UPPSALA UNIVERSITET

Deep Learning

Lecture 1 – Introduction, linear models

Niklas Wahlström
Division of Systems and Control
Department of Information Technology
Uppsala University

niklas.wahlstrom@it.uu.se
www.it.uu.se/katalog/nikwa778

UPPSALA UNIVERSITET

What is the course about?

Machine learning

"Machine learning is about learning, reasoning and acting based on data."

"It is one of today's most rapidly growing technical fields, lying at the intersection of computer science and statistics, and at the core of artificial intelligence and data science."

Ghahramani, Z. **Probabilistic machine learning and artificial intelligence**. *Nature* 521:452-459, 2015.

Jordan, M. I. and Mitchell, T. M. **Machine Learning: Trends, perspectives and prospects**. *Science*, 349(6245):255-260, 2015.

UPPSALA UNIVERSITET

Deep learning

"Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction."

LeCun, Y., Bengio, Y. and Hinton, G. **Deep learning**. *Nature* 521:436-444, 2015.

Example: Image classification

Input: pixels of an **image**
Output: **object identity**
Each hidden layer extracts increasingly abstract features.

Zeiler, M. D. and Fergus, R. **Visualizing and understanding convolutional networks**

Computer Vision - ECCV (2014).

Image removed in handout version due to copyright

UPPSALA UNIVERSITET

ex) Cancer diagnosis

Systems for detecting cell divisions (mitosis) in histology images can be used to improve (or automate) cancer diagnosis.

Image removed in handout version due to copyright
©Springer-Verlag Berlin Heidelberg 2013

- Learn a model with

  input: RBG histology image (pixel values)
  output: number and locations (in the image) of mitosis detections

- Training data: Histology images labeled by experts.

D. C. Cireșan, A. Giusti, L. M. Gambardella and J. Schmidhuber. **Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks**. In *Medical Image Computing and Computer Assisted Intervention*, 411-418, 2013.

Course information

Lecture outline

1. Introduction
2. Feedforward neural networks
3. Optimization
4. Convolutional neural networks 1
5. Convolutional neural networks 2
6. Over-/underfitting, bias-variance, regularizationf
7. Practical methodology
8. Deep time series models 1
9. Deep time series models 2
10. Project proposal presentation
    (for the project extension of the course)

All lectures will be over zoom and recorded (only accessible for enrolled students)

Hand-in assignments

3 hand-in assignments (HAs):

- Covers (mainly) implementation aspects of deep learning.
- Deadlines for all HAs available on course homepage.
- You are encouraged to collaborate...
- ... but you should write and submit your own report and code.
- Each submission should be a proper report
- The reports will be peer-reviewed

Helpdesks are scheduled after each lecture.

Optional project

- After the course you are encouraged to carry out a project
- Preferably 1-4 students in each team
- Awarded 3hp extra
- Conducted during summer, see course homepage for the dates

Course literature

- Ian Goodfellow, Yoshua Bengio and Aaron Courville *Deep Learning*, MIT Press, 2016. www.deeplearningbook.org

  We will not follow the book strictly though. Another great resource is

- Michael A. Nielson *Neural Networks and Deep Learning* Determiniation Press, 2015. www.neuralnetworksanddeeplearning.com

  Some lectures (1-4,7) are partly covered by

- Andreas Lindholm, Niklas Wahlström, Fredrik Lindsten, and Thomas B. Schön *Supervised machine learning* www.smlbook.org

All of them available online.

Previous course evaluation (I/III)

General comments

The teachers took a complicated topic and explained so that it was easy to follow

The lectures were good and informative

I really enjoyed them [assignments]! I specifically liked assignment 1a and 1b, it was a bit like taking a drivers license.

It's interesting to apply the concepts and I learned there more than in class, but they [assignments] are too long and too demanding for the number of credits.

Previous course evaluation (II/III)

What in this course has been particularly good?

No writing exam is the best part

Course scope and content, and the exercises. Thank you all for organizing this great course :)

The understanding of what a nn actually is and how the magic works.

The hand-in assignment were really good. Also, the the course itself was very well structured and organized.

The effort the teachers made so that everyone could follow

Previous course evaluation (III/III)

What specific measures are most important to improve the course?

Less annoying homework.

Furthermore, knowing the course is for PhD students and we might have project deadlines, conferences, defences... I found the deadlines quite stressing

The focus on image analysis was maybe a little too heavy.

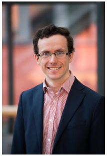
In my opinion, some materials could be discussed faster

Changes for this year

- Three assignment instead of four, and course one week longer $\Rightarrow$ hopefully deadlines less densely packed.
- One new module on time series models added (and previous one on variational autoencoder removed)
- The phase is a tricky one, due to broad spectra of pre-knowledge, but we will do our best to find a good balance.

Who are we?

Teachers involved in the course (in approximate order of appearance):

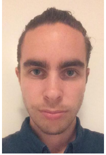


Niklas Wahlström

Thomas Schön

Joakim Lindblad

Carl Andersson



Håkan Wieslander

Ankit Gupta

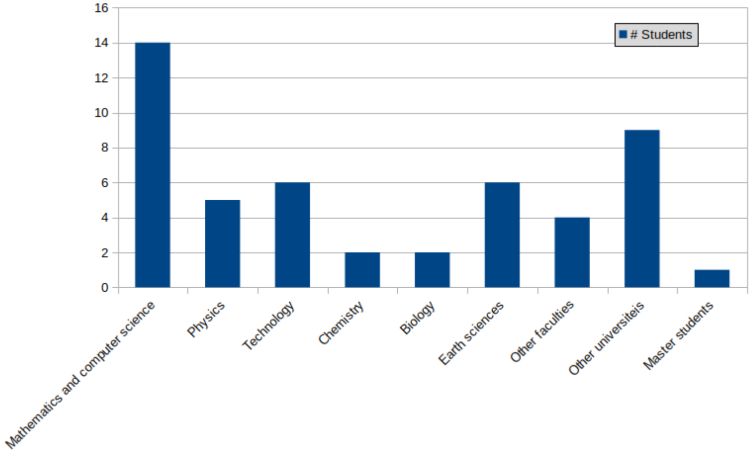
Eduard Chelebian

Lecturers

Teaching assistants

You can reach us by email: `<firstname.lastname>@it.uu.se`.

Who are you?



16
14
12
10
8
6
4
2
0
Students
Mathematics and computer science
Physics
Technology
Chemistry
Biology
Earth sciences
Other faculties
Other universiteis
Master students

Your expectations (I/II)

Learn more about the theory behind deep learning and how to apply the ideas.

To really understand the details in the setup of a neural network.

I would like to dig the basic concepts of DL in more detail

I wish to understand the core concepts of deep learning and gain hands-on experience with implementing them. I think this experience shall be useful for my PhD and also for my subsequent career.

UPPSALA UNIVERSITET

Your expectations (II/II)

To finally get to know deep learning that everybody talks about. Potentially to get so much knowledge so that I can include some of these methods in my research.

Get a theoretical aspect of the deep learning methods and how to implement some of these methods and maybe apply it to my own research project.

Hope to gain a good practical and theoretical understanding of ANNs and to get some insight into the current research of ANNs

How could deep learning could be applied in your research project? (I/II)

- A fairly broad spectra of answers.
- Most you you had some thought of how deep learning could be applied to your research, spanning from some more general ideas...

  ... there is an increasing tendency in using neural networks (of varying depths) in my field ...

  I know there have been some papers using machine learning in my research area....

  I see that many researchers are leaning towards DL in our field where large data is available.

How could deep learning could be applied in your research project? (II/II)

- ... to some quite specific ideas central to your research.

  Yes, I am working with the industry and one part of my research data is images from industrial parts that is taken by several cameras.

  My research project would be developing a deep learning based robot assistant to automatically diagnose depression. It would use multimodal data as input features.

  My research project relies heavily on deep learning. The goal of my PHD is to semantically segment and map individual tree crowns from satellite images at 3m resolution, and run this at continental scale.

  Using CNN as a feature extractor is a central part of my PhD project

UPPSALA UNIVERSITET

Linear regression and classification

Supervised machine learning

Methods for automatically learning (training, estimating, ...) **a model** for the relationship between

- the **input** x, and the
- the **output** y

from observed **training data**

$$\mathcal{T} \stackrel{\mathsf{def}}{=} \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}.$$

Regression vs. classification

We will distinguish between two types of problems: **regression** and **classification**

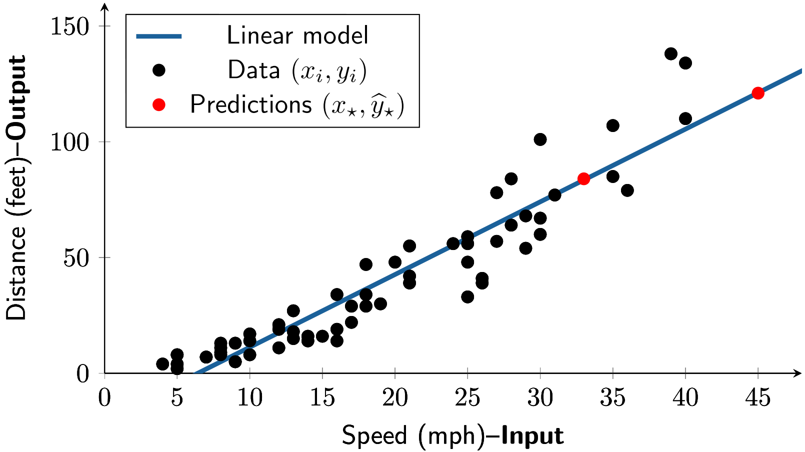
Regression is when the output y is quantitative, e.g.

- Climate models ($y =$ "increase in global temperature")
- Economic models ($y =$ "change in GDP")

Classification is when the output y is qualitative, e.g.

- Spam filters ($y \in \{\texttt{spam}, \texttt{good email}\}$)
- Diagnosis systems ($y \in \{\texttt{ALL}, \texttt{AML}, \texttt{CLL}, \texttt{CML}, \texttt{no leukemia}\}$)
- Fingerprint verification ($y \in \{\texttt{match}, \texttt{no match}\}$)

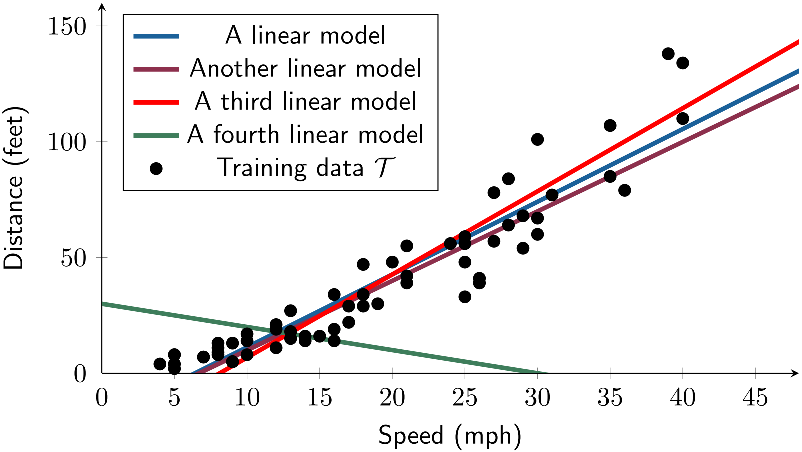
ex) A regression problem



Consider a linear model to explain the data

$$\hat{y} = wx + b$$

What is a good model?



Learning using maximum likelihood

Assume that each data points can be described by a linear model + noise

$$y_i = wx_i + b + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

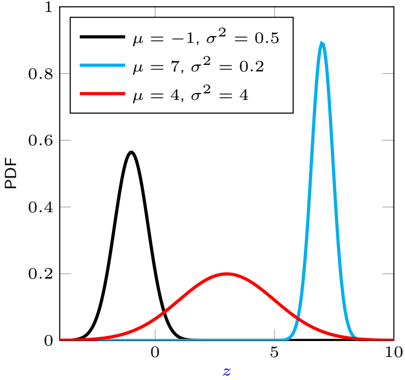
Maximum likelihood: Think of ε (dotted) as Gaussain random variables, and **choose the model** (solid) **such that the resulting ε are as likely as possible**.

Gaussian (Normal) distribution

Probability density function (PDF) for the scalar Gaussian distribution

$$\mathcal{N}(z; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$$

- μ is the mean (expected value of the distribution)
- σ is the standard deviation
- σ^2 is the variance



$z \sim \mathcal{N}(z; \mu, \sigma^2)$ means that z is a Gaussian random variable with mean μ and variance σ^2. $\sim$ reads "distributed according to".

Maximum likelihood

A linear model with Gaussian noise

$$y_i = wx_i + b + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \qquad i = 1, \dots, n$$

The model can also be expressed as

$$p(y_i | x_i, w, b) = \mathcal{N}(y_i; wx_i + b, \sigma^2)$$

Pick w and b which makes the data as likely as possible

$$\widehat{w}, \widehat{b} = \underset{w,b}{\operatorname{argmax}}\, p(y_1, \dots, y_n | x_1, \dots, x_n, w, b)$$

Assume all ε_i to be independent

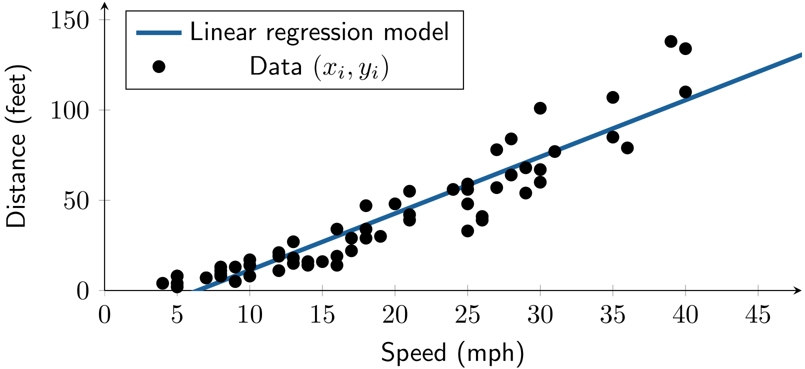
y and z independent
$\Rightarrow p(y, z) = p(y)p(z)$

$$\begin{aligned} p(y_1, \dots, y_n | x_1, \dots, x_n, w, b) &= \prod_{i=1}^{n} p(y_i | x_i, w, b) \\ &= \prod_{i=1}^{n} \mathcal{N}(y_i; wx_i + b, \sigma^2) \\ &\propto e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - wx_i - b)^2} \end{aligned}$$

Linear regression

The least squares problem

$$\widehat{w}, \widehat{b} = \underset{w,b}{\text{argmin}} \sum_{i=1}^{n} (y_i - wx_i - b)^2$$

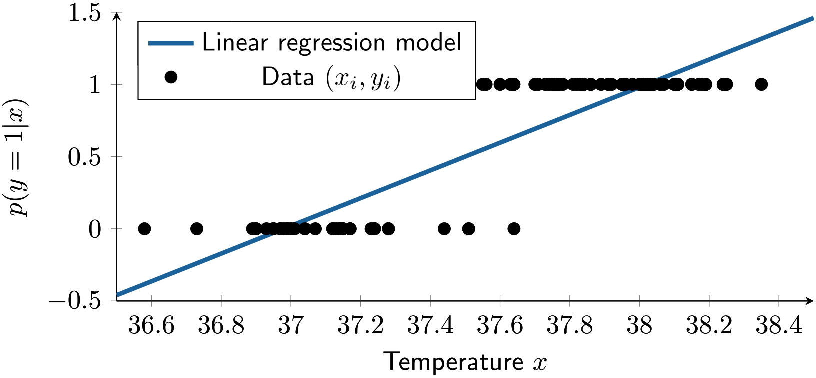


ex) A classification problem

Beaver Body Temperature Data

Input Body temperature x

Output Beaver is outside $y = 1$, or inside $y = 0$, of the retreat.



Linear regression is not suitable since it is not constrained to $[0,\ 1]$

Logistic regression

Consider a binary classification problem $y \in \{0, 1\}$.

$$p_i = p(y_i = 1|x_i) \qquad \text{and thus} \qquad p(y_i = 0|x_i) = 1 - p_i$$

Let the **odds** be the ration between the two class probabilities

$$\frac{p(y_i = 1|x_i)}{p(y_i = 0|x_i)} = \frac{p_i}{1 - p_i} \in (0, \infty)$$

and **log odds** consequently

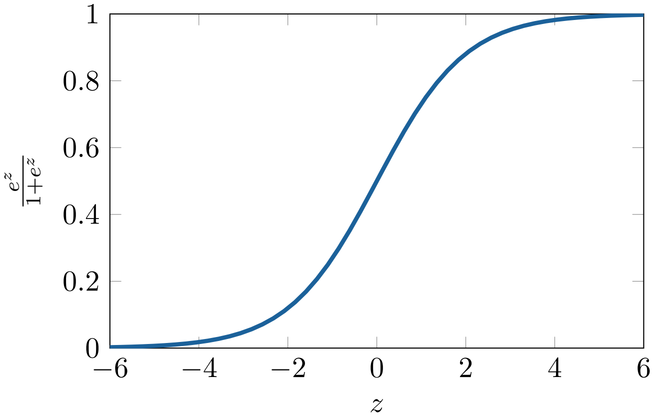
$$\ln \frac{p_i}{1 - p_i} \in (-\infty, \infty)$$

Logistic regression Assume a linear model for the log odds

$$\ln \frac{p_i}{1 - p_i} = wx_i + b \qquad \Rightarrow \qquad p_i = \frac{e^{z_i}}{1 + e^{z_i}}, \quad z_i = wx_i + b$$

Logistic function (aka sigmoid function)

The function $f : \mathbb{R} \mapsto [0,1]$ defined as $f(z) = \frac{e^z}{1+e^z}$ is known as the **logistic function**.



Logistic regression using maximum likelihood

Pick w and b which make data as likely as possible

$$\widehat{w}, \widehat{b} = \underset{w,b}{\operatorname{argmax}} \; \ln p(y_1, \dots, y_n | x_1, \dots, x_n, w, b)$$

$\boxed{\begin{array}{l} \ln(\cdot) \text{ strictly increasing} \Rightarrow \\ \underset{x}{\operatorname{argmax}} f(x) = \underset{x}{\operatorname{argmax}} \ln f(x) \end{array}}$

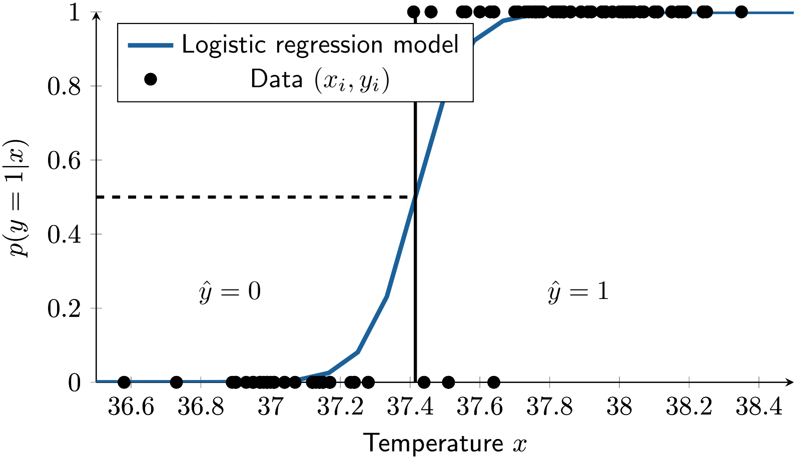
Assume all y_i to be independent

$$\begin{aligned} &\ln p(y_1, \dots, y_n | x_1, \dots, x_n, w, b) \\ &= \sum_{i=1}^{n} \ln p(y_i | x_i, w, b) \\ &= \sum_{\substack{i=1 \\ \text{where} \\ y_i=1}}^{n} \ln \underbrace{p(y_i = 1 | x_i, w, b)}_{=p_i} + \sum_{\substack{i=1: \\ y_i=0}}^{n} \ln \underbrace{p(y_i = 0 | x_i, w, b)}_{=1-p_i} \end{aligned}$$

This leads to the following optimization problem

$$\widehat{w}, \widehat{b} = \underset{w,b}{\operatorname{argmin}} \; -\sum_{i=1}^{n} y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)$$

The Beaver data example



Linear and logistic regression Multidimensional input

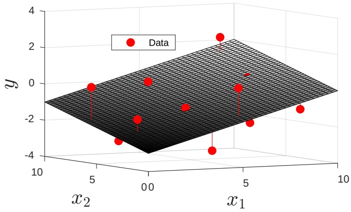
Linear and logistic regression also work for multidimensional inputs

$$\mathbf{x}_i = [x_{i1}, \dots, x_{ip}]^\mathsf{T}, \qquad i = 1, \dots, n$$

We assign one parameter for each input dimension

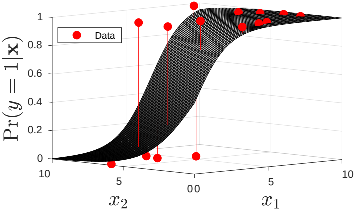
$$\mathbf{w} = [w_1, \dots, w_p]^\mathsf{T}$$

Linear regression



$$\hat{y}_i = \mathbf{w}^\mathsf{T}\mathbf{x}_i + b$$

Logistic regression



$$p(y_i = 1|x_i) = \frac{e^{\mathbf{w}^\mathsf{T}\mathbf{x}_i+b}}{1+e^{\mathbf{w}^\mathsf{T}\mathbf{x}_i+b}}$$

Linear and logistic regression

UPPSALA UNIVERSITET

Linear regression

Output

$y_i \in \mathbb{R}$

Model

$\hat{y}_i = \mathbf{w}^\mathsf{T}\mathbf{x}_i + b$

Loss

$L_i = (y_i - \hat{y}_i)^2$

Logistic regression

Output

$y_i \in \{0, 1\}$

Model

$p_i = p(y_i = 1|\mathbf{x}_i) = \frac{e^{\mathbf{w}^\mathsf{T}\mathbf{x}_i + b}}{1 + e^{\mathbf{w}^\mathsf{T}\mathbf{x}_i + b}}$

Loss

$L_i = -y_i \ln(p_i) - (1 - y_i)\ln(1 - p_i)$

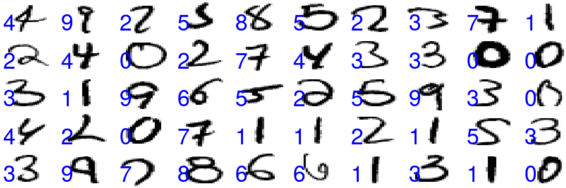
We find $\mathbf{w}$ and b by minimizing the sum of the losses

$$\widehat{\mathbf{w}}, \widehat{b} = \underset{\mathbf{w}, b}{\text{argmin}} \frac{1}{n}\sum_{i=1}^{n} L_i$$

- For linear regression the problem can be solved analytically.
- For logistic regression we need to use numerical optimization (will talk about it in the next lecture).

Hand-in assignment 1 - Classifying hand-written digits with neural networks

Input $\mathbf{x}$: Images of hand-written digits with $p = 784$ pixels
Output y: The digit that the image depicts: 0,1,2,3,4,5,6,7,8 or 9



Extend your code from pre-course assignment in three aspects:

1. Adding **softmax** function on the output to handle a classification problem with $M > 2$ classes.
2. Train with **minibatch** instead of batch gradient descent.
3. Extend the model to include **multiple layers**.

A few concepts to summarize lecture 1

Machine Learning: Deals with learning, reasoning and acting based on data.

Deep Learning: A set of machine learning methods that allow models composed of multiple processing layers.

Regression: Learning problem where the *output* is quantitative.

Classification: Learning problem where the *output* is qualitative.

Maximum likelihood: Learning objective based on probability theory.

Linear Regression: Linear model for regression problems.

Logistic Regression: Linear model for classification problems.