UPPSALA
UNIVERSITET

Machine learning – Block 1(a)

Måns Magnusson
Department of Statistics, Uppsala University

Autumn 2025

This block

- What is AI and Machine Learning?
- Course Information and Practicalities
- Introduction to Supervised Learning
- (Stochastic) Gradient Descent
- Regularization

UPPSALA UNIVERSITET

Section 1

What is AI and ML?

What exactly is machine learning and artificial intelligence?

The word "AI" is often used quite loosely:

> To briefly explain how Linear Regression helped us reverse engineer the BSR equation, let's break it down. Linear Regression is an AI equation that finds the proper coefficients for an equation by sorting through massive amounts of data. The equation looks something like BSR = X(a)+ Y(b) + Z(c)..... and so and and so forth.

What is Artificial Intelligence?

Artificial intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, unlike the natural intelligence displayed by humans and animals. – Wikipedia

What is Artificial Intelligence?

Artificial intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, unlike the natural intelligence displayed by humans and animals. – Wikipedia Artificial intelligence (AI), the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings. – Encyclopedia Brittanica

What is Artificial General Intelligence?

Artificial general intelligence (AGI) is the hypothetical intelligence of a machine that has the capacity to understand or learn any intellectual task that a human being can. – Wikipedia

Also called:

1. Strong AI
2. General AI
3. Full AI

What is Artificial General Intelligence?

Artificial general intelligence (AGI) is the hypothetical intelligence of a machine that has the capacity to understand or learn any intellectual task that a human being can. – Wikipedia

Also called:

1. Strong AI
2. General AI
3. Full AI

Artificial super intelligence (ASI) is "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest" – Nick Bostrom

A Definition of AGI (Hendrycks et al., 2025)

- AGI: a model that matches or exceeds the **cognitive versatility and proficiency** of a well-educated adult.
- Cattell–Horn–Carroll (CHC) theory of human cognition.
- Evaluates ten broad cognitive domains:
  - Knowledge, Reading & Writing, Math, Reasoning
  - Working/Long-term Memory, Visual, Auditory, Speed
- Current models total score: GPT-4: 27%, GPT-5: 57%.

A Definition of AGI (Hendrycks et al., 2025)

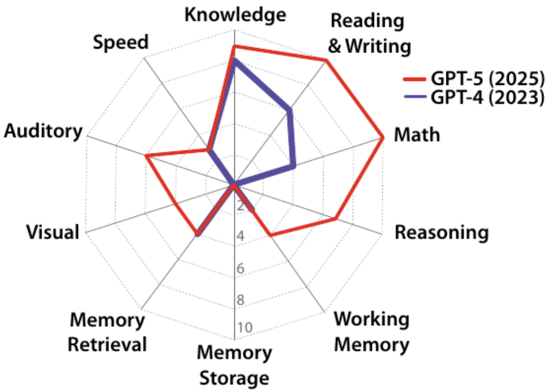


Figure 1: Cognitive abilities of GPT-4 vs GPT-5.

What is Artificial Super Intelligence?

Artificial super intelligence (ASI) is "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest" – Nick Bostrom

What is Machine Learning?

Machine Learning is the field of study that gives the computer the ability to learn without being explicitly programmed. – Arthur Samuel (1959)

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 Mitchell (1998)

What is Machine Learning?

Machine Learning is the field of study that gives the computer the ability to learn without being explicitly programmed. – Arthur Samuel (1959)

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 Mitchell (1998)

Learning from data. – Hastie, Tibshirani, Friedman (2009)

What is Machine Learning?

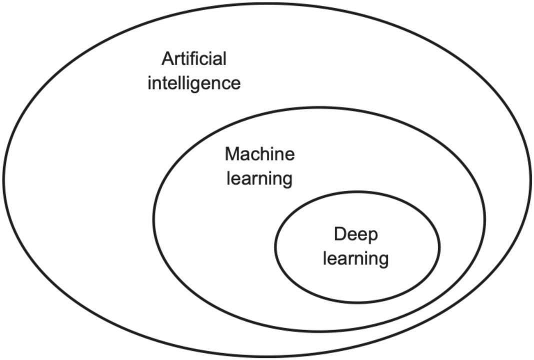
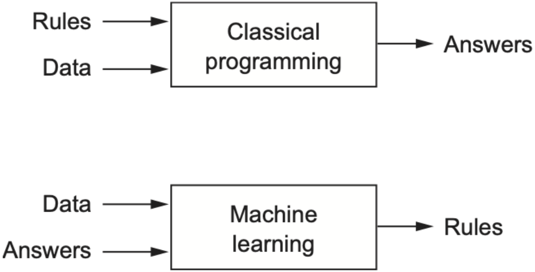


Figure: ML, AI and DL (Chollet, 2018, Figure 1.1)

Computer Science and Machine Learning

Figure: A new paradigm? (Chollet, 2018, Figure 1.2)



Statistics and Machine Learning

Figure: Regression vs. Pure Predictions (Efron, 2020, Table 5)

Table 5. A comparison checklist of differences between traditional regression methods and pure prediction algorithms.

	Traditional regressions methods	Pure prediction algorithms
1.	Surface plus noise models (continuous, smooth)	Direct prediction (possibly discrete, jagged)
2.	Scientific truth (long-term)	Empirical prediction accuracy (possibly short-term)
3.	Parametric modeling (causality)	Nonparametric (black box)
4.	Parsimonious modeling (researchers choose covariates)	Anti-parsimony (algorithm chooses predictors)
5.	$\mathbf{x}$ $p \times n$: with $p \ll n$ (homogeneous data)	$p \gg n$, both possibly enormous (mixed data)
6.	Theory of optimal inference (mle, Neyman–Pearson)	Training/test paradigm (Common Task Framework)

UPPSALA UNIVERSITET

Different names for the same things

- Machine learning has developed in parallel with Statistics
- Common with different names for the same thing:
  1. Time series classification (ML) vs. Functional data classification (Stats)

Different names for the same things

- Machine learning has developed in parallel with Statistics
- Common with different names for the same thing:
  1. Time series classification (ML) vs. Functional data classification (Stats)
  2. Time series regression (ML) vs. Scalar-on-function regression (Stats)

Different names for the same things

- Machine learning has developed in parallel with Statistics
- Common with different names for the same thing:
  1. Time series classification (ML) vs. Functional data classification (Stats)
  2. Time series regression (ML) vs. Scalar-on-function regression (Stats)
  3. Learning (ML) vs. Estimation (Stats)

Different names for the same things

UPPSALA UNIVERSITET

- Machine learning has developed in parallel with Statistics
- Common with different names for the same thing:
  1. Time series classification (ML) vs. Functional data classification (Stats)
  2. Time series regression (ML) vs. Scalar-on-function regression (Stats)
  3. Learning (ML) vs. Estimation (Stats)
  4. Weights (ML) vs. Parameters (Stats)

Different names for the same things

- Machine learning has developed in parallel with Statistics
- Common with different names for the same thing:
  1. Time series classification (ML) vs. Functional data classification (Stats)
  2. Time series regression (ML) vs. Scalar-on-function regression (Stats)
  3. Learning (ML) vs. Estimation (Stats)
  4. Weights (ML) vs. Parameters (Stats)
  5. Features (ML) vs. Covariates (Stats)

Different flavors of ML

- Supervised learning
- Unsupervised learning
- Reinforcement learning

The Bitter Lesson (Rich Sutton, 2019)

- Over 70 years of AI research show that *methods leveraging computation* outperform those using human knowledge.

The Bitter Lesson (Rich Sutton, 2019)

- Over 70 years of AI research show that *methods leveraging computation* outperform those using human knowledge.
- Researchers often try to build human understanding into systems — this helps short term, but limits long-term progress.

The Bitter Lesson (Rich Sutton, 2019)

- Over 70 years of AI research show that *methods leveraging computation* outperform those using human knowledge.
- Researchers often try to build human understanding into systems — this helps short term, but limits long-term progress.
- *Computation, search, and learning* scale with Moore's law — they keep improving as compute increases.

The Bitter Lesson (Rich Sutton, 2019)

- Over 70 years of AI research show that *methods leveraging computation* outperform those using human knowledge.
- Researchers often try to build human understanding into systems — this helps short term, but limits long-term progress.
- *Computation, search, and learning* scale with Moore's law — they keep improving as compute increases.
- Examples: Chess, Go, speech recognition, vision, language models — all advanced by scaling compute, not by encoding expert knowledge.

The Bitter Lesson (Rich Sutton, 2019)

- Over 70 years of AI research show that *methods leveraging computation* outperform those using human knowledge.
- Researchers often try to build human understanding into systems — this helps short term, but limits long-term progress.
- *Computation, search, and learning* scale with Moore's law — they keep improving as compute increases.
- Examples: Chess, Go, speech recognition, vision, language models — all advanced by scaling compute, not by encoding expert knowledge.
- The bitter truth: General-purpose, scalable methods win; human-designed structure eventually holds us back.

Questions?

UPPSALA
UNIVERSITET

Questions?

Section 2

Course information

Course information

The aims of this course are that you should:

Course information

The aims of this course are that you should:

1. get a good knowledge of a large number of machine learning models,
2. become able to use methods for evaluating and improving predictive models,
3. become able to handle big data,
4. become able to train and use machine learning models in R,
5. become able to train and use neural networks using Keras/TensorFlow.
6. become able to describe and discuss ethical aspects of big data and black box-models,

Course Outline

Two main parts:

- Core Content (8 lecture blocks):
  - Supervised learning (5 blocks)
    - Introduction, statistical learning (1 block)
    - Tree-based methods (1 block)
    - Neural Networks (3 block)
  - Unsupervised learning (2 blocks)
  - Reinforcement learning (1 block)

Course Outline

Two main parts:

- Core Content (8 lecture blocks):
  - Supervised learning (5 blocks)
    - Introduction, statistical learning (1 block)
    - Tree-based methods (1 block)
    - Neural Networks (3 block)
  - Unsupervised learning (2 blocks)
  - Reinforcement learning (1 block)
- Assignments (8 individual assignments)

Course Outline

Two main parts:

- Core Content (8 lecture blocks):
  - Supervised learning (5 blocks)
    - Introduction, statistical learning (1 block)
    - Tree-based methods (1 block)
    - Neural Networks (3 block)
  - Unsupervised learning (2 blocks)
  - Reinforcement learning (1 block)
- Assignments (8 individual assignments)
- Oral exam (Block 1-5) New for this year

Course Outline

Two main parts:

- Core Content (8 lecture blocks):
  - Supervised learning (5 blocks)
    - Introduction, statistical learning (1 block)
    - Tree-based methods (1 block)
    - Neural Networks (3 block)
  - Unsupervised learning (2 blocks)
  - Reinforcement learning (1 block)
- Assignments (8 individual assignments)
- Oral exam (Block 1-5) New for this year
- Mini-project on a supervised project (2-3 students)

Course Outline

Two main parts:

- Core Content (8 lecture blocks):
  - Supervised learning (5 blocks)
    - Introduction, statistical learning (1 block)
    - Tree-based methods (1 block)
    - Neural Networks (3 block)
  - Unsupervised learning (2 blocks)
  - Reinforcement learning (1 block)
- Assignments (8 individual assignments)
- Oral exam (Block 1-5) New for this year
- Mini-project on a supervised project (2-3 students)

Exact dates and details; see the course page.

Core Content

- Each block consist of:
  - Online video material (optional)
  - Reading assignments (approx. 2-4h, 50-90 pages a week)
  - One-two Lecture(s) (optional)
  - An individual computer assignment (approx. 14-16h).
  - Three Zoom computer lab sessions (optional)

UPPSALA UNIVERSITET

Core Content

- Each block consist of:
  - Online video material (optional)
  - Reading assignments (approx. 2-4h, 50-90 pages a week)
  - One-two Lecture(s) (optional)
  - An individual computer assignment (approx. 14-16h).
  - Three Zoom computer lab sessions (optional)
- Reading: Mandatory and optional (overlap)

Core Content

- Each block consist of:
  - Online video material (optional)
  - Reading assignments (approx. 2-4h, 50-90 pages a week)
  - One-two Lecture(s) (optional)
  - An individual computer assignment (approx. 14-16h).
  - Three Zoom computer lab sessions (optional)
- Reading: Mandatory and optional (overlap)
- Recommended workflow for each block
  - Do the reading assignments
  - Watch the videos (optional)
  - Attend the lecture (optional) to ask questions.
  - Do the computer assignment
  - Attend the zoom lab session (optional) to ask questions.

UPPSALA UNIVERSITET

Lectures

- Present overall theory, concepts and content (overview)

Lectures

- Present overall theory, concepts and content (overview)
- Ask questions during the lecture!
- Guest lectures on the course (worthwhile):
  1. Jonas Wallin, Lund University (regularization)

UPPSALA UNIVERSITET

Lectures

- Present overall theory, concepts and content (overview)
- Ask questions during the lecture!
- Guest lectures on the course (worthwhile):
  1. Jonas Wallin, Lund University (regularization)
  2. Erik Fredlund, CEO Codon AI (industry applications)

Lectures

- Present overall theory, concepts and content (overview)
- Ask questions during the lecture!
- Guest lectures on the course (worthwhile):
  1. Jonas Wallin, Lund University (regularization)
  2. Erik Fredlund, CEO Codon AI (industry applications)
  3. Holli Sargeant, Cambridge University (fairness and law)

Lectures

- Present overall theory, concepts and content (overview)
- Ask questions during the lecture!
- Guest lectures on the course (worthwhile):
  1. Jonas Wallin, Lund University (regularization)
  2. Erik Fredlund, CEO Codon AI (industry applications)
  3. Holli Sargeant, Cambridge University (fairness and law)
  4. Karim Jebari, The Institute for Futures Studies (AI and ethics)

Lectures

- Present overall theory, concepts and content (overview)
- Ask questions during the lecture!
- Guest lectures on the course (worthwhile):
  1. Jonas Wallin, Lund University (regularization)
  2. Erik Fredlund, CEO Codon AI (industry applications)
  3. Holli Sargeant, Cambridge University (fairness and law)
  4. Karim Jebari, The Institute for Futures Studies (AI and ethics)
  5. Väinö Yrjänäinen, UU (word embeddings)

Lectures

- Present overall theory, concepts and content (overview)
- Ask questions during the lecture!
- Guest lectures on the course (worthwhile):
  1. Jonas Wallin, Lund University (regularization)
  2. Erik Fredlund, CEO Codon AI (industry applications)
  3. Holli Sargeant, Cambridge University (fairness and law)
  4. Karim Jebari, The Institute for Futures Studies (AI and ethics)
  5. Väinö Yrjänäinen, UU (word embeddings)
- No lectures after 19th/20th of December

Examination

- To pass (G): All labs (at 75%), mini-project and oral exam needs to be passed

Examination

- To pass (G): All labs (at 75%), mini-project and oral exam needs to be passed
- To pass with distinction (VG): 7/10 VG points

Examination

- To pass (G): All labs (at 75%), mini-project and oral exam needs to be passed
- To pass with distinction (VG): 7/10 VG points
- Each assignment has an extra (VG) task worth 1 VG point.

Examination

- To pass (G): All labs (at 75%), mini-project and oral exam needs to be passed
- To pass with distinction (VG): 7/10 VG points
- Each assignment has an extra (VG) task worth 1 VG point.
- The mini-project is worth 2 VG-points (if it is passed with distinction).
- Ph.D. students: I suggest you get VG to pass the course. Make the project a potential paper.
- Reassesement of grades (supply form to course admin)
- Failing the course: You will need to redo all assignments, oral exam and mini-project.

UPPSALA UNIVERSITET

Computer Assignments

- Main part of the course
  Learning by doing
- Machine learning = Statistics + Computer Science
  Hence a lot of programming

UPPSALA UNIVERSITET

Computer Assignments

- Main part of the course
  Learning by doing
- Machine learning = Statistics + Computer Science
  Hence a lot of programming
- Both implementation of core components and state-of-the-art methods

Computer Assignments

- Main part of the course
  Learning by doing
- Machine learning = Statistics + Computer Science
  Hence a lot of programming
- Both implementation of core components and state-of-the-art methods
- *Warning!* There might be bugs in the assignments! Don't hesitate to ask questions!

UPPSALA UNIVERSITET

UPPSALA UNIVERSITET

Computer Assignments

- Main part of the course
  Learning by doing
- Machine learning = Statistics + Computer Science
  Hence a lot of programming
- Both implementation of core components and state-of-the-art methods
- *Warning!* There might be bugs in the assignments! Don't hesitate to ask questions!
- Deadline Sundays 23.59.

Computer Assignments

- Main part of the course
  Learning by doing
- Machine learning = Statistics + Computer Science
  Hence a lot of programming
- Both implementation of core components and state-of-the-art methods
- *Warning!* There might be bugs in the assignments! Don't hesitate to ask questions!
- Deadline Sundays 23.59.
- All assignments can be turned in a three times. 2nd deadline last day of course. 3rd deadline approx 2-4 weeks after the course. If failed, resubmit right away!

Computer Assignments

- Main part of the course
  Learning by doing
- Machine learning = Statistics + Computer Science
  Hence a lot of programming
- Both implementation of core components and state-of-the-art methods
- *Warning!* There might be bugs in the assignments! Don't hesitate to ask questions!
- Deadline Sundays 23.59.
- All assignments can be turned in a three times. 2nd deadline last day of course. 3rd deadline approx 2-4 weeks after the course. If failed, resubmit right away!
- We will mark and return each assignment within 10 working days.

Computer Assignments

- Don't write your name anywhere!

Computer Assignments

- Don't write your name anywhere!
- Do the assignment evaluation

Computer Assignments

- Don't write your name anywhere!
- Do the assignment evaluation
- *Important!* Don't show your assignment to any other student. But feel free to discuss!

Computer Assignments

- Don't write your name anywhere!
- Do the assignment evaluation
- *Important!* Don't show your assignment to any other student. But feel free to discuss!
- Zoom sessions:
  1. First lab each week will include a 15 min introduction
  2. Our focus: Help during computer labs - less focus on written feedback
  3. Ask questions! This is your time.

Computer Assignments

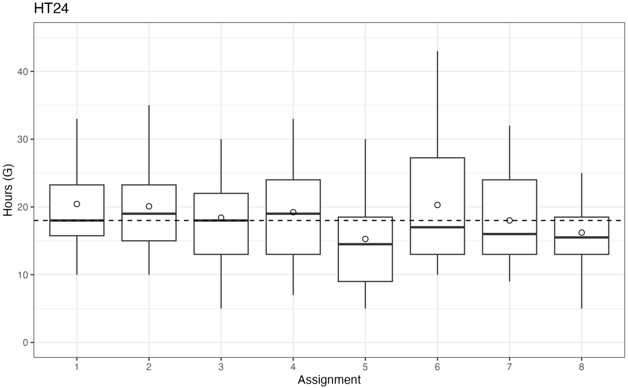
- Don't write your name anywhere!
- Do the assignment evaluation
- *Important!* Don't show your assignment to any other student. But feel free to discuss!
- Zoom sessions:
  1. First lab each week will include a 15 min introduction
  2. Our focus: Help during computer labs - less focus on written feedback
  3. Ask questions! This is your time.

UPPSALA UNIVERSITET

- What is AI and ML?
- Course information
- Introduction to Supervised Learning
  - Example: Logistic regression
- Optimization in Machine Learning

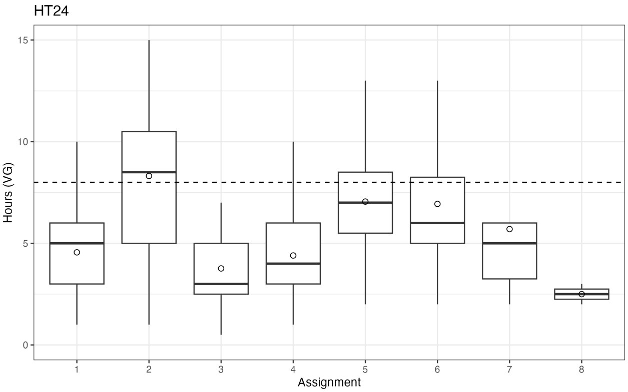
Computer Assignments

Figure: Workload last year to pass (G)



Computer Assignments

Figure: Workload last year to pass with distinction (VG)



- What is AI and ML?
- Course information
- Introduction to Supervised Learning
  - Example: Logistic regression
- Optimization in Machine Learning

Oral Exam

- New this year. Purpose: Additional exam of assignments (ChatGPT proof)

Oral Exam

- New this year. Purpose: Additional exam of assignments (ChatGPT proof)
- Approx 60-70 questions based on the assignment 1-5 and the reading assignments.
- If a question is not covered by the material - reach out!
- The 9th of November 13.15-16.00 H319
- Random student will answer 1-2 random questions.
- Pass/Fail at the spot.
- Examiner Måns and Hannes.

Mini-project

- See project instructions on webpage for details.

Mini-project

- See project instructions on webpage for details.
- Supervised problem of choice on real data.
- 2-3 students.

Mini-project

- See project instructions on webpage for details.
- Supervised problem of choice on real data.
- 2-3 students.
- Supply step 1 proposal at the end of block 3.

UPPSALA UNIVERSITET

- What is AI and ML?
- Course information
- Introduction to Supervised Learning
  — Example: Logistic regression
- Optimization in Machine Learning

Mini-project

- See project instructions on webpage for details.
- Supervised problem of choice on real data.
- 2-3 students.
- Supply step 1 proposal at the end of block 3.
- Supply step 2 proposal at the end of block 6.

Mini-project

- See project instructions on webpage for details.
- Supervised problem of choice on real data.
- 2-3 students.
- Supply step 1 proposal at the end of block 3.
- Supply step 2 proposal at the end of block 6.
- Project will last two weeks (half time) - but start earlier. Good case to show potential employers.
- Recommended data: Images, text or tabular data (e.g. avoid time series).
- Feel free to build upon the Bayesian project (e.g. compare with Bayesian methods).
- Approximate 40 hours of work *per student*.

UPPSALA UNIVERSITET

Mini-project

- See project instructions on webpage for details.
- Supervised problem of choice on real data.
- 2-3 students.
- Supply step 1 proposal at the end of block 3.
- Supply step 2 proposal at the end of block 6.
- Project will last two weeks (half time) - but start earlier. Good case to show potential employers.
- Recommended data: Images, text or tabular data (e.g. avoid time series).
- Feel free to build upon the Bayesian project (e.g. compare with Bayesian methods).
- Approximate 40 hours of work *per student*.
- The project should result in a 4 page report (PDF) using the ICML LaTeX template (see course page).
- Project oral presentation (10-15 minutes)

Mini-project

- See project instructions on webpage for details.
- Supervised problem of choice on real data.
- 2-3 students.
- Supply step 1 proposal at the end of block 3.
- Supply step 2 proposal at the end of block 6.
- Project will last two weeks (half time) - but start earlier. Good case to show potential employers.
- Recommended data: Images, text or tabular data (e.g. avoid time series).
- Feel free to build upon the Bayesian project (e.g. compare with Bayesian methods).
- Approximate 40 hours of work *per student*.
- The project should result in a 4 page report (PDF) using the ICML LaTeX template (see course page).
- Project oral presentation (10-15 minutes)
- The first author is corresponding author

Mini-project

- See project instructions on webpage for details.
- Supervised problem of choice on real data.
- 2-3 students.
- Supply step 1 proposal at the end of block 3.
- Supply step 2 proposal at the end of block 6.
- Project will last two weeks (half time) - but start earlier. Good case to show potential employers.
- Recommended data: Images, text or tabular data (e.g. avoid time series).
- Feel free to build upon the Bayesian project (e.g. compare with Bayesian methods).
- Approximate 40 hours of work *per student*.
- The project should result in a 4 page report (PDF) using the ICML LaTeX template (see course page).
- Project oral presentation (10-15 minutes)
- The first author is corresponding author
- Mini-project and master thesis:
  - The mini-project can be used to explore thesis project

UPPSALA UNIVERSITET

Practicalities

- Course page: Github – please do a PR if something is wrong!

UPPSALA UNIVERSITET

- What is AI and ML?
- Course information
- Introduction to Supervised Learning
  - Example: Logistic regression
- Optimization in Machine Learning

Practicalities

- Course page: Github – please do a PR if something is wrong!
- Acknowledgements: Måns Thulin, Josef Wilzén, Anders Eklund

UPPSALA UNIVERSITET

Practicalities

- Course page: Github – please do a PR if something is wrong!
- Acknowledgements: Måns Thulin, Josef Wilzén, Anders Eklund
- Schedule: Time Edit/Studium
- Assignments: Studium

Practicalities

- Course page: Github – please do a PR if something is wrong!
- Acknowledgements: Måns Thulin, Josef Wilzén, Anders Eklund
- Schedule: Time Edit/Studium
- Assignments: Studium
- Literature
  - Bishop & Bishop (2024) *Deep Learning - Foundations and Concepts*.
  - Hastie, Tibshirani & Friedman (2009). *Elements of Statistical Learning*.
  - Sutton and Barto (2020) *Reinforcement learning: An introduction*
  - Chollet & Allaire (2018) *Deep Learning with R* (optional)
  - Additional articles, tuturials, videos etc. posted on course (github) homepage
  - Mandatory and optional material: Overlap exists!

Practicalities

- Course page: Github – please do a PR if something is wrong!
- Acknowledgements: Måns Thulin, Josef Wilzén, Anders Eklund
- Schedule: Time Edit/Studium
- Assignments: Studium
- Literature
  - Bishop & Bishop (2024) *Deep Learning - Foundations and Concepts*.
  - Hastie, Tibshirani & Friedman (2009). *Elements of Statistical Learning*.
  - Sutton and Barto (2020) *Reinforcement learning: An introduction*
  - Chollet & Allaire (2018) *Deep Learning with R* (optional)
  - Additional articles, tuturials, videos etc. posted on course (github) homepage
  - Mandatory and optional material: Overlap exists!
- If you have complaint - reach out to me in assignment evaluations!

Practicalities

- Course page: Github – please do a PR if something is wrong!
- Acknowledgements: Måns Thulin, Josef Wilzén, Anders Eklund
- Schedule: Time Edit/Studium
- Assignments: Studium
- Literature
  - Bishop & Bishop (2024) *Deep Learning - Foundations and Concepts*.
  - Hastie, Tibshirani & Friedman (2009). *Elements of Statistical Learning*.
  - Sutton and Barto (2020) *Reinforcement learning: An introduction*
  - Chollet & Allaire (2018) *Deep Learning with R* (optional)
  - Additional articles, tuturials, videos etc. posted on course (github) homepage
  - Mandatory and optional material: Overlap exists!
- If you have complaint - reach out to me in assignment evaluations!
- Notation and literature is difficult... hence a mess.

Course improvements since last year

- Oral exam (I hope it works!)
- New lecture (maybe) on diffusion models

Questions?

Questions?

Section 3

Introduction to Supervised Learning

Supervised learning

Figure: Relationship between appartment size and price (source)

Problem: We want to predict the price of a new apartment.

Supervised learning

- General problem: We have *training* data

$$\mathbf{d} = \{(y_i, \mathbf{x}_i), i = 1, ..., n\}.$$

- $\mathbf{x}_i$ = features/input/predictors/features/independent variables
- y_i = labels/output/dependent variable
- We want to *learn* a function $\hat{y} = f(x_{new})$ with as good performance as possible.

Supervised learning

- General problem: We have *training* data

$$\mathbf{d} = \{(y_i, \mathbf{x}_i), i = 1, ..., n\}.$$

- $\mathbf{x}_i$ = features/input/predictors/features/independent variables
- y_i = labels/output/dependent variable
- We want to *learn* a function $\hat{y} = f(x_{new})$ with as good performance as possible.
- Regression problems: $y_i \in \mathbb{R}$
- Classification problems: $y_i \in a, b, c, ...$ where $a, b, c...$ are discrete classes.

Example of supervised problems

Any examples of applications?

Example of supervised problems

Any examples of applications?

- Is this e-mail message spam or not?

UPPSALA
UNIVERSITET

Example of supervised problems

Any examples of applications?

- Is this e-mail message spam or not?
- Image recognition/classification

Example of supervised problems

Any examples of applications?

- Is this e-mail message spam or not?
- Image recognition/classification
- Image object traction (position in a video)

Example of supervised problems

Any examples of applications?

- Is this e-mail message spam or not?
- Image recognition/classification
- Image object traction (position in a video)
- Will this patient recover from their illness or not?

Example of supervised problems

Any examples of applications?

- Is this e-mail message spam or not?
- Image recognition/classification
- Image object traction (position in a video)
- Will this patient recover from their illness or not?
- Does this fingerprint belong to an employee or not?

Example of supervised problems

Any examples of applications?

- Is this e-mail message spam or not?
- Image recognition/classification
- Image object traction (position in a video)
- Will this patient recover from their illness or not?
- Does this fingerprint belong to an employee or not?
- Does this customer have stable finances or not?

Example of supervised problems

Any examples of applications?

- Is this e-mail message spam or not?
- Image recognition/classification
- Image object traction (position in a video)
- Will this patient recover from their illness or not?
- Does this fingerprint belong to an employee or not?
- Does this customer have stable finances or not?
- Face recognition

Example of supervised problems

Any examples of applications?

- Is this e-mail message spam or not?
- Image recognition/classification
- Image object traction (position in a video)
- Will this patient recover from their illness or not?
- Does this fingerprint belong to an employee or not?
- Does this customer have stable finances or not?
- Face recognition
- Is this tumour malign or not?

Example of supervised problems

Any examples of applications?

- Is this e-mail message spam or not?
- Image recognition/classification
- Image object traction (position in a video)
- Will this patient recover from their illness or not?
- Does this fingerprint belong to an employee or not?
- Does this customer have stable finances or not?
- Face recognition
- Is this tumour malign or not?

Logistic regression and classification

When the y_i in a regression problem is binary (or more generally, categorical), it becomes a classification problem.

Logistic regression and classification

When the y_i in a regression problem is binary (or more generally, categorical), it becomes a classification problem.

The question that the model tries to answer is: does this observation belong to class 0 or class 1?

Logistic regression and classification

When the y_i in a regression problem is binary (or more generally, categorical), it becomes a classification problem.

The question that the model tries to answer is: does this observation belong to class 0 or class 1?

Logistic regression is a workhorse for classification problems.

UPPSALA UNIVERSITET

Logistic regression

When analysing binary data $y_1, \ldots, y_N$, we usually assume that the Y_i follow binomial (or Bernoulli) distributions.

Logistic regression

When analysing binary data $y_1, \ldots, y_N$, we usually assume that the Y_i follow binomial (or Bernoulli) distributions.

Assume that $Y_1, \ldots, Y_N$ are independent with $Y_i \sim Bernoulli(\pi_i)$.

Logistic regression

When analysing binary data $y_1, \dots, y_N$, we usually assume that the Y_i follow binomial (or Bernoulli) distributions.

Assume that $Y_1, \dots, Y_N$ are independent with $Y_i \sim Bernoulli(\pi_i)$.

$Y_i \in 0, 1$ with success probability π_i and $\mu_i = E(Y_i) = \pi_i$.

Logistic regression

When analysing binary data $y_1, \ldots, y_N$, we usually assume that the Y_i follow binomial (or Bernoulli) distributions.

Assume that $Y_1, \ldots, Y_N$ are independent with $Y_i \sim Bernoulli(\pi_i)$.

$Y_i \in 0, 1$ with success probability π_i and $\mu_i = E(Y_i) = \pi_i$.

- The natural parameter of the binomial distribution is

$$g(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right),$$

called the logit or log odds.

Logistic regression

When analysing binary data $y_1, \ldots, y_N$, we usually assume that the Y_i follow binomial (or Bernoulli) distributions.

Assume that $Y_1, \ldots, Y_N$ are independent with $Y_i \sim Bernoulli(\pi_i)$.

$Y_i \in 0, 1$ with success probability π_i and $\mu_i = E(Y_i) = \pi_i$.

- The natural parameter of the binomial distribution is

$$g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right),$$

  called the logit or log odds.

- A GLM using this link function is called logistic regression, but other link functions are also often used in practice. Many times we use likelihood functions

Logistic regression

There are two equivalent formulas for logistic regression:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}, \qquad i = 1,\ldots,N$$

and

$$\pi_i = \frac{\exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}{1+\exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}.$$

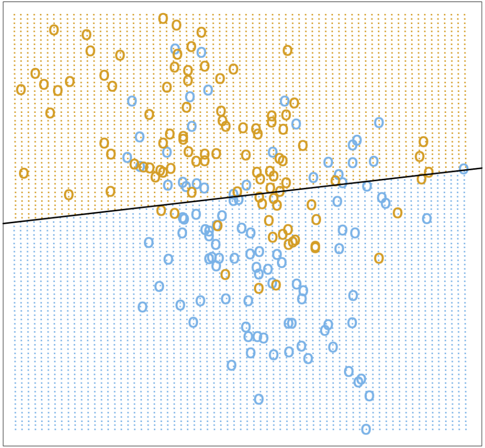
Logistic regression: Prediction

- We *train* a logistic regression model using MLE using the training data.
- Our estimation/traing output the MLE $\hat{\theta}$
- We the compute $\hat{p}_i = g^{-1}(\hat{\theta} x_{new})$ a for a new observation
- We use a decision rule to predict value 0 or 1:

$$\hat{y}_i(\hat{p}_i) = \begin{cases} 1, & \text{if } \hat{p}_i \geq 0.5 \\ 0, & \text{otherwise} \end{cases}$$

Logistic regression: Example

Figure: Decision boundry with two covariates (Hastie et al, 2009, Figure 2.1)



An example: E-mail classification

E-mail Spam
Predict if a new, unseen, e-mail is spam or ham.

UPPSALA UNIVERSITET

Questions?

Questions?

Section 4

Optimization in Machine Learning

Training of ML algorithms

1. Training is usually done by minimizing the objective/loss/cost function $L(\theta)$ for $\theta \in \mathbf{R}^P$.
2. Example: Logistic regression, here we can use the negative log-likelihood as loss function:

$$L(\theta, \mathbf{y}, \mathbf{X}) = -\log \prod_{i=1}^{N} p_i^{y_i} (1 - p_i)^{1-y_i} \,,$$

where

$$\log \frac{p_i}{1 - p_i} = \mathbf{x}_i \theta \,,$$

Training of ML algorithms

1. Training is usually done by minimizing the objective/loss/cost function $L(\theta)$ for $\theta \in \mathbf{R}^P$.
2. Example: Logistic regression, here we can use the negative log-likelihood as loss function:

$$L(\theta, \mathbf{y}, \mathbf{X}) = -\log \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i} \,,$$

where

$$\log \frac{p_i}{1-p_i} = \mathbf{x}_i \theta \,,$$

3. In Machine Learning: P and N might be very large...

Gradient Decent

1. The workhorse of Machine Learning

$$\theta_t = \theta_{t-1} - \eta \nabla L(\theta_{t-1}, \mathbf{X}, \mathbf{y}),$$

   where

$$\nabla f(p) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(p) \\ \vdots \\ \frac{\partial f}{\partial x_n}(p) \end{bmatrix}$$

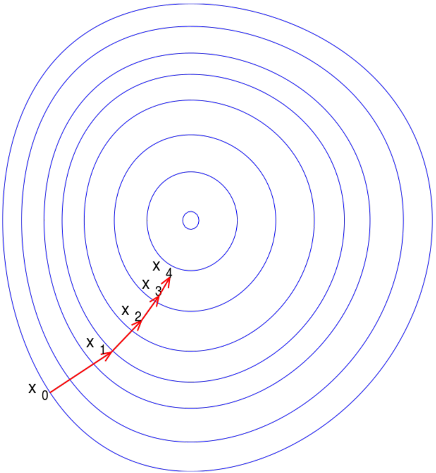
2. $L(\theta)$ needs to be differentiable

Gradient Descent Analogy

Figure: Gradient Descent Analogy (source)

Gradient Descent (cont.)

Figure: Gradient Descent (source)



UPPSALA UNIVERSITET

Why Gradient Descent?

- Gradient Descent is a poor algorithm (Newtons method, Iteratively Reweighted Least Squares are 'better')
- So why is gradient descent relevant?

Why Gradient Descent?

- Gradient Descent is a poor algorithm (Newtons method, Iteratively Reweighted Least Squares are 'better')
- So why is gradient descent relevant?
- The two benefits with Gradient Descent:
  1. Only uses the gradient—scales to large P
  2. Can scale to large data with Stochastic Gradient Descent—scales to large N

Stochastic Gradient Descent

- Many loss functions (and gradients) are a sum over N observations (e.g. log-likelihoods).

UPPSALA UNIVERSITET

Stochastic Gradient Descent

- Many loss functions (and gradients) are a sum over N observations (e.g. log-likelihoods).
- We can estimate $\nabla L(\theta, X_i, y_i)$ by choosing a random observation (with index i)

$$E(\nabla L(\theta, X_i, y_i)) = \frac{1}{Z}\nabla L(\theta, \mathbf{X}, \mathbf{y}),$$

for some constant Z.

Stochastic Gradient Descent

- This give us the following algorithm:

$$\theta_t = \theta_{t-1} - \eta_t \hat{\nabla} L(\theta_{t-1}, X_i, y_i) \,,$$

  where i is random sampled index.

- *Note!*
  We need to have an unbiased estimator for $\nabla L(\theta, \mathbf{X}, \mathbf{y})$

Stochastic Gradient Descent

UPPSALA UNIVERSITET

- This give us the following algorithm:

$$\theta_t = \theta_{t-1} - \eta_t \hat{\nabla} L(\theta_{t-1}, X_i, y_i)\,,$$

  where i is random sampled index.
- *Note!*
  We need to have an unbiased estimator for $\nabla L(\theta, \mathbf{X}, \mathbf{y})$
- What is an iteration?

Stochastic Gradient Descent

UPPSALA UNIVERSITET

- This give us the following algorithm:

$$\theta_t = \theta_{t-1} - \eta_t \hat{\nabla} L(\theta_{t-1}, X_i, y_i)\,,$$

  where i is random sampled index.
- *Note!*
  We need to have an unbiased estimator for $\nabla L(\theta, \mathbf{X}, \mathbf{y})$
- What is an iteration?
- Epochs vs. Iterations

Stochastic Gradient Descent

- Learning rate η_t is important
- Will it converge to an optimum?

UPPSALA UNIVERSITET

Stochastic Gradient Descent

- Learning rate η_t is important
- Will it converge to an optimum?
- We need to reduce η_t over time

Stochastic Gradient Descent

- Learning rate η_t is important
- Will it converge to an optimum?
- We need to reduce η_t over time
- Robbins–Monro (1951) conditions:
  1. $\eta_t \geq 0 \; \forall t \geq 0$
  2. $\sum_t^\infty \eta_t = \infty$
  3. $\sum_t^\infty \eta_t^2 < \infty$

Stochastic Gradient Descent

- Learning rate η_t is important
- Will it converge to an optimum?
- We need to reduce η_t over time
- Robbins–Monro (1951) conditions:
  1. $\eta_t \geq 0 \; \forall t \geq 0$
  2. $\sum_t^\infty \eta_t = \infty$
  3. $\sum_t^\infty \eta_t^2 < \infty$

A simple solution is:

$$\eta_t = \frac{1}{t}, \qquad (t \geq 1)$$

UPPSALA UNIVERSITET

Mini-batch gradient descent

- Can we estimate the gradient in a better way?

Mini-batch gradient descent

- Can we estimate the gradient in a better way?
- We take a mini-batch of size B:

$$\theta_t = \theta_{t-1} - \eta_t \nabla L(\theta, \mathbf{X}_{(S)_i}, y_{(S)_i}) \,,$$

where $(S)_i$ is a set of random sample (without replacement) indices and $|(S)_i| = B$.

Mini-batch gradient descent

- Can we estimate the gradient in a better way?
- We take a mini-batch of size B:

$$\theta_t = \theta_{t-1} - \eta_t \nabla L(\theta, \mathbf{X}_{(S)_i}, y_{(S)_i}),$$

where $(S)_i$ is a set of random sample (without replacement) indices and $|(S)_i| = B$.

- B is usually set to optimize hardware

UPPSALA UNIVERSITET

SGD with momentum

- SGD can be slow to converge due to 'jumping' behaviour

SGD with momentum

- SGD can be slow to converge due to 'jumping' behaviour
- Can improve behaviour using momentum – the rolling mean of gradients

UPPSALA UNIVERSITET

SGD with momentum

- SGD can be slow to converge due to 'jumping' behaviour
- Can improve behaviour using momentum – the rolling mean of gradients
- Additional hyperparemeter α to control the momentum

$$m_t = \alpha m_{t-1} + \eta_t \hat{\nabla} L(\theta_{t-1}, X_i, y_i)\,,$$

$$\theta_t = \theta_{t-1} - m_t,$$

Figure: SGD with momentum

(a) SGD without momentum (b) SGD with momentum

SGD with momentum, Intuition

Figure: SGD with momentum, Intuition (CC)

UPPSALA UNIVERSITET

- What is AI and ML?
- Course information
- Introduction to Supervised Learning
  - Example: Logistic regression
- Optimization in Machine Learning

SGD with momentum

Example of SGD with momentum here.

Adaptive Moment Estimation (Adam)

UPPSALA UNIVERSITET

- What is AI and ML?
- Course information
- Introduction to Supervised Learning
  - Example: Logistic regression
- Optimization in Machine Learning

- Adapt η_t to individual parameters

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1)\hat{\nabla} L(\theta_{t-1}, X_i, y_i)$$
$$v_t = \beta_2 v_{t-1} + (1 - \beta_2)\hat{\nabla} L(\theta_{t-1}, X_i, y_i)^2$$

- Bias correction
  (due to initialization at 0)

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}$$
$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

- Update

$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon}\hat{m}_t,$$

- Common values: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$

Adaptive Moment Estimation (Adam)

- Adapt η_t to individual parameters

$$
\begin{aligned}
m_t =& \beta_1 m_{t-1} + (1-\beta_1)\hat{\nabla} L(\theta_{t-1}, X_i, y_i) \\
v_t =& \beta_2 v_{t-1} + (1-\beta_2)\hat{\nabla} L(\theta_{t-1}, X_i, y_i)^2
\end{aligned}
$$

- Bias correction
  (due to initialization at 0)

$$
\begin{aligned}
\hat{m}_t =& \frac{m_t}{1-\beta_1^t} \\
\hat{v}_t =& \frac{v_t}{1-\beta_2^t}
\end{aligned}
$$

- Update

$$
\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon}\hat{m}_t,
$$

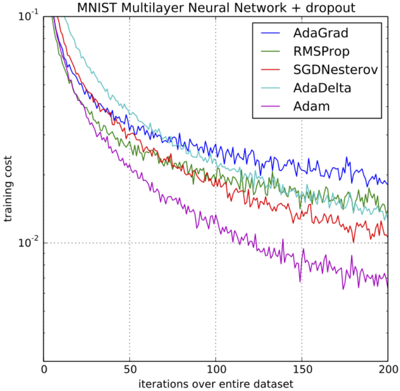
- Common values: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$
- RMSprop is another (similar) alternative

UPPSALA UNIVERSITET

- What is AI and ML?
- Course information
- Introduction to Supervised Learning
  - Example: Logistic regression
- Optimization in Machine Learning

Adam

Figure: The Adam Optimizer (Kingma and Ba, 2014)



For convergence proofs, see:
Defossez et al (2020) "A Simple Convergence Proof of Adam and Adagrad"