues
- Understand the bag of word representation
- Be able to apply principal component analysis for data visualization and feature extraction

What is data?

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

| Attributes | | | |
|------------|-----|--------|--------|
| 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 countably infinite) set of values
- Examples:
 - Zip codes
 - Counts
 - Set of words in a collection of documents
- Often represented as integer variables

- **Continuous**

- Has real numbers as attribute 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:** Objects can be ranked (Greater than / Less than)
 - Taste of potato chips on a scale from 1-10
 - Grades
 - Height in {short, medium, tall}
- **Interval:** Distance between objects 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

Qualitative

Quantitative



Discussion

- **Classify the following attributes**
 - a) Military rank
 - b) Angles measured in degrees
 - c) A persons year of birth
 - d) A persons age in years
 - e) Coat check number
 - f) Distance from center of campus
 - g) Number of patients 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 1: Attribute types (Spring 2012)

| 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.

Solution:

There are a finite set of brands thus NAME is discrete and as the only operators that can be applied to NAME is equal or not equal NAME is nominal. PROT, FAT and SOD are all continuous and since they have that zero means absence they are ratio. TYPE and VIT are both discrete, however, TYPE is not ordinal, i.e. Hot is not better than Cold, thus

TYPE must be considered nominal, VIT on the other hand is ordinal as 0% is less than 25% which in turn is less than 100%. An attribute that is ratio will also be both interval, ordinal and nominal, i.e. we can apply all the operations $=, \neq, >, <, +, -, *, /$ to a ratio attribute.

Types of data sets

- **Record data**
 - 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

Record data example: Market basket data

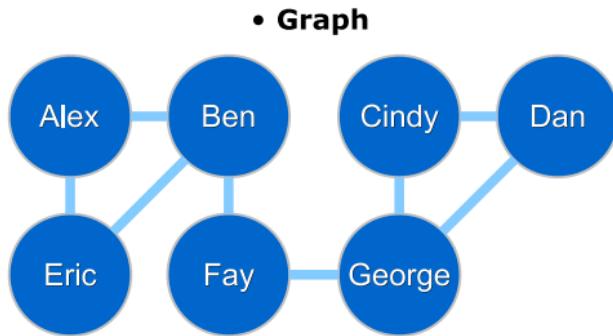
- Transaction data table

| 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 |

Relational data example: Who knows who?

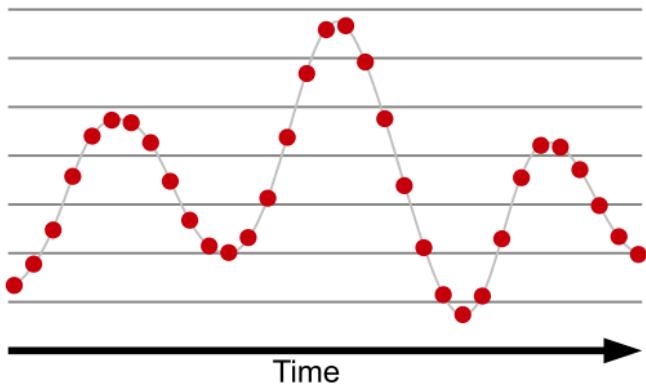


• Matrix

| | 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 |

Ordered data example: 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 |

Data quality

- **Data is of high quality if they**
 - Are fit for their intended use
 - Correctly represent the phenomena they correspond to

- **Examples of quality problems**

- Noise
- Outliers
- Missing values



Noise

- **Definition**

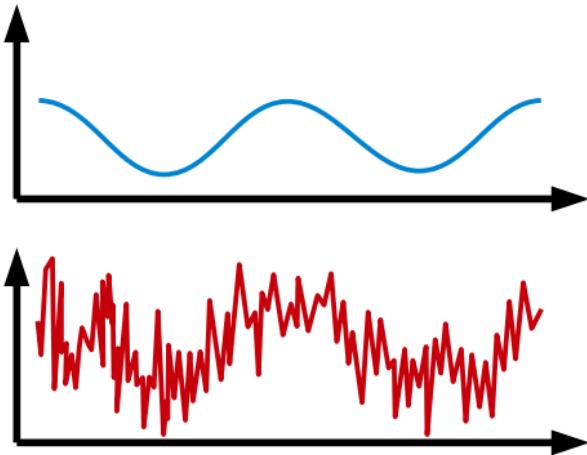
- Unwanted perturbation to a signal
- Unwanted data

- **Reasons for noise**

- Limits in measurement accuracy
- Interference from other signals
- Measurement of attributes not related to the data modeling task

- **Handling noise**

- Exclude noisy attributes
- Remove noise by filtering
- Include a model of the noise



Outliers

- **Definition**

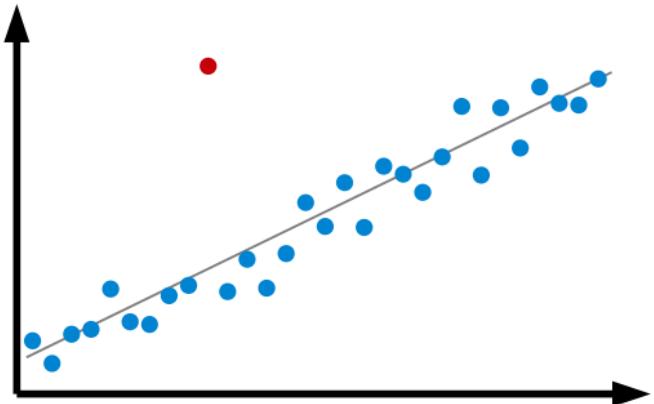
- Data objects which are significantly different from most others

- **Reasons for outliers**

- Measurement error
 - Natural property of data

- **Handling outliers**

- Identify & exclude outliers
 - Model the outliers



Missing values

- **Definition**

- No value is stored for an attribute in a data object

- **Reasons for missing values**

- Information is not collected or measured
 - People decline to give their age
- Attribute is not applicable
 - Annual income is not applicable to children

- **Handling missing values**

- Eliminate data objects
- Eliminate attributes
- Estimate missing values (e.g. an average)
- Ignore the missing value in analysis
- Model the missing values

| 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

3 1 NONE 2 7 3 ,5 2 1 3 2 zero *

- A research 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 occurred, and how should they be handled?

Dataset manipulations

- **Sampling**

- Selecting a representative subset of data points

- **Feature subset selection**

- Choose a subset of attributes

- **Feature extraction/transformation**

- Create new features from existing attributes

- Discretization and binarization

- Apply a fixed transformation to an attribute

- Aggregation several attributes into a single attribute

$x_a x_H$

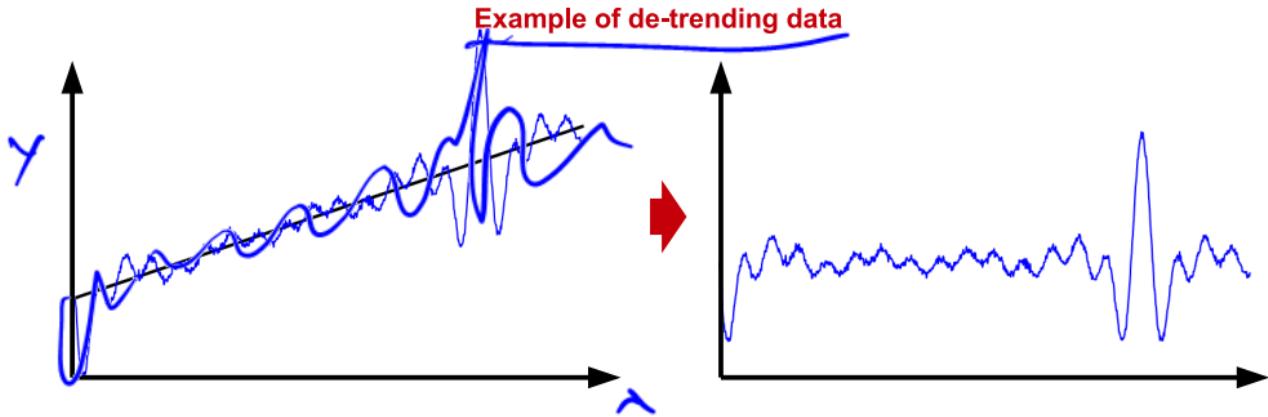
$\log(x_a)$

- **Dimensionality reduction**

- Project 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)



Common feature transformations

| 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 |

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

Standardize:

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

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

Binarize/threshold:

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

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

One-out-of K encoding

One-out-of-K coding

| <u>Age</u> | <u>Height</u> | <u>Weight</u> | <u>Nationality</u> |
|-------------|---------------|---------------|--------------------|
| -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' |
| X = -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' |

| <u>Age</u> | <u>Height</u> | <u>Weight</u> | | | |
|------------|---------------|---------------|---|---|---|
| -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 |

Bag of words representation

- First three sentences on **wikipedia.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



(Image source: <https://pixabay.com/p-297223/>)

Bag of words representation

- First three sentences on **wikipedia.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

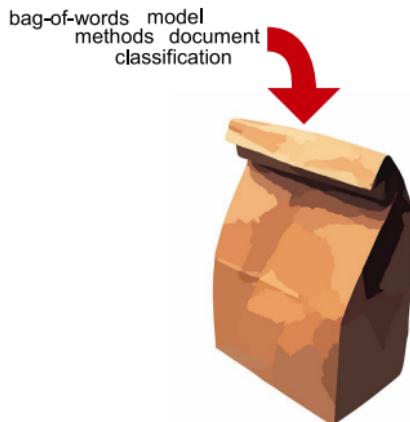


- We will treat **this text** as a data set and create a bag-of-words model of it



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



Bag of words representation

- Representation as 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 | |

Bag of words representation

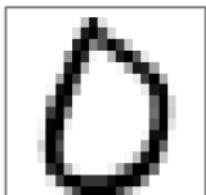
- Stemming

| 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 |

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

$$1 \times 784 \quad x_0 = [0 \quad \dots \quad 0 \quad 0.3 \quad 1 \quad 0.2 \quad 0 \quad \dots \quad 0]^\top$$

- Matrix representation of data set

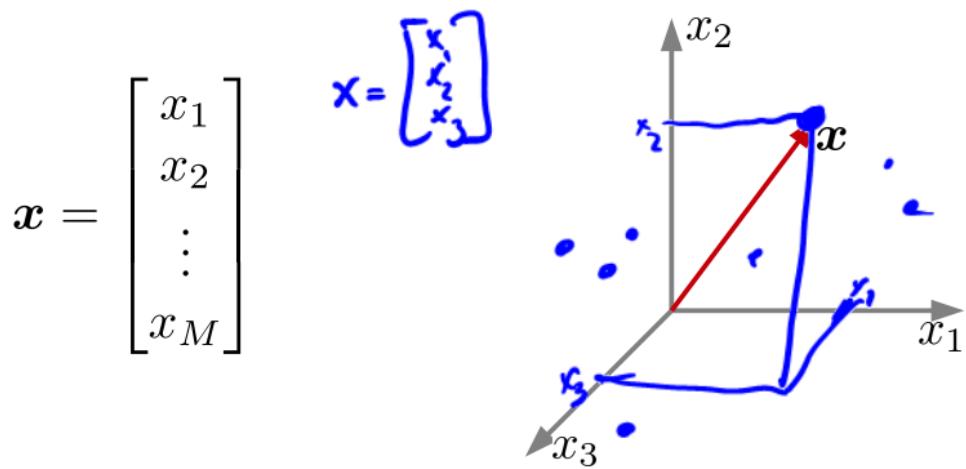
$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$$



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

Vector space representation

- All these data objects have a vector space representation



14:05



Plan for the rest of today:

- Linear algebra recap (subspaces and projections)
- The **goal** of Principal Component Analysis (PCA)
- Derivation of PCA
- Singular Value Decomposition used to implement PCA
- Use of PCA for data visualization

Vectors and matrices

- Common matrix notation

A , A , \overline{A} , \underline{A}

$$A = \begin{bmatrix} a_{1,1} & \cdots & a_{1,M} \\ \vdots & & \vdots \\ a_{N,1} & \cdots & a_{N,M} \end{bmatrix} \in \mathbb{R}^{N \times M}$$

- Common vector notation

x , x , \bar{x} , \vec{x}

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix} \in \mathbb{R}^M$$

Matrix multiplication

- Two matrices can be multiplied $\mathbf{AB} = \mathbf{C}$
 - if the number of columns in the first equals the number of rows in the second

$$\begin{array}{c} \text{A} \times \text{B} = \text{C} \\ L \times M \quad M \times N \quad L \times N \\ \text{3} \times 4 \text{ matrix} \quad \text{4} \times 5 \text{ matrix} \quad \text{3} \times 5 \text{ matrix} \\ \left[\begin{array}{cccc} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & 2 & 3 & 4 \end{array} \right] \left[\begin{array}{ccccc} \cdot & \cdot & \cdot & a & \cdot \\ \cdot & \cdot & \cdot & b & \cdot \\ \cdot & \cdot & \cdot & c & \cdot \\ \cdot & \cdot & \cdot & d & \cdot \end{array} \right] = \left[\begin{array}{ccccc} \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & x_{3,4} & \cdot \end{array} \right] \end{array}$$
$$x_{3,4} = 1 \cdot a + 2 \cdot b + 3 \cdot c + 4 \cdot d$$

Example:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix}$$

Matrix transpose

- The transpose of a matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 6 \\ 7 & 8 & 9 \end{bmatrix} \quad \mathbf{A}^\top = \begin{bmatrix} 1 & 3 & 7 \\ 2 & 4 & 8 \\ 3 & 6 & 9 \end{bmatrix}$$

- Transpose of a sum

$$(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$$

- Transpose of a product

$$(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$$

$$(\mathbf{Ax})^\top \mathbf{y} = \mathbf{x}^\top \mathbf{A}^\top \mathbf{y} = \mathbf{x}^\top (\mathbf{A}^\top \mathbf{y})$$

The identity matrix

- Ones on the diagonal and zeros everywhere else

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \quad \mathbf{I}^\top = \mathbf{I}$$

- Multiplying by the identity does not change anything

$$\mathbf{IA} = \mathbf{A}$$
$$\mathbf{I}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$
$$\mathbf{I}_2 \mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

- For a square matrix, the inverse satisfies

$$\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

Norms

- The (Euclidian) norm of a vector measures it's length (magnitude):

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{n=1}^N |x_n|^2} = \sqrt{\mathbf{x}^\top \mathbf{x}}$$

- The Frobenius norm of a matrix measures it's magnitude:

$$\|\mathbf{X}\|_F^2 = \sum_{i,j} x_{i,j}^2 = \text{trace}(\mathbf{X} \mathbf{X}^T) = \text{trace}(\mathbf{X}^T \mathbf{X})$$

Where trace takes the sum of the diagonal elements, i.e. $\text{trace}(\mathbf{A}) = \sum_{i=1}^N a_{i,i}$

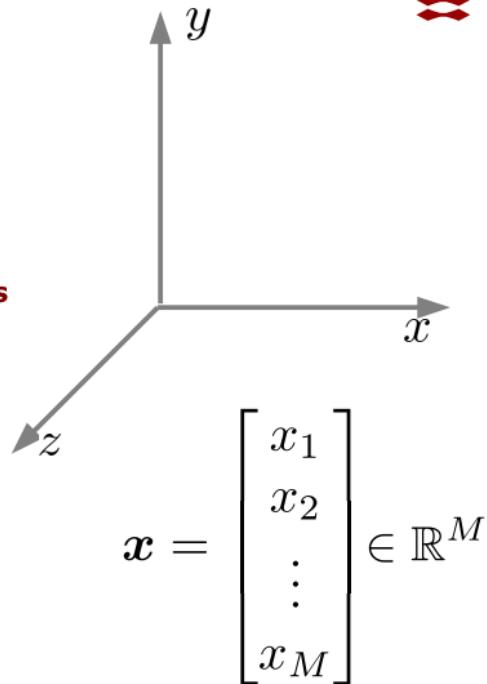
Vector spaces

- A M-dimensional vector space is just \mathbb{R}^M
- This is the set of all M-dimensional vectors
- A vector space is closed under **linear combinations**

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_n\mathbf{x}_n$$

$\mathbf{x}_1, \dots, \mathbf{x}_n$ Vectors

a_1, \dots, a_n Numbers



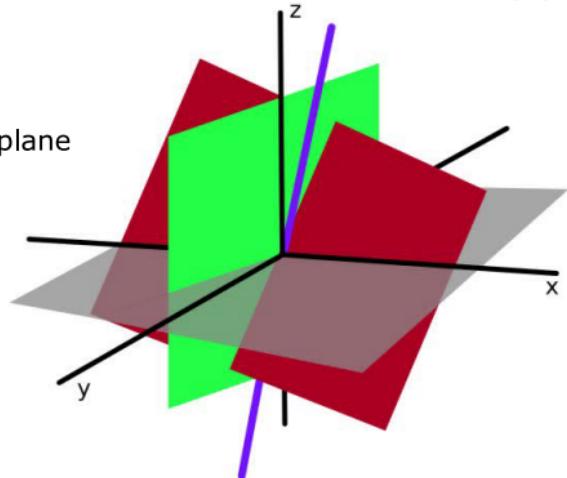
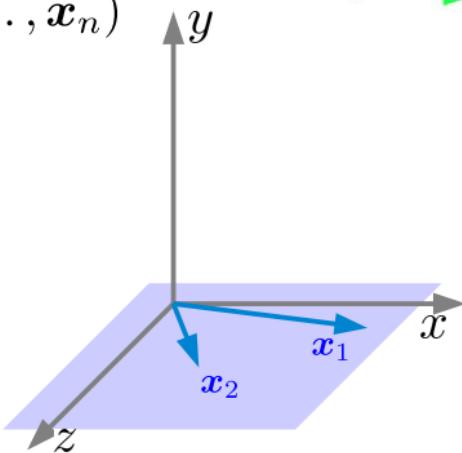
Subspaces

- A **subspace** generalizes the concept of a line/plane
- If we consider n vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ the **span** is then all linear combinations

$$\mathbf{z} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_n\mathbf{x}_n$$

and it is said to be a **subspace**

$$V = \text{span}(\mathbf{x}_1, \dots, \mathbf{x}_n)$$

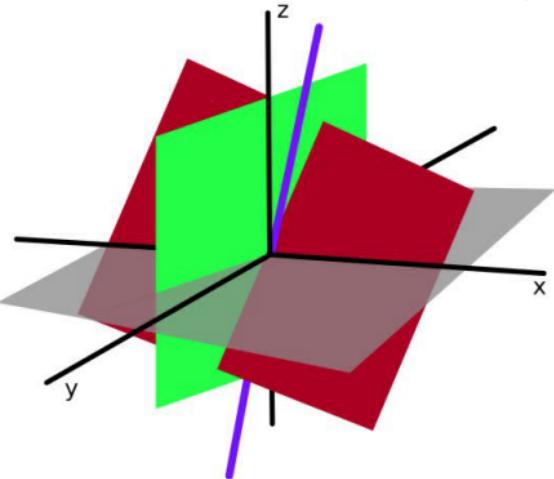


Basis of a (sub)space

- Vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ are said to be **linearly independent** if

$$\mathbf{0} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_n\mathbf{x}_n$$

implies $a_1 = a_2 = \cdots = a_n = 0$



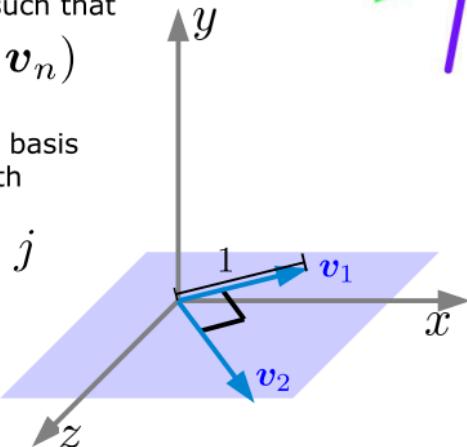
- A **basis** of a vector space V are n linearly independent vectors such that

$$V = \text{span}(\mathbf{v}_1, \dots, \mathbf{v}_n)$$

- A basis is **orthonormal** if the basis is orthogonal and of unit length

$$\mathbf{v}_i^T \mathbf{v}_j = 0 \text{ for } i \neq j$$

$$\|\mathbf{v}_i\| = 1$$



Basis of a (sub)space

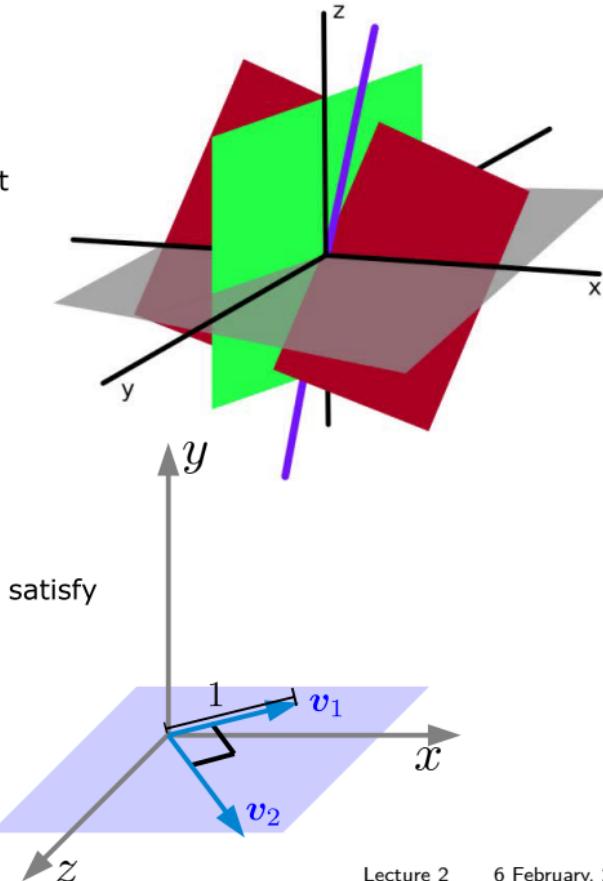
- A **basis** of a vector space V are n linearly independent vectors such that $V = \text{span}(\mathbf{v}_1, \dots, \mathbf{v}_n)$

- We collect the basis into a matrix

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_N \end{bmatrix}$$

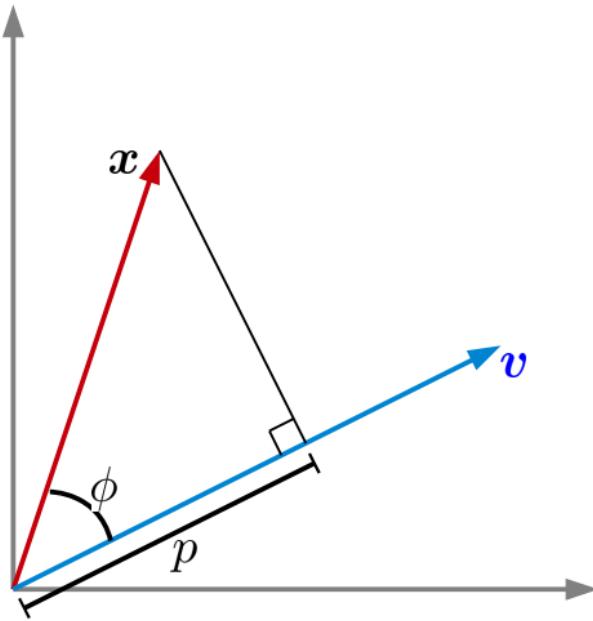
- If the basis is orthonormal the matrix satisfy

$$\mathbf{V}^\top \mathbf{V} = \mathbf{I}, \quad \mathbf{V}^\top = \mathbf{V}^{-1}$$



Projection

- Projection onto a vector



- Angle between vectors

$$\cos(\phi) = \frac{\mathbf{v}^\top \mathbf{x}}{\|\mathbf{x}\|_2 \|\mathbf{v}\|_2}$$

- Length of projection

$$p = \|\mathbf{x}\|_2 \cos(\phi) = \frac{\mathbf{v}^\top \mathbf{x}}{\|\mathbf{v}\|_2}$$

- Projection onto unit vector

$$p = \mathbf{v}^\top \mathbf{x}$$

Projection onto a subspace

- **Projection onto a subspace**

- Subspace of dimension n defined by a orthonormal basis matrix \mathbf{V}
- Projection of \mathbf{x} (M dimensional) onto V given by

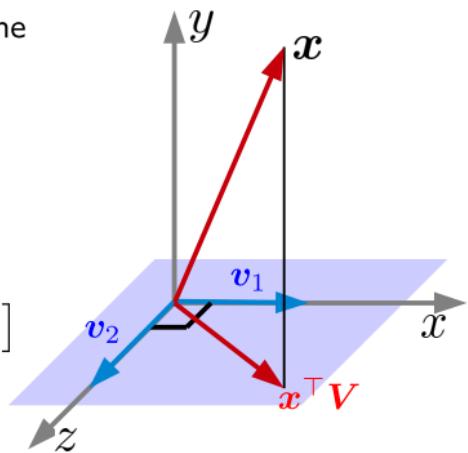
$$[\mathbf{b}_1, \mathbf{b}_2] = \mathbf{b}^T = \underline{\mathbf{x}}^T \underline{\mathbf{V}}$$

- 'Reconstruction' can be found as: $\mathbf{x}' = \mathbf{V}\mathbf{b}$

Example: Projection of 3-D vector onto the (x,z) plane

$$\mathbf{V} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

$$\mathbf{x}^T \mathbf{V} = [x \ y \ z] \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} = [x \ z]$$



Projection onto a subspace

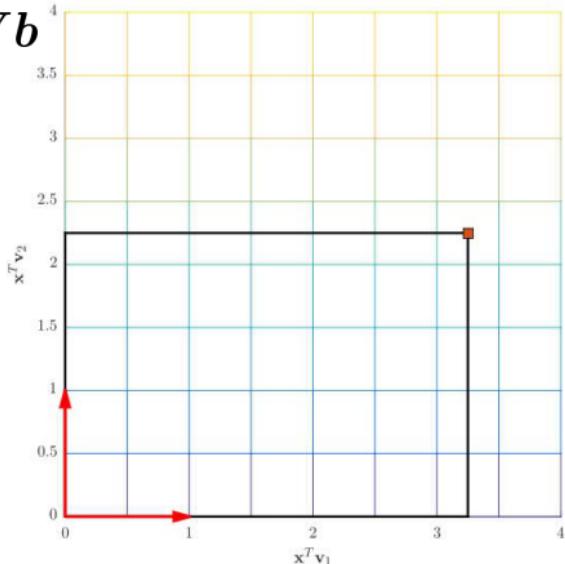
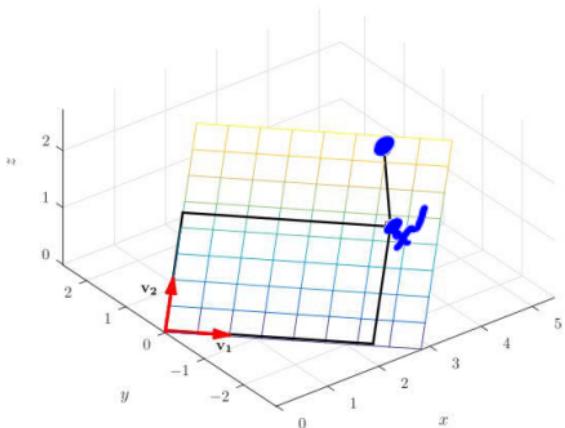
- **Projection onto a subspace**

- Subspace of dimension n defined by a orthonormal basis matrix V
- Projection of x (M dimensional) onto V given by

$$b^T = x^T V$$

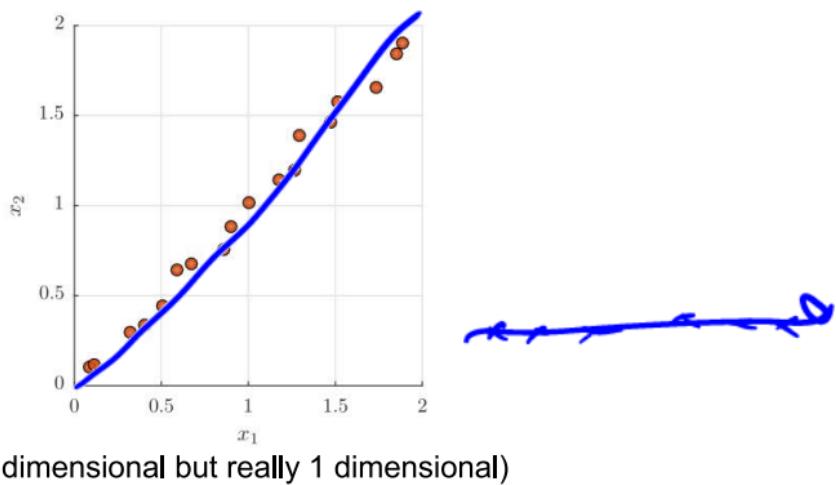
– 'Reconstruction' can be found as: $x' = Vb$

Example 2:



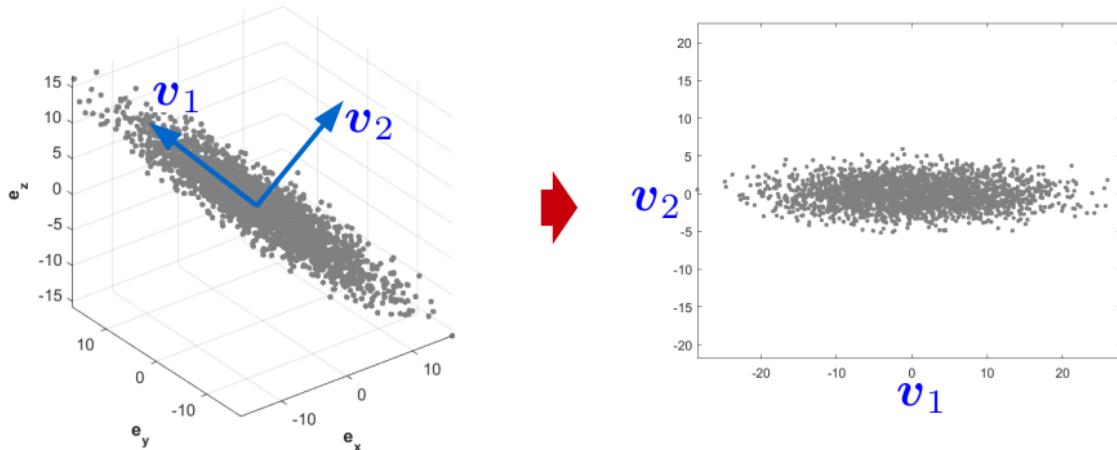
PCA for high-dimensional data

- Much data is high-dimensional
- We want to find a **lower**-dimensional representation of the **high**-dimensional data



PCA for high-dimensional data

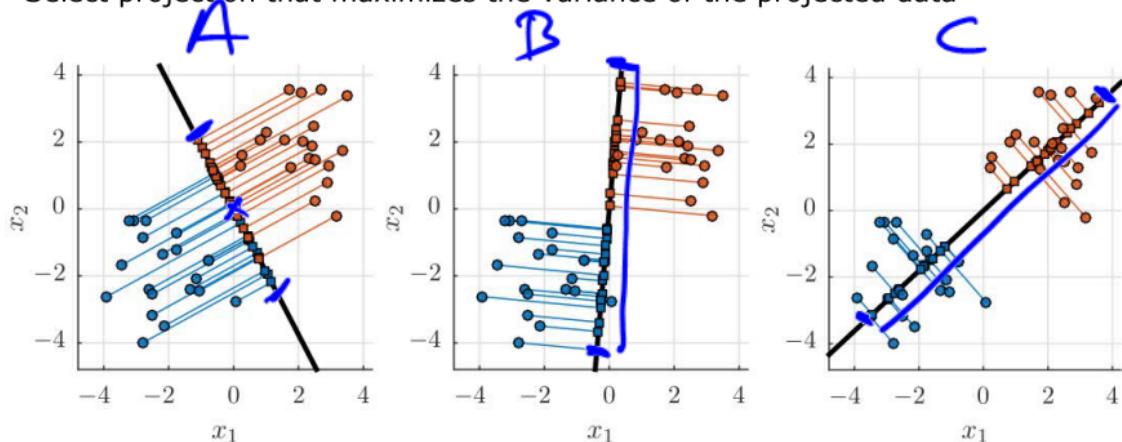
- Much data is high-dimensional
- We can **project high** dimensional data to a **lower** dimensional **subspace**
- But what is a good projection?



PCA for high-dimensional data

$$b^T = x^T V$$

- Much data is high-dimensional
- We can project high dimensional data to a lower dimensional subspace
- But what is a good projection?
- Select projection that maximizes the variance of the projected data



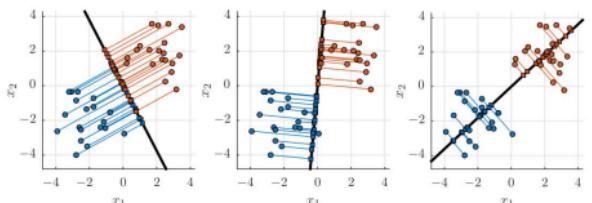
$$\text{Var}[b_i] = \frac{1}{N-1} \sum_{j=1}^N (b_j - \bar{b})^2$$

PCA derivation

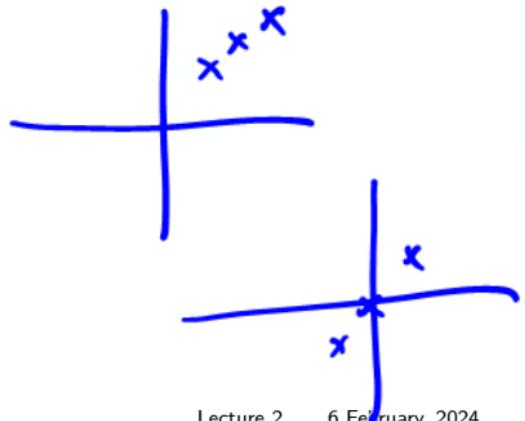
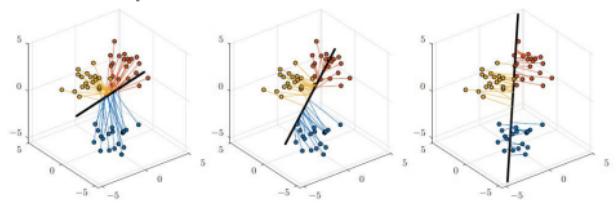
Projection of \mathbf{x}_i onto unit vector \mathbf{v} : $b_i = \mathbf{x}_i^\top \mathbf{v}$

$$\begin{aligned}\text{Var}[b] &= \frac{1}{N-1} \sum_{i=1}^N \left[b_i - \frac{1}{N} \sum_{j=1}^N b_j \right]^2 \\ &= \frac{1}{N-1} \sum_{i=1}^N \left[\mathbf{x}_i^\top \mathbf{v} - \frac{1}{N} \sum_{j=1}^N \mathbf{x}_j^\top \mathbf{v} \right]^2 \\ &= \frac{1}{N-1} \sum_{i=1}^N \left[\left(\mathbf{x}_i^\top - \frac{1}{N} \sum_{j=1}^N \mathbf{x}_j^\top \right) \mathbf{v} \right]^2 \\ &= \frac{1}{N-1} \sum_{i=1}^N \left(\tilde{\mathbf{x}}_i^\top \mathbf{v} \right)^2 \quad \boxed{\tilde{\mathbf{x}}_i = \mathbf{x}_i - \mathbf{m}} \\ &= \frac{1}{N-1} \sum_{i=1}^N \mathbf{v}^\top \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top \mathbf{v} = \frac{1}{N-1} \mathbf{v}^\top \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \mathbf{v}\end{aligned}$$

2D example



3D example

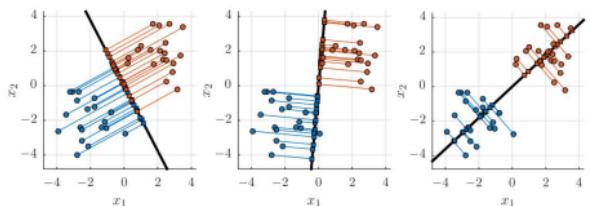


PCA derivation

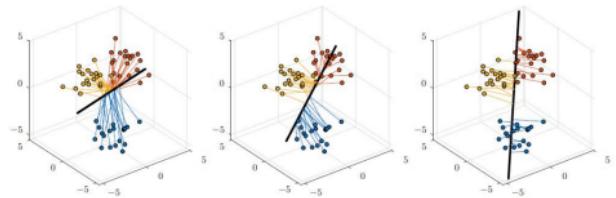
Projection of \mathbf{x}_i onto unit vector \mathbf{v} : $b_i = \mathbf{x}_i^\top \mathbf{v}$

$$\begin{aligned} \text{Var}[b] &= \frac{1}{N-1} \sum_{i=1}^N \left[b_i - \frac{1}{N} \sum_{j=1}^N b_j \right]^2 \\ &= \frac{1}{N-1} \sum_{i=1}^N \left[\mathbf{x}_i^\top \mathbf{v} - \frac{1}{N} \sum_{j=1}^N \mathbf{x}_j^\top \mathbf{v} \right]^2 \\ &= \frac{1}{N-1} \sum_{i=1}^N \left[\left(\mathbf{x}_i^\top - \frac{1}{N} \sum_{j=1}^N \mathbf{x}_j^\top \right) \mathbf{v} \right]^2 \\ &= \frac{1}{N-1} \sum_{i=1}^N \left(\tilde{\mathbf{x}}_i^\top \mathbf{v} \right)^2 \quad \boxed{\tilde{\mathbf{x}}_i = \mathbf{x}_i - \mathbf{m}} \\ &= \frac{1}{N-1} \sum_{i=1}^N \mathbf{v}^\top \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top \mathbf{v} = \frac{1}{N-1} \mathbf{v}^\top \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \mathbf{v} \end{aligned}$$

2D example



3D example



$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}, \quad \mathbf{A} \text{ is a } N \times N \text{ matrix}$$

We say \mathbf{v} is an eigenvector with eigenvalue λ

$$\arg \max_{\mathbf{v}} \text{Var}[b] = \arg \max_{\mathbf{v}} \mathbf{v}^T \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \mathbf{v} \quad \text{s.t.} \quad \|\mathbf{v}\|^2 = \mathbf{v}^T \mathbf{v} = 1$$

$$\mathcal{L}(\mathbf{v}, \lambda) = \mathbf{v}^T \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \mathbf{v} - \lambda(\mathbf{v}^T \mathbf{v} - 1), \quad \frac{\partial \mathcal{L}}{\partial \mathbf{v}} = 2 \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \mathbf{v} - 2\lambda \mathbf{v} = 0$$

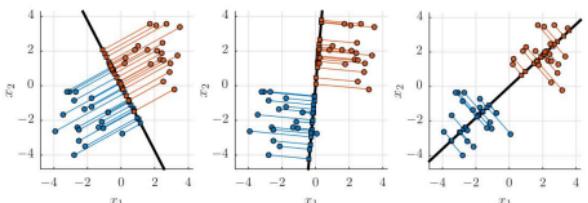
$$\text{or } \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \mathbf{v} = \lambda \mathbf{v}$$

PCA derivation

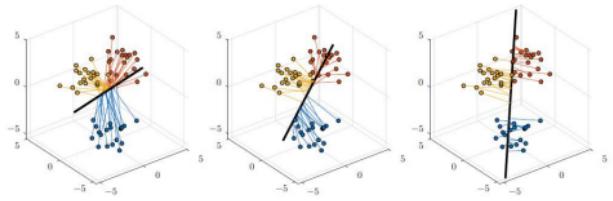
Projection of \mathbf{x}_i onto unit vector \mathbf{v} : $b_i = \mathbf{x}_i^\top \mathbf{v}$

$$\begin{aligned}\text{Var}[b] &= \frac{1}{N-1} \sum_{i=1}^N \left[b_i - \frac{1}{N} \sum_{j=1}^N b_j \right]^2 \\ &= \frac{1}{N-1} \sum_{i=1}^N \left[\mathbf{x}_i^\top \mathbf{v} - \frac{1}{N} \sum_{j=1}^N \mathbf{x}_j^\top \mathbf{v} \right]^2 \\ &= \frac{1}{N-1} \sum_{i=1}^N \left[\left(\mathbf{x}_i^\top - \frac{1}{N} \sum_{j=1}^N \mathbf{x}_j^\top \right) \mathbf{v} \right]^2 \\ &= \frac{1}{N-1} \sum_{i=1}^N \left(\tilde{\mathbf{x}}_i^\top \mathbf{v} \right)^2 \quad \boxed{\tilde{\mathbf{x}}_i = \mathbf{x}_i - \mathbf{m}} \\ &= \frac{1}{N-1} \sum_{i=1}^N \mathbf{v}^\top \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top \mathbf{v} = \frac{1}{N-1} \mathbf{v}^\top \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \mathbf{v}\end{aligned}$$

2D example



3D example



$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}, \quad \mathbf{A} \text{ is a } N \times N \text{ matrix}$$

We say \mathbf{v} is an eigenvector with eigenvalue λ

$$\frac{\partial \mathcal{L}}{\partial \mathbf{v}} = 2\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \mathbf{v} - 2\lambda \mathbf{v} = 0 \quad \text{or} \quad \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \mathbf{v} = \lambda \mathbf{v}$$

This means that $\text{Var}[b] = \frac{1}{N-1} \mathbf{v}^\top \lambda \mathbf{v} = \frac{1}{N-1} \lambda$

The Singular Value Decomposition (SVD)

(Eugino Beltrami & Camille Jordan, independently, 1873-1874)

Any $N \times M$ matrix can be decomposed as follows:

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top$$

$$\begin{matrix} \mathbf{X} \\ N \times M \end{matrix} = \begin{matrix} \mathbf{U} \\ N \times N \end{matrix} \begin{matrix} \Sigma \\ N \times M \end{matrix} \begin{matrix} \mathbf{V}^\top \\ M \times M \end{matrix}$$

Orthonormal Diagonal Orthonormal

$\sigma_1, \dots, \sigma_M$
is known as the singular values

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_M \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_N \end{bmatrix} \quad \mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_M \end{bmatrix}$$

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_M \geq 0$$

$$\text{if } i \neq j: \Sigma_{i,j} = 0, \quad \mathbf{U}^\top \mathbf{U} = \mathbf{I}_{N \times N}, \quad \mathbf{V}^\top \mathbf{V} = \mathbf{I}_{M \times M}$$

The Singular Value Decomposition (SVD)

$$\tilde{X}^T \tilde{X} v = \lambda v$$

(Eugino Beltrami & Camille Jordan, independently, 1873-1874)

Any $N \times M$ matrix can be decomposed as follows:

$$\tilde{X} = U \Sigma V^\top$$

\tilde{X}
 U
 Σ
 V^\top

$N \times M$
 $N \times N$
 $N \times M$
 $M \times M$

$$v_i^\top v_i = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_M \geq 0$$

is known as the singular values

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_M \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

$$\begin{aligned} (\tilde{X}^\top \tilde{X}) v_i &= (\underline{U} \Sigma \underline{V}^\top)^\top \underline{U} \Sigma \underline{V}^\top v_i \\ &= (\underline{V} \Sigma^\top \underline{U}^\top \underline{U} \Sigma \underline{V}^\top) v_i \\ &= \underline{V} \Sigma^\top \Sigma e_i = \sigma_{ii}^2 v_i \end{aligned}$$

$$A v = \lambda v, \quad A \text{ is a } N \times N \text{ matrix}$$

We say v is an eigenvector with eigenvalue λ



Principal component analysis (PCA)

(Karl Pearson, 1901)

$$\tilde{\mathbf{b}}^T = \tilde{\mathbf{x}}^T \mathbf{V}_{(K)}$$

$$\tilde{\mathbf{x}}^T \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_K \end{bmatrix}$$

- 1) Subtract the mean from each observation $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \mathbf{m}$
- 2) Apply singular value decomposition (SVD) $\tilde{\mathbf{X}} = \mathbf{U} \Sigma \mathbf{V}^T$

$$\tilde{\mathbf{X}} = \mathbf{U} \Sigma \mathbf{V}^T$$

$N \times M \quad N \times N \quad N \times M \quad M \times M$

- 3) Select first K columns of \mathbf{V} (the PCA projection operation) and first K columns of Σ .

$$\hat{\mathbf{X}} = \mathbf{U} \Sigma_{(K)}$$

$N \times K \qquad M \times K$

(PCA components or PCA projection of the data) (PCA loadings)

$$\mathbf{V}_{(K)} = \begin{bmatrix} \mathbf{v}_1 & \dots & \mathbf{v}_K \end{bmatrix}$$

Principal component analysis (PCA)

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top$$

$$\text{Var}[b] = \frac{1}{N-1} \lambda = \frac{1}{N-1} \sigma^2$$

- Entries in the diagonal matrix Σ are called **singular values**
 - They are sorted (largest first)
 - Indicate how much variability is explained by the corresponding component
 - 1st component explains most of the variability
 - 2nd component explains most of the remaining variability
 - Etc.

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N \end{bmatrix} \quad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_N \geq 0$$

Explained Variance

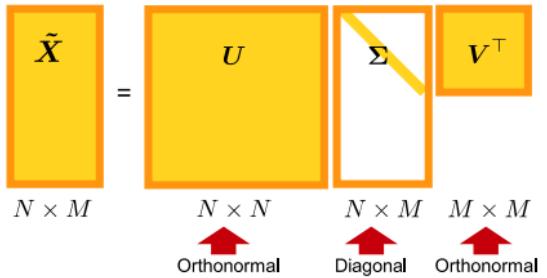
Recall that from SVD: $\tilde{\mathbf{X}} = \mathbf{U}\Sigma\mathbf{V}^T$

In the original space, the coordinates of $\tilde{\mathbf{X}}$ project onto the first K components are:

$$\mathbf{X}' = \mathbf{U}\Sigma_{(K)}\mathbf{V}_{(K)}^T$$

We can measure how much variance is retained in the reconstruction \mathbf{X}' :

$$\text{Explained var.} = \frac{\|\mathbf{X}'\|_F^2}{\|\tilde{\mathbf{X}}\|_F^2}$$



$$\text{cov}(\tilde{\mathbf{X}}) = \tilde{\mathbf{X}}^T \tilde{\mathbf{X}}$$

$$\text{trace}(\mathbf{AB}) = \text{trace}(\mathbf{BA})$$

$$\|\tilde{\mathbf{X}}\|_F^2 = \text{trace}(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}) = \text{trace}(\tilde{\mathbf{X}} \tilde{\mathbf{X}}^T)$$

Explained Variance

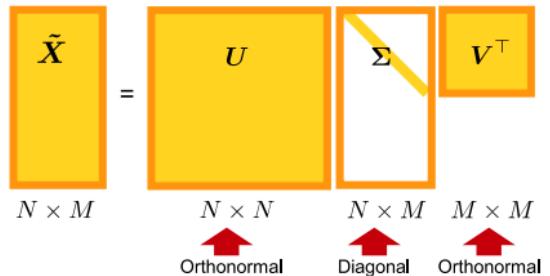
Recall that from SVD: $\tilde{\mathbf{X}} = \mathbf{U}\Sigma\mathbf{V}^T$

In the original space, the coordinates of $\tilde{\mathbf{X}}$ project onto the first K components are:

$$\mathbf{X}' = \mathbf{U}\Sigma_{(K)}\mathbf{V}_{(K)}^T$$

We can measure how much variance is retained in the reconstruction \mathbf{X}' :

$$\text{Explained var.} = \frac{\|\mathbf{X}'\|_F^2}{\|\tilde{\mathbf{X}}\|_F^2} = \frac{\sum_{i=1}^K \sigma_i^2}{\sum_{i=1}^M \sigma_i^2}$$



$$\text{cov}(\tilde{\mathbf{X}}) = \tilde{\mathbf{X}}^T \tilde{\mathbf{X}}$$

$$\text{trace}(\mathbf{AB}) = \text{trace}(\mathbf{BA})$$

$$\begin{aligned}\|\tilde{\mathbf{X}}\|_F^2 &= \text{trace}(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}) = \text{trace}(\tilde{\mathbf{X}} \tilde{\mathbf{X}}^T) \\ &= \text{trace}(\mathbf{U}\Sigma\mathbf{V}^T (\mathbf{U}\Sigma\mathbf{V}^T)^T) \\ &= \text{trace}(\mathbf{U}\Sigma\mathbf{V}^T \mathbf{V}\Sigma^T \mathbf{U}^T) \\ &= \text{trace}(\mathbf{U}\Sigma\Sigma^T \mathbf{U}^T) \\ &= \text{trace}(\mathbf{U}^T \mathbf{U}\Sigma\Sigma^T) \\ &= \text{trace}(\Sigma\Sigma^T) = \sum_i \sigma_i^2\end{aligned}$$

Explained Variance

Recall that from SVD: $\tilde{\mathbf{X}} = \mathbf{U}\Sigma\mathbf{V}^T$

In the original space, the coordinates of $\tilde{\mathbf{X}}$ project onto the first K components are:

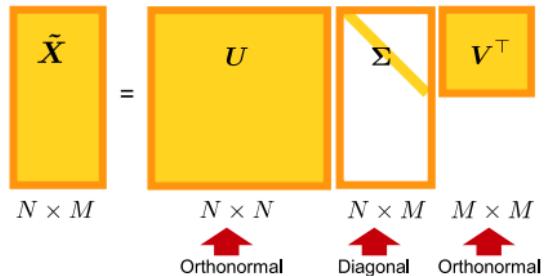
$$\mathbf{X}' = \mathbf{U}\Sigma_{(K)}\mathbf{V}_{(K)}^T$$

We can measure how much variance is retained in the reconstruction \mathbf{X}' :

$$\text{Explained var.} = \frac{\|\mathbf{X}'\|_F^2}{\|\tilde{\mathbf{X}}\|_F^2} = \frac{\sum_{i=1}^K \sigma_i^2}{\sum_{i=1}^M \sigma_i^2}$$

Similarly, the fraction of explained variance for the i 'th component is

$$\text{Explained var.} = \frac{\sigma_i^2}{\sum_{i=1}^M \sigma_i^2}$$



$$\text{cov}(\tilde{\mathbf{X}}) = \tilde{\mathbf{X}}^T \tilde{\mathbf{X}}$$

$$\text{trace}(\mathbf{AB}) = \text{trace}(\mathbf{BA})$$

$$\begin{aligned}\|\tilde{\mathbf{X}}\|_F^2 &= \text{trace}(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}) = \text{trace}(\tilde{\mathbf{X}} \tilde{\mathbf{X}}^T) \\ &= \text{trace}(\mathbf{U}\Sigma\mathbf{V}^T (\mathbf{U}\Sigma\mathbf{V}^T)^T) \\ &= \text{trace}(\mathbf{U}\Sigma\mathbf{V}^T \mathbf{V}\Sigma^T \mathbf{U}^T) \\ &= \text{trace}(\mathbf{U}\Sigma\Sigma^T \mathbf{U}^T) \\ &= \text{trace}(\mathbf{U}^T \mathbf{U}\Sigma\Sigma^T) \\ &= \text{trace}(\Sigma\Sigma^T) = \sum_i \sigma_i^2\end{aligned}$$

Quiz 2: PCA (Fall 2012)

$$\frac{40.1^2}{\|X\|_F^2} = \frac{1}{\sum_{i=1}^4 \sigma_i^2}$$

A PCA analysis is applied to the standardized data based on the attributes x_1-x_{10} . The squared Frobenius norm of the standardized data matrix \mathbf{X} is given by $\|\mathbf{X}\|_F^2 = 5780.0$. The first four singular values are $\sigma_1 = 40.1$, $\sigma_2 = 34.2$, $\sigma_3 = 28.1$, and $\sigma_4 = 24.8$. Which of the following statements is correct?

- A. The first PCA component accounts for more than 35 % of the variation.
- B. The second PCA component accounts for more than 30 % of the variation.
- C. The first three PCA components account for less than 70 % of the variation in the data.
- D. The fourth PCA component accounts for less than 10 % of the variation in the data.
- E. Don't know.

| No. | Attribute description | Abbrev. |
|----------|-------------------------------------|---------|
| x_1 | Age (in years) | AGE |
| x_2 | Gender (Female=0, Male=1) | GDR |
| x_3 | Total Bilirubin | TB |
| x_4 | Direct Bilirubin | DB |
| x_5 | Alkaline Phosphotase | AP |
| x_6 | Alamine Aminotransferase | ALA |
| x_7 | Aspartate Aminotransferase | ASA |
| x_8 | Total Proteins | TP |
| x_9 | Albumin | AB |
| x_{10} | Albumin to Globulin ratio | A/G |
| y | 0=No liver disease, 1=Liver disease | LD |

Table 1: Attributes in a study on liver disease among Indians living in the north eastern part of Andhra Pradesh, India. (taken from <http://archive.ics.uci.edu/ml/datasets/ILPD> +%28Indian+Liver+Patient+Dataset%29). The data has 10 input attributes x_1-x_{10} and one output variable y which defines whether the subject considered has a liver disease ($y = 1$) or not ($y = 0$). x_3-x_9 are non-negative measurements giving the concentrations of various quantities measured in a blood test. x_{10} gives the ratio of Albumin to Globulin in the blood.

Solution:

The i^{th} principal component accounts for $\frac{\sigma_i^2}{\sum_j \sigma_j^2} = \frac{\sigma_i^2}{\|X\|_F^2}$. We therefore have that the first PCA component accounts for $\frac{40.1^2}{5780.0} = 27.8\%$, the second $\frac{34.2^2}{5780.0} = 20.2\%$, and the first three principal components ac-

count for $\frac{40.1^2 + 34.2^2 + 28.1^2}{5780.0} = 61.7\%$ of the variation whereas the fourth principal component accounts for $\frac{24.8^2}{5780.0} = 10.6\%$. Thus, the first three PCA components account for less than 70% of the variation in the data.

Fishers Iris Data

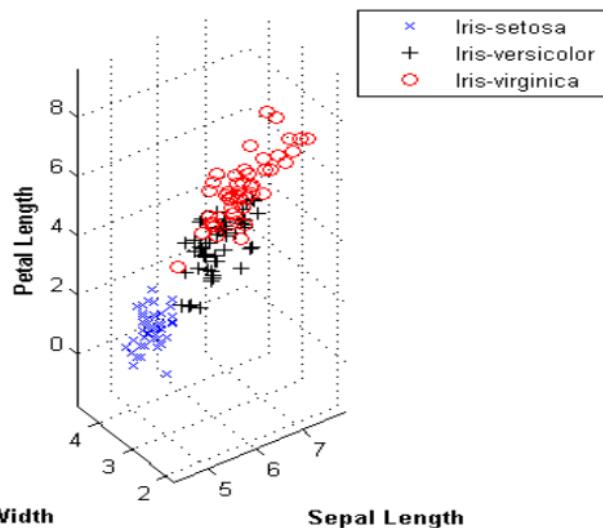


**Three types of flowers:
Iris Setosa, Iris Versicolor, Iris Virginica**

| Flower ID | Attribute | | | | Petal Width |
|-----------|--------------|-------------|--------------|-------------|-------------|
| | Sepal Length | Sepal Width | Petal Length | Petal Width | |
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | |
| 2 | 4.9 | 3.0 | 1.4 | 0.2 | |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 | |
| 4 | 4.6 | 3.1 | 1.5 | 0.2 | |
| . | . | . | . | . | |
| . | . | . | . | . | |
| 150 | 5.9 | 3.0 | 5.1 | 1.8 | |

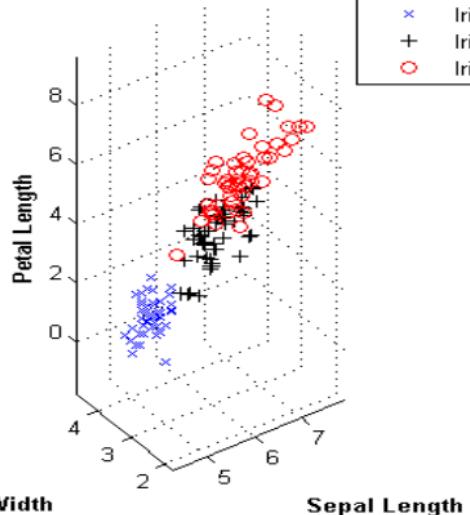
We will presently consider the first 3 attributes, i.e. Sepal length, Sepal Width and Petal Length.

3D scatter plot of Iris Data



What fraction of the total variation in the data will the first principal component account for?

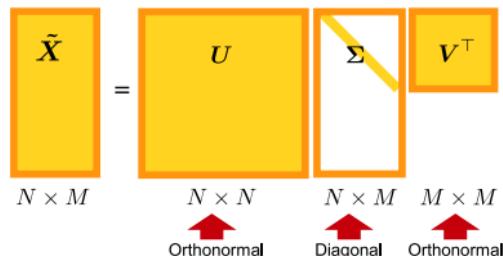
3D scatter plot of Iris Data



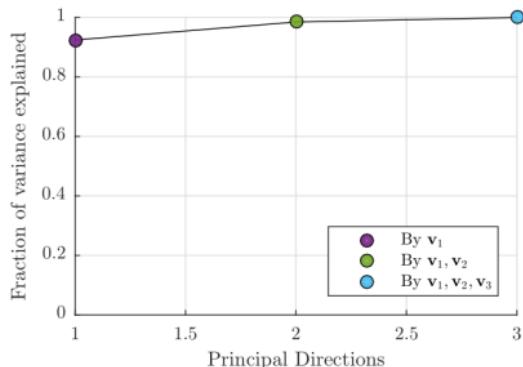
What fraction of the total variation in the data will the first principal component account for?

- 1) Subtract the mean
- 2) Apply singular value decomposition (SVD)

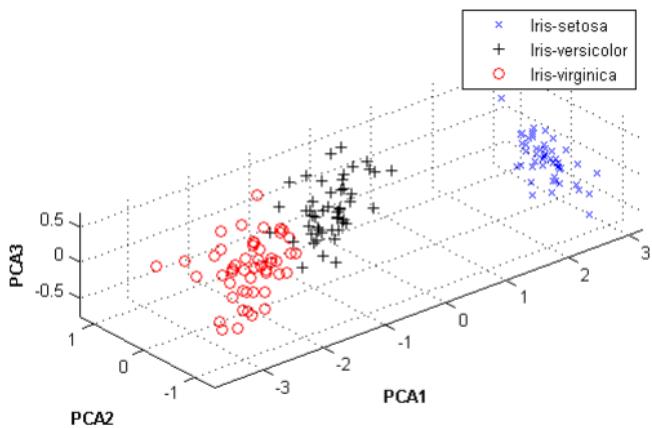
$$\tilde{X} = U \Sigma V^T$$



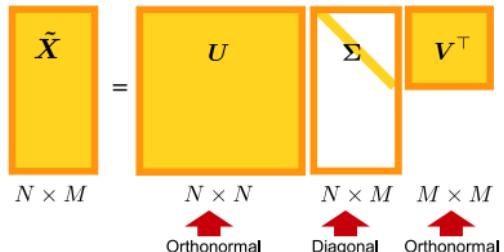
Evaluate the singular values to determine how much of the dynamics is lost when reducing the dimensionality



Visualization of the PCA projections of the data



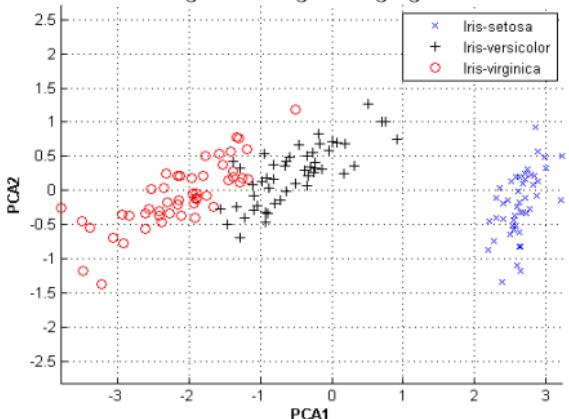
$$\tilde{X} = U\Sigma V^\top$$



$$PCA1: b_1 = \tilde{X}v_1 = u_1\sigma_1$$

$$PCA2: b_2 = \tilde{X}v_2 = u_2\sigma_2$$

$$PCA3: b_3 = \tilde{X}v_3 = u_3\sigma_3$$



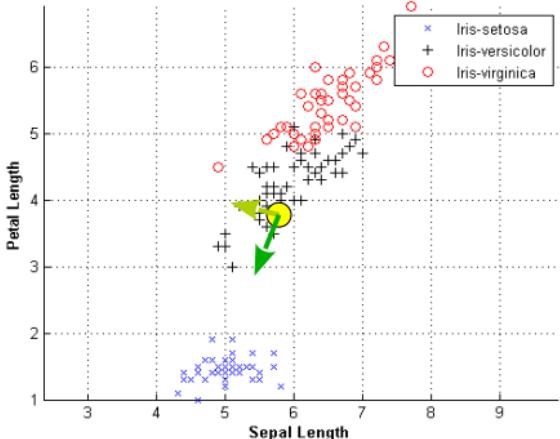
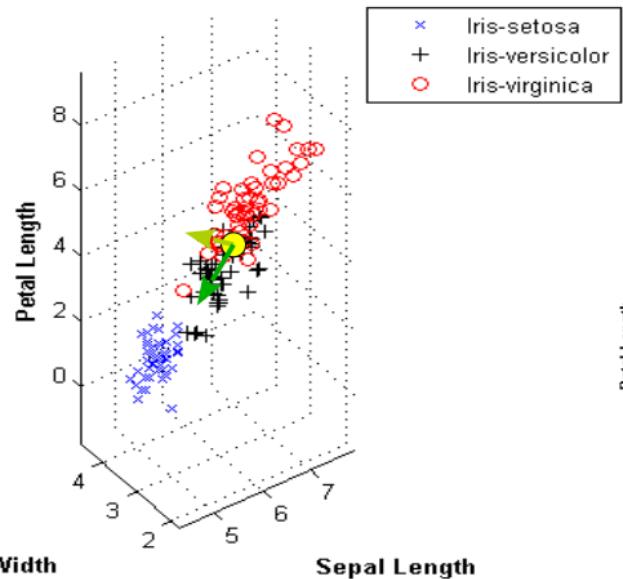
The principal directions V

Sepal Length
Sepal Width
Petal Length

$$V = \begin{bmatrix} -0.39 & -0.64 & -0.66 \\ 0.09 & -0.74 & 0.66 \\ -0.92 & 0.20 & 0.35 \end{bmatrix}$$

$$\mu = \begin{bmatrix} 5.8 \\ 3.1 \\ 3.8 \end{bmatrix}, \quad v_1 = \begin{bmatrix} -0.39 \\ 0.09 \\ -0.92 \end{bmatrix}, \quad v_2 = \begin{bmatrix} -0.64 \\ -0.74 \\ 0.20 \end{bmatrix}$$

Sepal Length
Sepal Width
Petal Length



Quiz 3: PCA Cont. (Fall 2012)

| No. | Attribute description | Abbrev. |
|----------|-------------------------------------|---------|
| x_1 | Age (in years) | AGE |
| x_2 | Gender (Female=0, Male=1) | GDR |
| x_3 | Total Bilirubin | TB |
| x_4 | Direct Bilirubin | DB |
| x_5 | Alkaline Phosphotase | AP |
| x_6 | Alanine Aminotransferase | AIA |
| x_7 | Aspartate Aminotransferase | AsA |
| x_8 | Total Proteins | TP |
| x_9 | Albumin | AB |
| x_{10} | Albumin to Globulin ratio | A/G |
| y | 0=No liver disease, 1=Liver disease | LD |

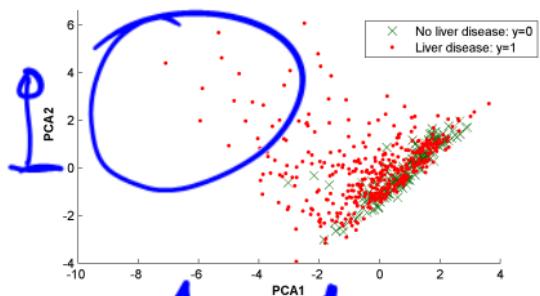


Figure 1: Principal component 1 (PCA1) plotted against principal component 2 (PCA2).

The first and second principal component directions

$$\hat{y} = \tilde{x}^T v,$$

of the liver-dataset are

$$v_1 = \begin{bmatrix} -0.1404 \\ -0.1090 \\ -0.4115 \\ -0.4179 \\ -0.2468 \\ -0.2682 \\ -0.3009 \\ 0.2781 \\ 0.4375 \\ 0.3638 \end{bmatrix} \quad v_2 = \begin{bmatrix} -0.2859 \\ 0.0130 \\ 0.2510 \\ 0.2622 \\ 0.0525 \\ 0.4162 \\ 0.3927 \\ 0.4197 \\ 0.4323 \\ 0.3052 \end{bmatrix}.$$

In the figure, the data projected onto the first two principal components is plotted, and the colors indicate the presence of liver disease. Which of the following statements is *correct*?

- A. Relatively high values of AGE, GDR, TB, DB, AP, AIA, and AsA and low values of TP, AB, and A/G will result in a positive projection onto the first principal component.
- B. Relatively low values of the projection onto PCA1 and high values of the projection onto PCA2 indicates the subject does not have a liver disease.
- C. PCA2 mainly discriminate between old subjects with low measurements of TB, DB, AIA, AsA, TP, AB, and A/G from young subjects with high values of TB, DB, AIA, AsA, TP, AB, and A/G.
- D. The principal component directions are not guaranteed to be orthogonal to each other since the data has been standardized.



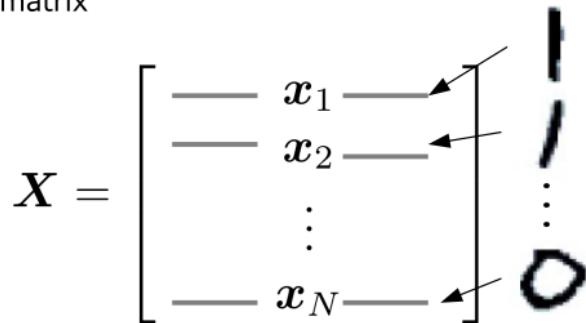
Solution:

AGE, GDR, TB, DB, AP, AlA, and AsA have negative coefficients of PCA1 whereas TP, AB, and A/G have positive coefficients resulting in a negative projection onto the first principal component, thus this is correct. From the figure we observe that observations with low values of PCA1 and high values of PCA2 in general have a red dot meaning they have a liver disease. For PCA2 we observe that AGE has a negative value whereas the remaining entities

have positive values while GDR and AP have small amplitudes. As a result PCA2 mainly discriminate between young subjects with high measurements of TD, DB, AlA, AsA, TP, AB, and A/G from old subjects with low values of TD, DB, AlA, AsA, TP, AB, and A/G hence this is correct. The principal component directions are always orthogonal to each other irrespective of the data preprocessing.

Visualization of hand written digits

- Data matrix

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


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

- Principal component analysis

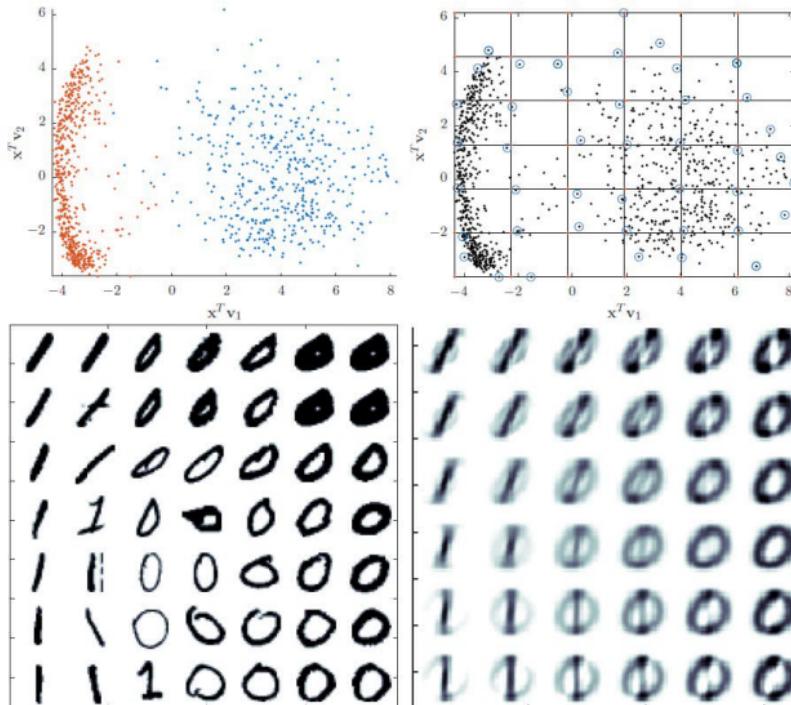
$$\tilde{\boldsymbol{X}} = \boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{V}^\top$$

$$\tilde{\boldsymbol{X}} = \begin{array}{c|c|c|c} \boldsymbol{U} & \boldsymbol{\Sigma} & \boldsymbol{V}^\top & \\ \hline N \times M & N \times N & N \times M & M \times M \end{array}$$

Orthonormal Diagonal Orthonormal

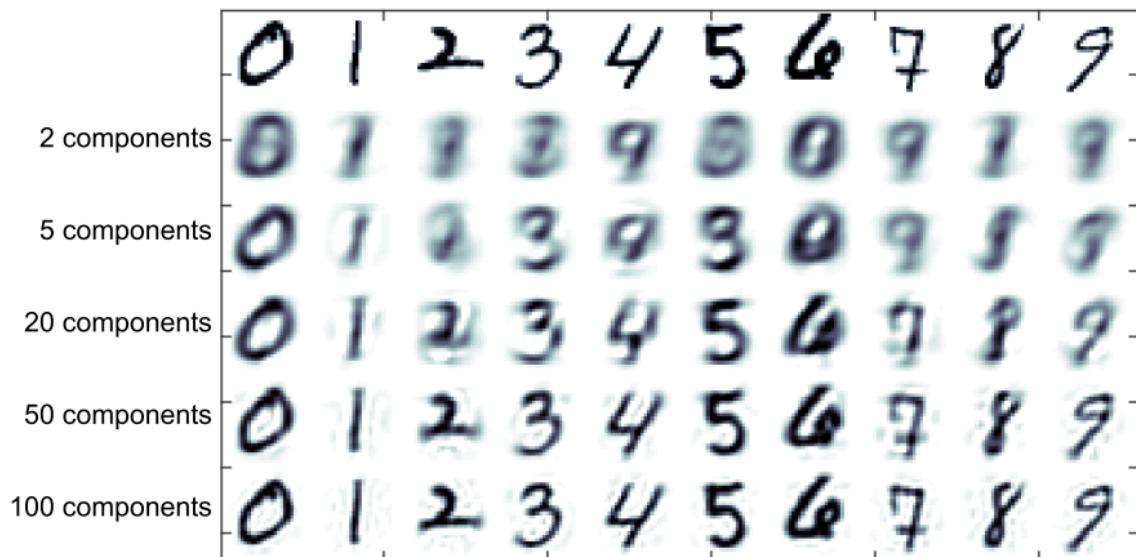
Visualization of hand written digits

...



PCA as compression

Only include a few components: $\hat{x}_i = Vb + m$ n=2,5,20,50,100



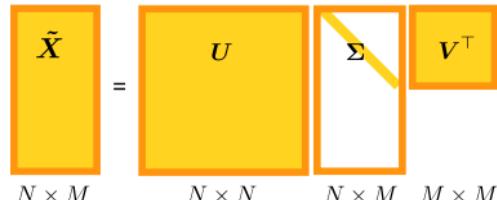
Data and domain driven feature extraction

PCA is an example of a data driven approach for feature extraction

i.e., we define from data the features extracted in terms of the projections $V^{(PCA)}$ that preserve most of the variance in the data

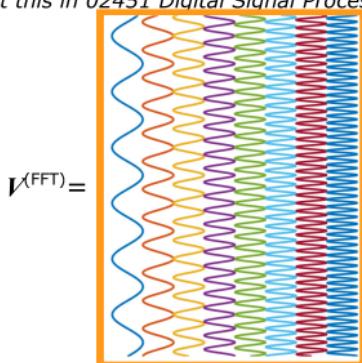
$$\tilde{X} = U \Sigma V^T$$

$N \times M$ $N \times N$ $N \times M$ $M \times M$



The fourier transform is an example of a domain driven approach for feature extraction

i.e., in the analysis of sound good features are often to use spectral representations. These can be found by projecting the data using the so-called fourier transform matrix $V^{(FFT)}$ where the components are defined as specific frequencies such that the projection of the data onto these frequencies defines the extend to which these frequencies are present in the data. (you can learn much more about this in 02451 Digital Signal Processing)



Resources

<http://www2.imm.dtu.dk> Our online PCA demo which highlights key concepts of PCA such as the effect of normalization, variance explained, and much more (<http://www2.imm.dtu.dk/courses/02450/DemoPCA.html>)

<https://arxiv.org> A great and more in-depth tutorial on PCA
(<https://arxiv.org/abs/1404.1100>)



<https://www.3blue1brown.com> An great, animated recap of linear algebra
(<https://www.3blue1brown.com/essence-of-linear-algebra-page/>)