UPPSALA
UNIVERSITET

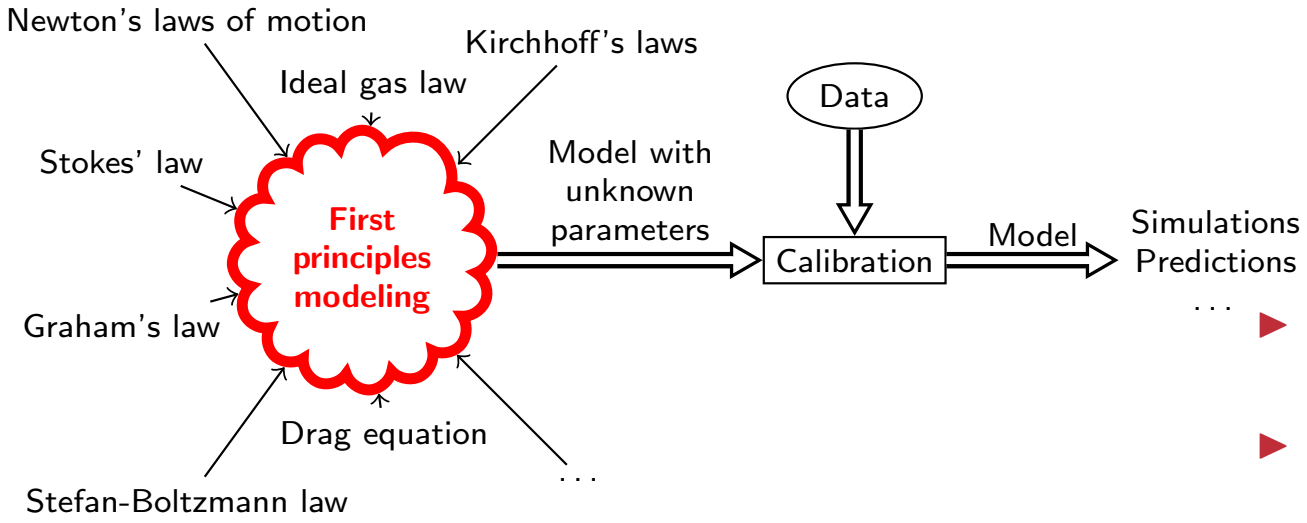
Statistical Machine Learning
an overview

Andreas Svensson

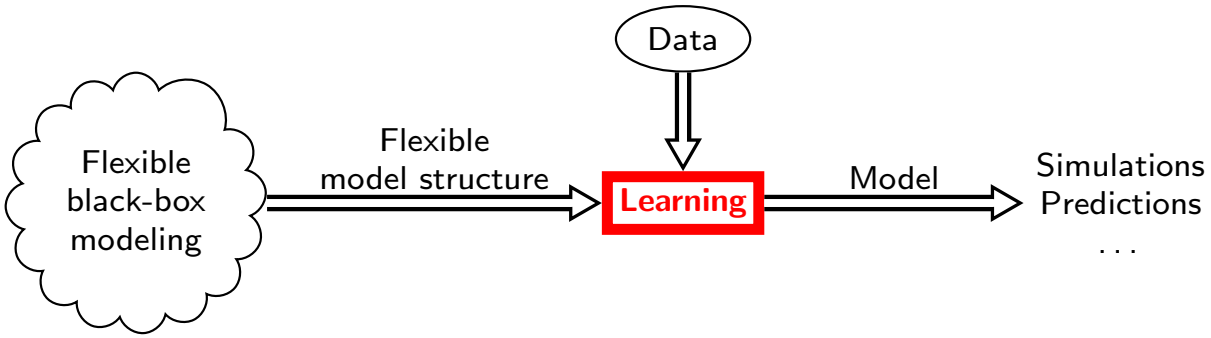
Department of Information Technology
Uppsala University

Machine Learning: a very subjective view

'Classical' approach



Statistical machine learning approach



- Level of detailed knowledge present?
- Model-reduction: manual or automated?
- When to make approximations?
- Interpretation of the model?
- Representing uncertainty

andreas.svensson@it.uu.se

Outline

- Introduction
- Deep learning
- Gaussian processes
- Outlook

andreas.svensson@it.uu.se
Department of Information Technology, Uppsala University

Introduction to neural networks (I/III)

What is a neural network?

A neural network (NN) is a nonlinear function $y = g_{\boldsymbol{\theta}}(\boldsymbol{x})$ from an input variable $\boldsymbol{x}$ to an output variable y, parameterized by $\boldsymbol{\theta}$.

Linear regression models the relationship between the input $\boldsymbol{x}$ and the output y as a linear combination

$$y = \sum_{i=1}^{n} x_i \theta_i + \theta_0 = \boldsymbol{x}^\mathsf{T} \boldsymbol{\theta},$$

where $\boldsymbol{\theta}$ is the parameters composed by the "weights" θ_i and the offset ("bias") term θ_0,

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_0 & \theta_1 & \theta_2 & \cdots & \theta_n \end{pmatrix}^\mathsf{T},$$
$$\boldsymbol{x} = \begin{pmatrix} 1 & x_1 & x_2 & \cdots & x_n \end{pmatrix}^\mathsf{T}.$$

andreas.svensson@it.uu.se Department of Information Technology, Uppsala University

Introduction to neural networks (II/III)

1. Form m_1 linear combinations of the input

$$a_j^{(1)} = \sum_{i=1}^{n} \theta_{ji}^{(1)} x_i + \theta_{j0}^{(1)} = \boldsymbol{x}^\mathsf{T} \boldsymbol{\theta}_j^{(1)}, \quad j = 1, \ldots, m_1.$$

2. Apply a (simple) nonlinear transformation

$$z_j^{(1)} = f\left(a_j^{(1)}\right), \quad j = 1, \ldots, m_1.$$

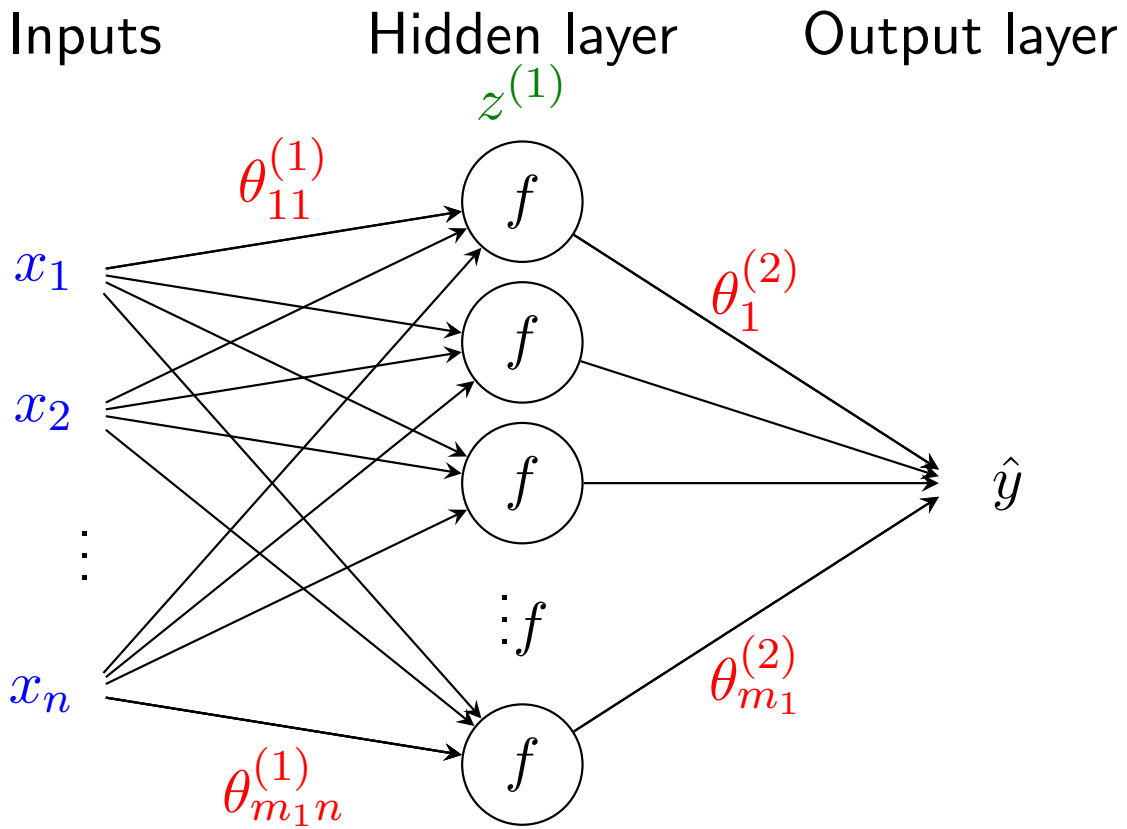
(common choices: $f(a) = 1/(1 + e^{-a})$, $f(a) = \tanh(a)$ or $f(a) = \max(0, a)$)

3. Form a linear combination of the z_js

$$y = \sum_{j=1}^{m_1} \theta_j^{(2)} z_j^{(1)} + \theta_0^{(2)} = \boldsymbol{z}^{(1)\mathsf{T}} \boldsymbol{\theta}_j^{(1)}$$

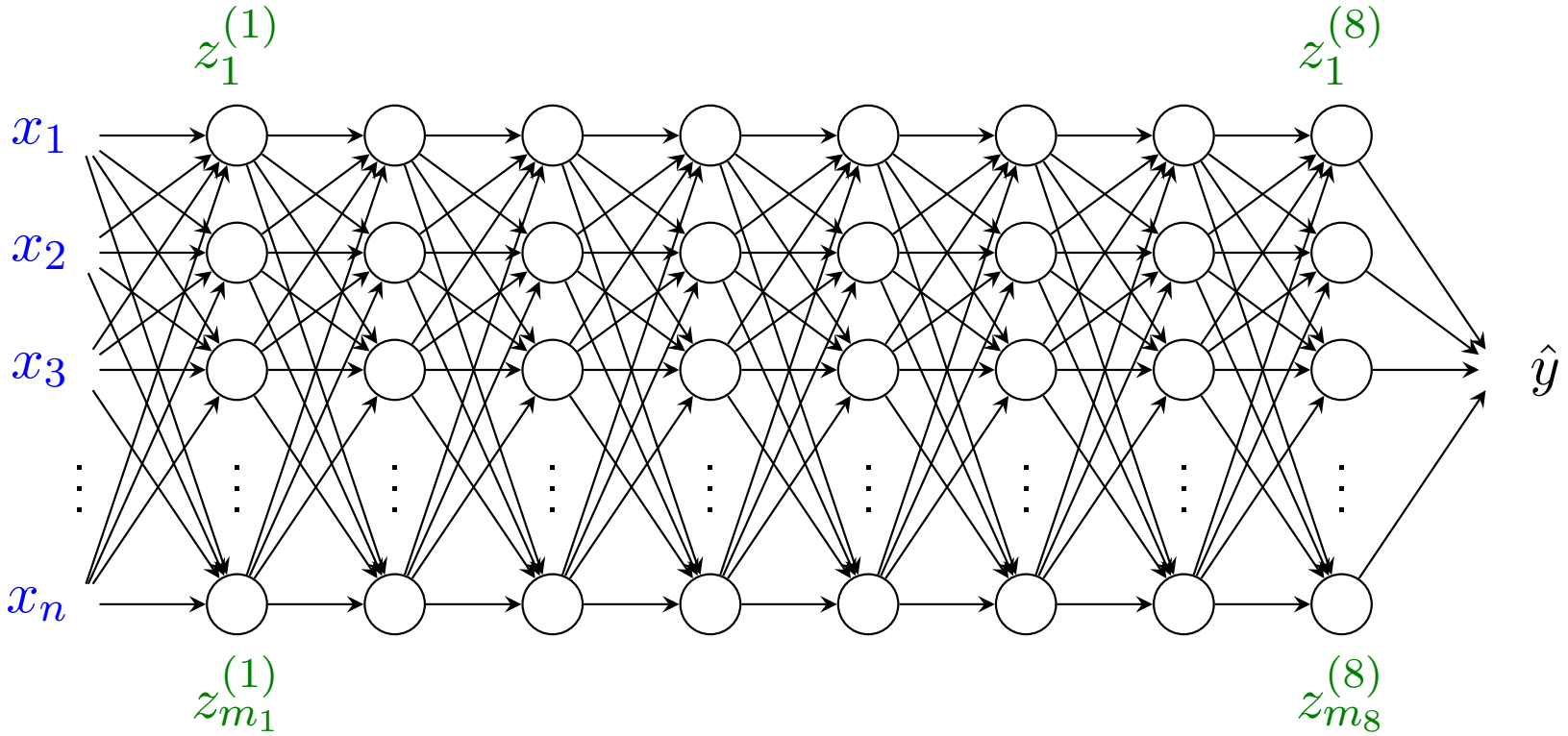
Introduction to neural networks (III/III)

$$\hat{y}(\theta) = \sum_{j=1}^{m_1} \theta_j^{(2)} \underbrace{f\left(\sum_{i=1}^{n} \theta_{ji}^{(1)} x_i + \theta_{j0}^{(1)}\right)}_{z_j^{(1)}} + \theta_0^{(2)}$$



andreas.svensson@it.uu.se Department of Information Technology, Uppsala University

Deep neural networks

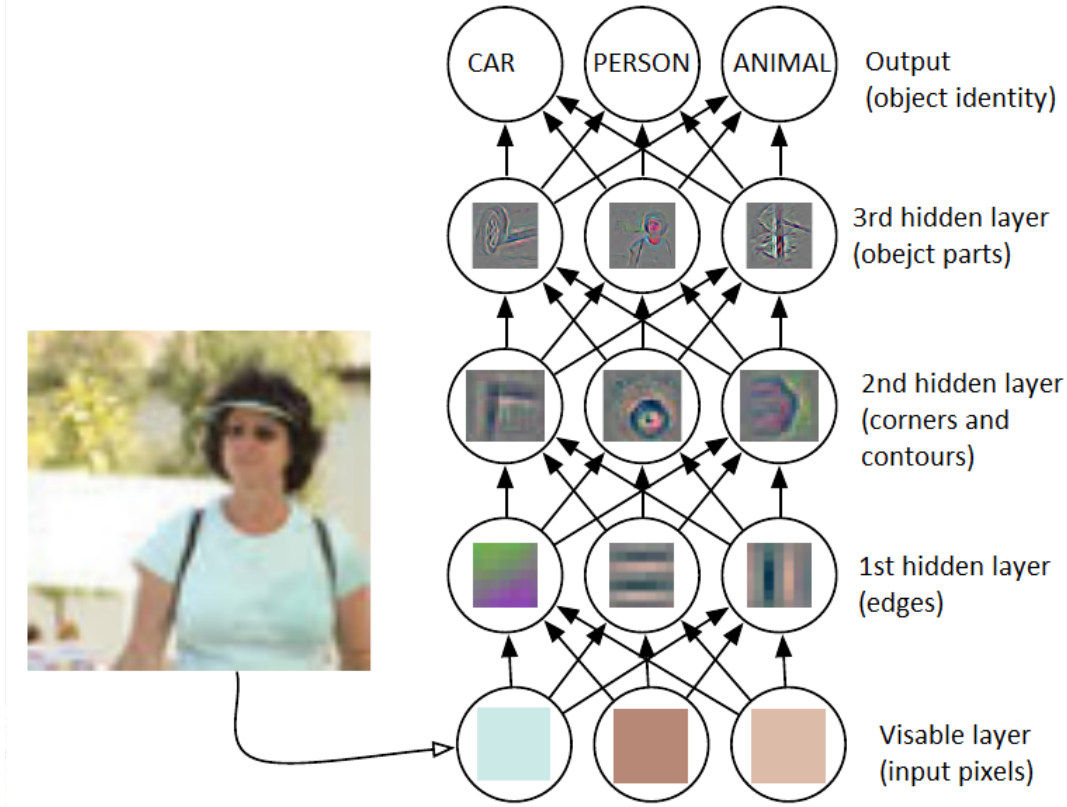


How to learn the unknown parameters $\theta_{ij}^{(d)}$?

Differentiation of $(\hat{y} - y)^2$ with respect to $\theta_{ij}^{(d)}$, *backpropagation*.

Deep learning: Image classification

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



M. D. Zeiler and R. Fergus (2014) **Visualizing and understanding convolutional networks** *Computer Vision - ECCV*.

Deep learning: A recent example

An AI defeated a human professional for the first time in the game of Go

D. Silver et al. (2016) **Mastering the game of Go with deep neural networks and tree search**, *Nature*, vol 529.

Deep learning—Why now?

Neural networks have been around for more than fifty years. Why have they become so popular now (again)?

To solve really interesting problems you need:

1. Efficient learning algorithms
2. Efficient computational hardware
3. A lot of labeled data!

This has not been fulfilled to a satisfactory level until the last 5-10 years.

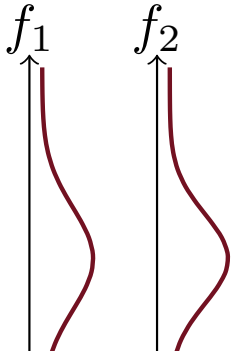
Outline

- Introduction
- Deep learning
- Gaussian processes
- Outlook

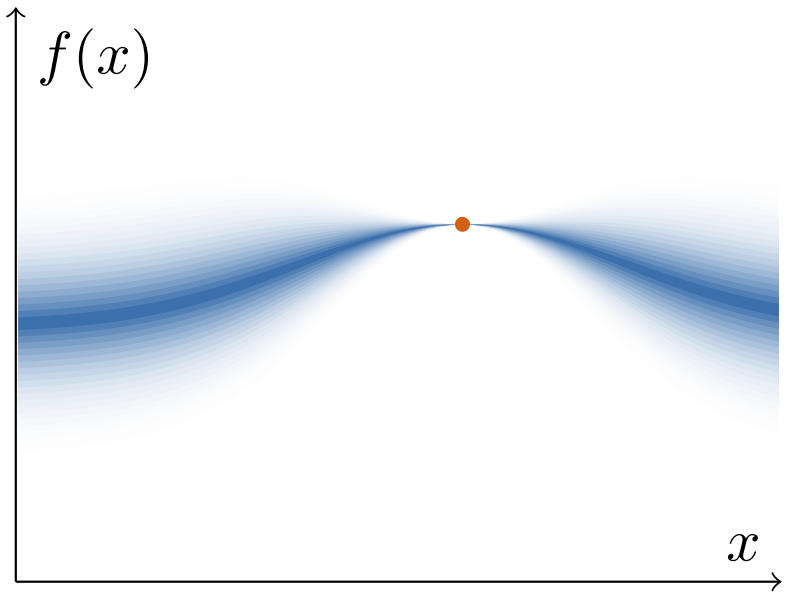
GP: a probability distribution over functions

Gaussian distribution for f_1,

$$p(f_1) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(f_1 - \mu)^2)$$

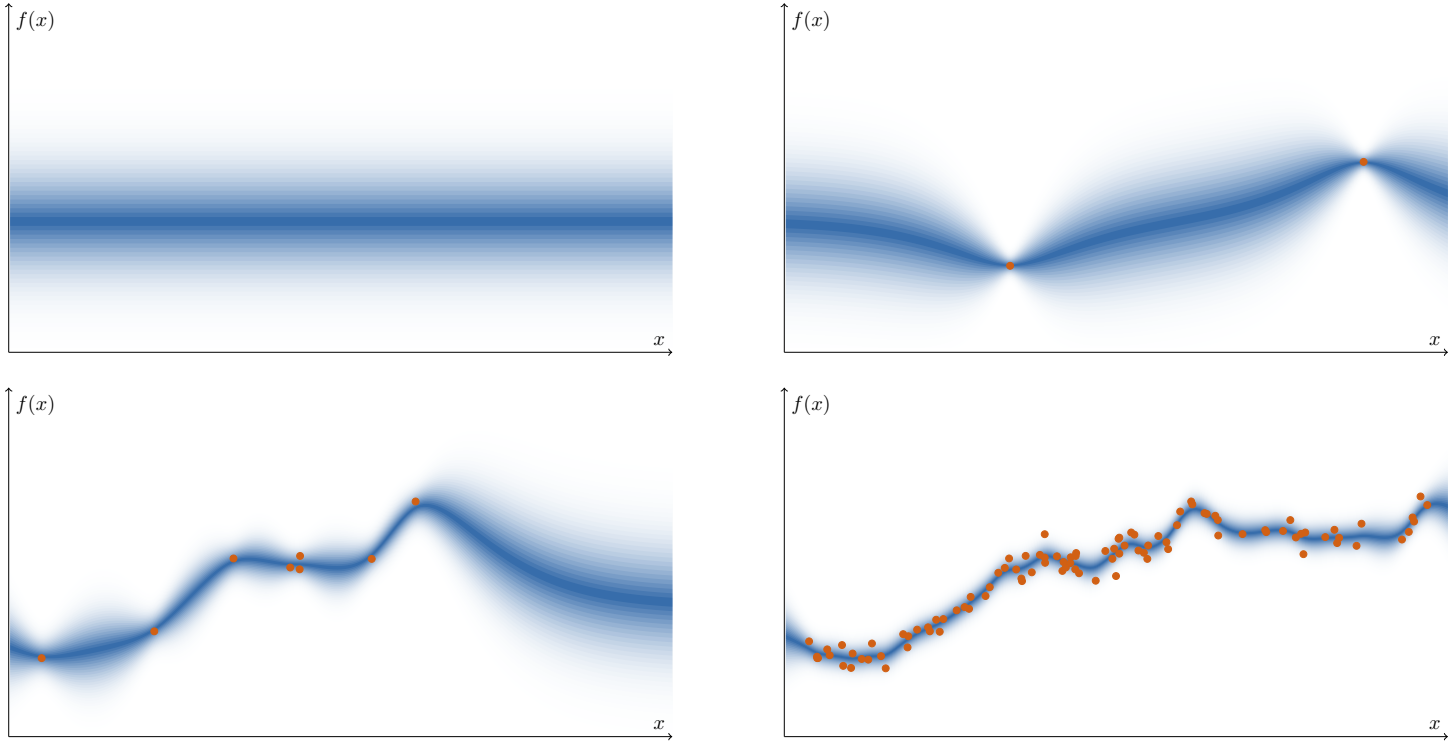


GP: a probability distribution over functions



Distribution for $f(x)$ *conditioned* on one observation (orange dot)

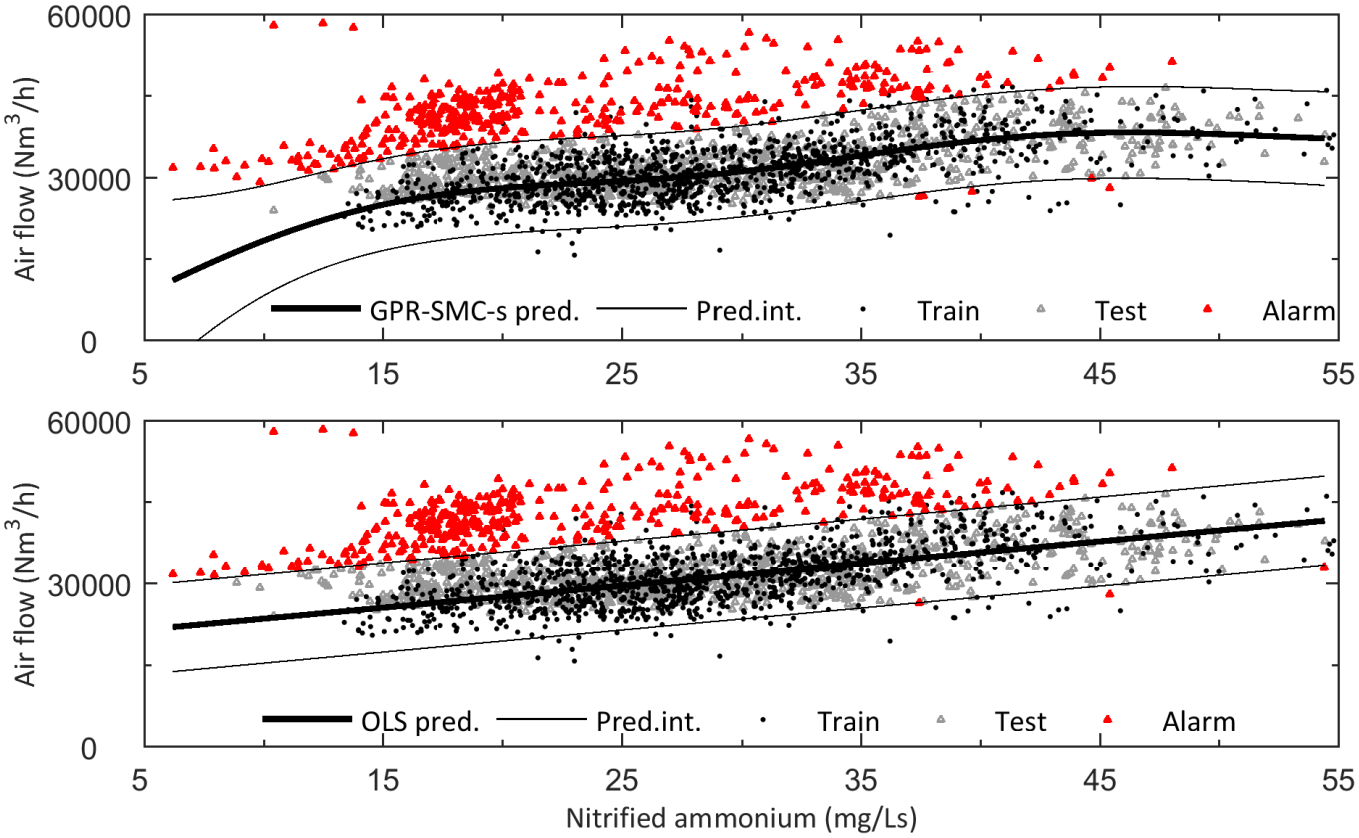
Gaussian Processes: a flexible model



There are 'tuning knobs' for the smoothness etc.

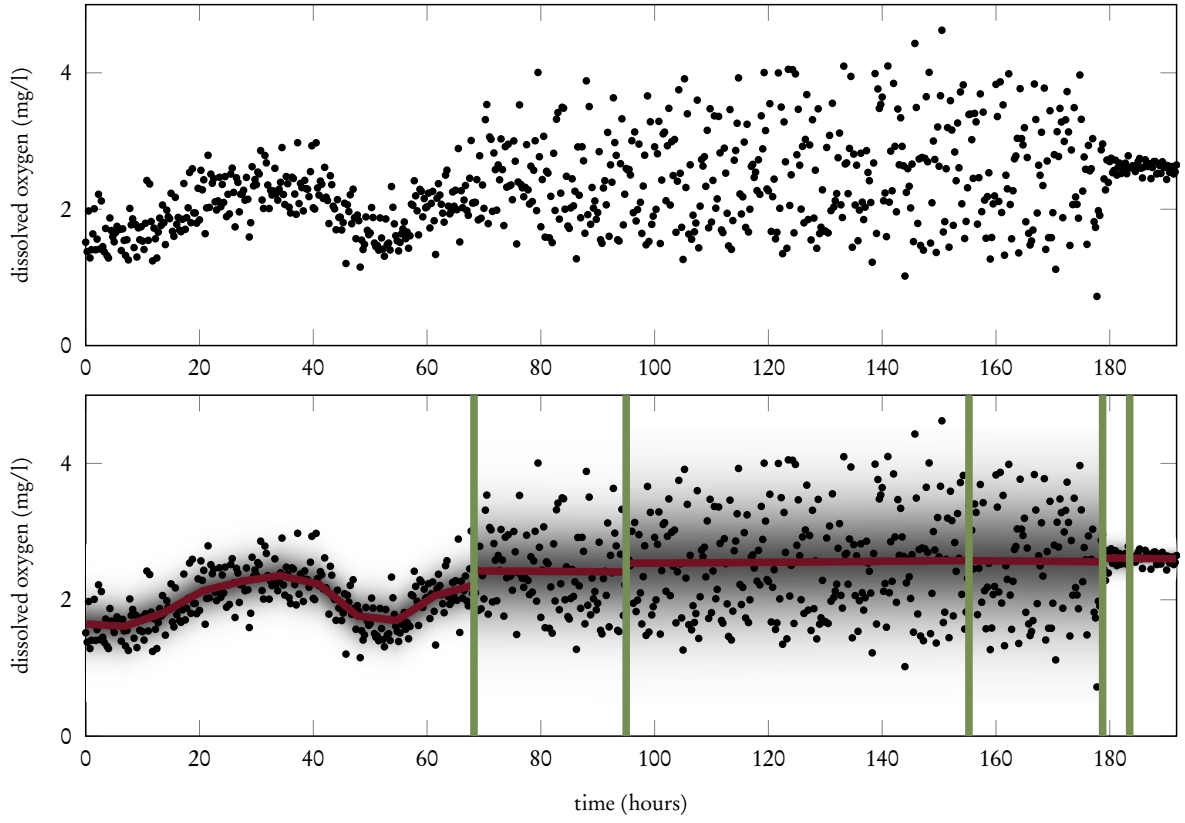
andreas.svensson@it.uu.se
Department of Information Technology, Uppsala University

Gaussian Processes: Fault detection I



O. Samuelsson et al. (2016), **Improved monitoring and fault detection of wastewater treatment processes with Gaussian process regression**, *manuscript*.

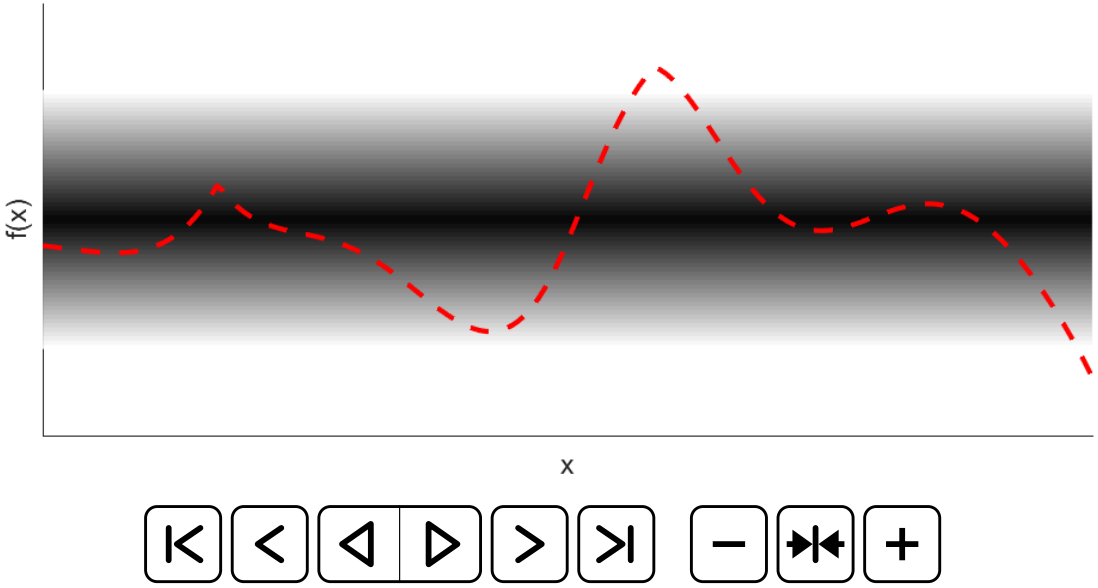
Gaussian Processes: Fault detection II



A. Svensson, J. Dahlin and T.B. Schön (2015) **Marginalizing Gaussian Process Hyperparameters using Sequential Monte Carlo**. *IEEE CAMSAP*.

GP optimization

- Optimization of some parameters
- Expensive to evaluate the objective function (e.g., a lengthy computer simulation)



M. Osborne, R. Garnett and S. Roberts (2009) **Gaussian Processes for Global Optimization**, *LION3*

B. Shahriari et al. (2016) **Taking the Human Out of the Loop: A Review of Bayesian Optimization**, *Proceedings of the IEEE* vol 104

Outline

- Introduction
- Deep learning
- Gaussian processes
- Outlook

andreas.svensson@it.uu.se
Department of Information Technology, Uppsala University

Of course there is more...

- Regression
- Classification
  - Support vector machines
  - Decision trees
  - Boosting
- Clustering
- Reinforcement learning
- Probabilistic programming
- Computational methods
  - Markov chain Monte Carlo
  - Sequential Monte Carlo
  - Variational inference

Learn more

General

Z. Ghahramani (2015) **Probabilistic machine learning and artificial intelligence**, *Nature*, Vol 521

G. James, D. Witten, T. Hastie and R. Tibshirani (2013) **An introduction to statistical learning with applications in R**, *Springer*

C. Bishop (2006) **Pattern recognition and machine learning**, *Springer*

Deep learning

I. Goodfellow, Y. Bengio and A. Courville (2016) **Deep learning**, *Book in preparation for MIT Press*, `http://www.deeplearningbook.org/`

Y. LeCun, Y. Bengio and G. Hinton (2015) **Deep learning**, *Nature*, vol 521

`http://deeplearning.net/`

Gaussian processes

C. Rasmussen and K. Williams (2006) **Gaussian processes for machine learning**, *MIT Press*

`http://gaussianprocess.org/`

Conferences: Neural Information Processing Systems (**NIPS**, `http://nips.cc/`), the International Conference on Machine Learning (**ICML**, `http://icml.cc/`)

Journals: Journal of Machine Learning Research (JMLR), IEEE Transactions Pattern Analysis and Machine Intelligence (PAMI)

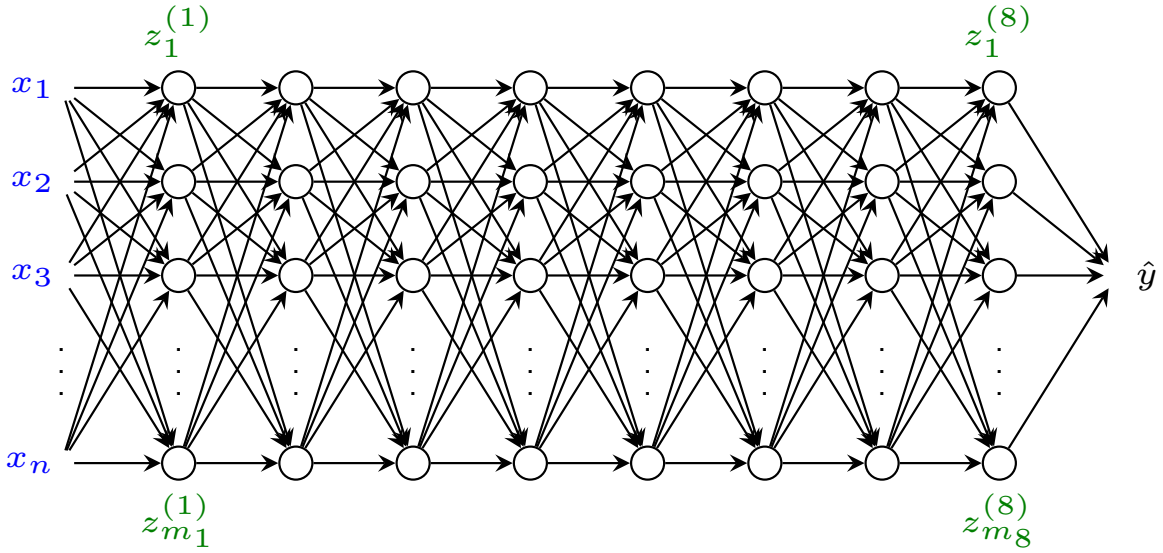
New course

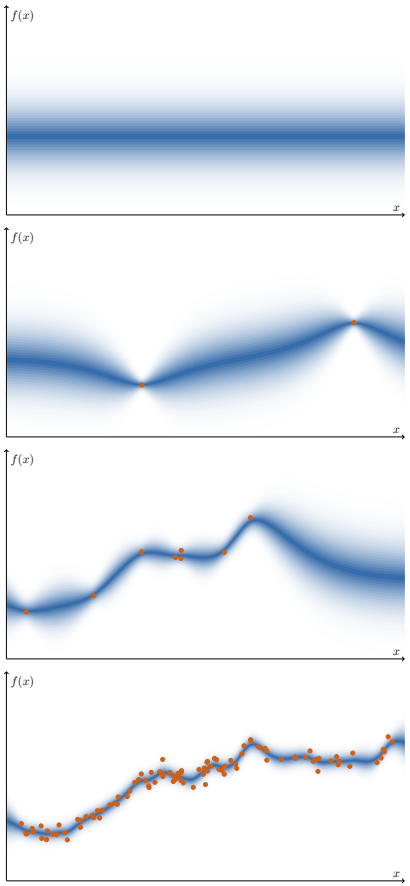
Statistical Machine Learning 5 hp, January-March 2017

- Classical and Bayesian linear regression
- Classification via logistic regression
- Linear discriminant analysis
- Gaussian processes and kernel methods
- Cross-validation and model selection techniques
- Regularization (ridge regression and the LASSO)
- Regression and classification trees
- Principal component analysis
- k-means clustering
- Neural networks

`http://www.it.uu.se/edu/course/homepage/sml`

Thank you!





Questions?

andreas.svensson@it.uu.se
Department of Information Technology, Uppsala University