



UPPSALA  
UNIVERSITET

# Machine learning, big data and artificial intelligence – Block 1

Måns Magnusson  
Department of Statistics, Uppsala University

November 2020

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



UPPSALA  
UNIVERSITET

# This week's lectures

---

- What is AI and ML?
  - Course information
  - Introduction to Supervised Learning
    - Example: Logistic regression
  - Optimization Algorithms for Machine Learning
- What is AI and Machine Learning?
  - Course Information and Practicalities
  - Introduction to Supervised Learning
  - (Stochastic) Gradient Descent



UPPSALA  
UNIVERSITET

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

## 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 so forth.

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



# What is Artificial Intelligence?

---

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

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 Britannica



# 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 AI and ML?
- Course information
- Introduction to Supervised Learning
  - Example: Logistic regression
- Optimization Algorithms for Machine Learning



# What is Machine Learning?

---

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

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

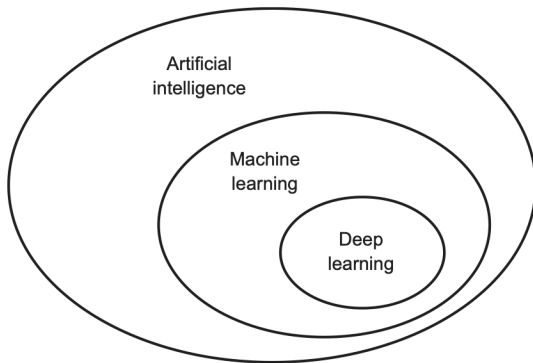
A computer program is said to learn from experience  $E$  with respect to some class of tasks  $T$  and performance measure  $P$ , if its performance at tasks in  $T$ , as measured by  $P$ , improves with experience  $E$ . – Tom Mitchell (1998)

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



# What is Machine Learning?

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



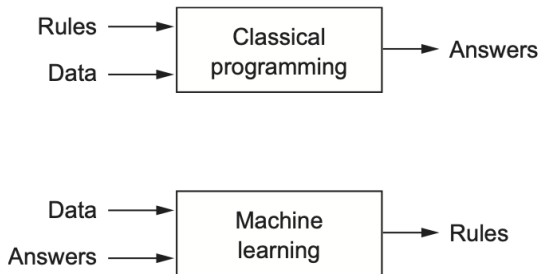
- What is AI and ML?
- Course information
- Introduction to Supervised Learning
  - Example: Logistic regression
- Optimization Algorithms for Machine Learning





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

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





**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)

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



UPPSALA  
UNIVERSITET

# Different flavors of ML

---

- What is AI and ML?
  - Course information
  - Introduction to Supervised Learning
    - Example: Logistic regression
  - Optimization Algorithms for Machine Learning
- Supervised learning
  - Unsupervised learning
    - Self-(un)supervised learning
  - Reinforcement learning



UPPSALA  
UNIVERSITET

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

## Section 2

### Course information



UPPSALA  
UNIVERSITET

## Course information

---

The aims of this course are that you should:

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

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,



UPPSALA  
UNIVERSITET

# Course Outline

---

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

Two main parts:

- Core Content (8 blocks):
  - Supervised learning (5 blocks)
  - Unsupervised learning (2 blocks)
  - Reinforcement learning (1 block)
- Mini-project on a supervised project (2-3 students)

Exact dates and details; see the course page.



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

- Two lectures/computer labs (approx. 4h)
- Online video material and reading assignments (approx. 4-6h, 50-90 pages a week)
- *Note!* There might be some overlap between reading instructions.
- An individual computer assignment (approx. 12-16h). Deadline Sundays 23.59.
- Recommended workflow for each block
  - Watch the videos (although, optional)
  - Do the reading assignments
  - Do the computer assignment



UPPSALA  
UNIVERSITET

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

# Computer Assignments

---

1. Main part of the course  
Learning by doing
2. Machine learning = Statistics + Computer Science  
Hence a lot of programming
3. Both implementation of core components and  
state-of-the-art methods
4. *Warning!* There might be bugs in the assignments!
5. All labs can be turned in a second time. Deadline 17th of  
January.
6. *Important!* Do not write your name anywhere and do the  
assignment evaluation!





UPPSALA  
UNIVERSITET

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

## Mini-project

---

- See project instructions on webpage for details.
- Supervised problem of choice on real data.
- 2-3 students.
- Supply a half-page project proposal of data and problem at the end of block 6.
- Project will last two weeks (half time) - but start earlier.
- Approximate 40 hours of work *per student*.
- The project should result in a 4 page report (PDF) using the ICML LaTeX template.
- Project oral presentation (10-15 minutes)
- Each student will review one other project:
  1. Write a 1-2 page review of the report.
  2. Discuss at the oral presentations.



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

## Practicalities

---

- Course page: Github – please do a PR if something is wrong!
- Communication: Course Slack
- Acknowledgements: Måns Thulin, Josef Wilzén, Anders Eklund
- Schedule: Time Edit/Studium
- Assignments: Studium
- Literature
  - Hastie, Tibshirani & Friedman (2009). *Elements of Statistical Learning*. Springer. Available at <https://web.stanford.edu/~hastie/ElemStatLearn/>
  - Chollet & Allaire (2018). *Deep Learning with R*. Manning. Available for reading at <https://www.manning.com/books/deep-learning-with-r>
  - Goodfellow, Bengio & Courville (2017). *Deep Learning*. Available at <https://www.deeplearningbook.org/>
  - Additional articles, tutorials, videos etc. posted on course (github) homepage



UPPSALA  
UNIVERSITET

# Examination

---

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

1. To pass (G): All labs, mini-project, and project review need to be passed
2. To pass with distinction (VG): 7/10 VG points
3. Each assignment has an extra (VG) task worth 1 VG point.
4. The mini-project is worth 2 VG-points (if it is passed with distinction).
5. Ph.D. students: I suggest you get VG to pass the course. Make the project a potential paper.



UPPSALA  
UNIVERSITET

# Course improvements since last time

---

- What is AI and ML?
  - **Course information**
  - Introduction to Supervised Learning
    - Example: Logistic regression
  - Optimization Algorithms for Machine Learning
- Optional VG assignments and clearer info on how to get VG
  - More even burden (hopefully)
  - General updates of unclear parts in assignments
  - General question in each assignment (4-5 paragraphs)
  - *Note!* There can be some new bugs after these updates.



UPPSALA  
UNIVERSITET

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

## Section 3

# Introduction to Supervised Learning



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

# Supervised learning

Figure: Relationship between apartment size and price ([source](#))



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



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

- General problem: We have *training* data

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

- $\mathbf{x}_i$  = 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:  $y_i \in a, b, c, \dots$  where  $a, b, c, \dots$  are discrete classes.



# Example of supervised problems

---

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

- Is this e-mail message spam (1) or not (0)?
- 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 (1) or not (0)?





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

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 in classification problems.



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

# 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  $Y_i \sim \text{Bernoulli}(\pi_i)$ .

$Y_i \in 0, 1$  with success probability  $\pi_i$  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.



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

# Logistic regression

---

There are two equivalent formulas for **logistic regression**:

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

and

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



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

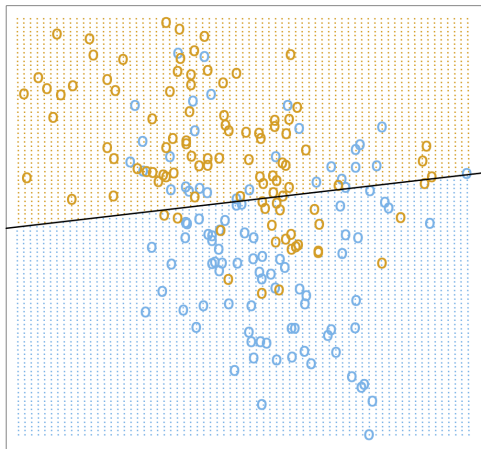
- We *train* a logistic regression model using MLE using the training data.
- Our estimation/training output the MLE  $\hat{\theta}$
- We then compute  $\hat{p}_i = g^{-1}(\hat{\theta}x_{new})$  for a new observation
- We use a **decision rule** to predict value 0 or 1:

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



# Logistic regression: Example

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



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



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

# An example: Spam and Ham

## E-mail Spam

An e-mail provider what to help classify e-mails as spam (1) or ham (0). They have many previous e-mails that customers have already classified as spam, and e-mails people have responded (ham). They want to predict if a new, unseen e-mail is spam or ham.





UPPSALA  
UNIVERSITET

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

## Section 4

# Optimization Algorithms for Machine Learning



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

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

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





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

# 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



UPPSALA  
UNIVERSITET

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

# Gradient Descent Analogy

---

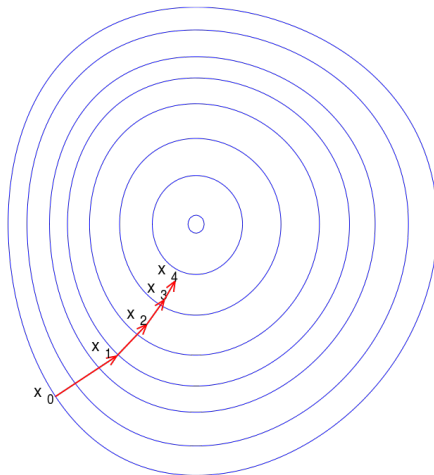
Figure: Gradient Descent Analogy ([source](#))





# Gradient Descent (cont.)

Figure: Gradient Descent ([source](#))



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



# Why Gradient Descent?

---

- What is AI and ML?
  - Course information
  - Introduction to Supervised Learning
    - Example: Logistic regression
  - Optimization Algorithms for Machine Learning
- Gradient Descent is a poor algorithm (Newton's 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  $N$



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

# Stochastic Gradient Descent

- Many loss functions (and gradients) are a sum over  $N$  observations.
- 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  $Z$ .

- Think survey sampling – we want to estimate a total.
- 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  $\nabla L(\theta, \mathbf{X}, \mathbf{y})$
- Epochs vs. Iterations



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

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



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

- Can we estimate the gradient better?
- 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



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

- SGD can be slow to converge due to 'jumping' behaviour
- Can improve behaviour using the velocity – the rolling mean of gradients
- Additional hyperparameter  $\alpha$  to control velocity

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

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





UPPSALA  
UNIVERSITET

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

# 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 Algorithms for Machine Learning

# SGD with momentum

---

Example of SGD with momentum [here](#).



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

- Want the optimizer to adapt to the learning rate  $\eta_t$  to individual parameters
- Common approaches are
  - RMSprop
  - Adaptive Moment Estimation (Adam)