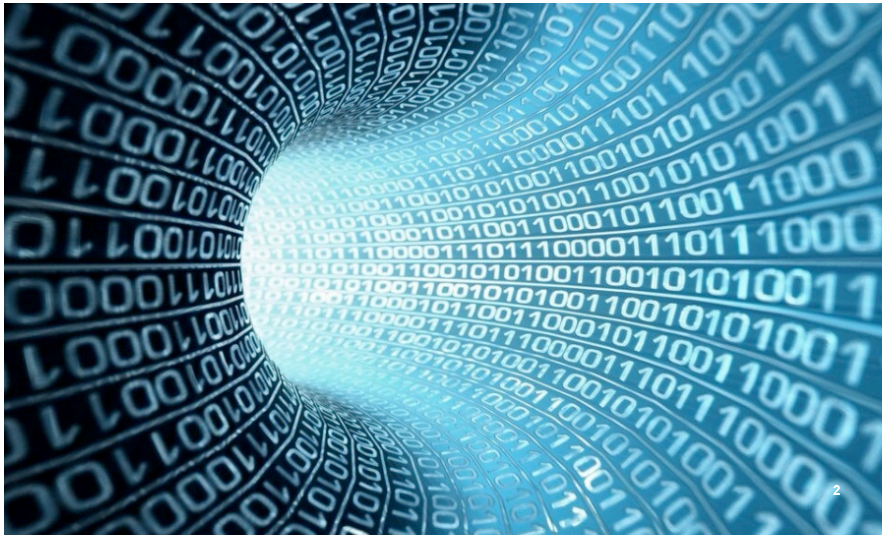
DD2434 Machine Learning, Advanced Course

Lecture 1: Introduction

Hedvig Kjellström
hedvig@kth.se
https://www.kth.se/social/course/DD2434/

Big Data

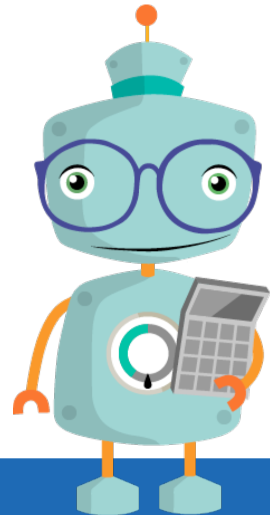


Making sense of signals (RGB-D video): Hand Tracking from MSR Cambridge

https://www.youtube.com/watch?v=A-xXrMpOHyc

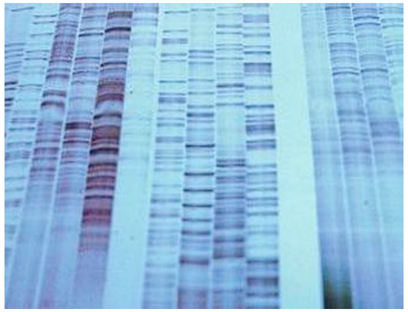
Predicting future events knowing the history: Botten Ada from Linköping U

http://bottenada.se



Learning to see suddle patterns in huge amounts of data: Cancer Therapy based on DNA Sequencing from IBM

https://www.youtube.com/watch?v=0M1DMdc1mQ0



Today

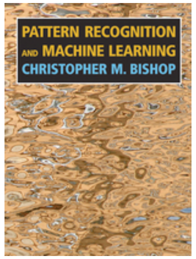
Check the homepage at least 2 times / week!
Or set it to send you emails!

Course preliminaries

All info at https://www.kth.se/social/course/DD2434/

Ask questions through the News forum!

Buy the book by Chris Bishop:



The three teachers

Carl Henrik Ek
Jens Lagergren
Hedvig Kjellström

Introduction to Machine Learning (Bishop 1)

Course Preliminaries

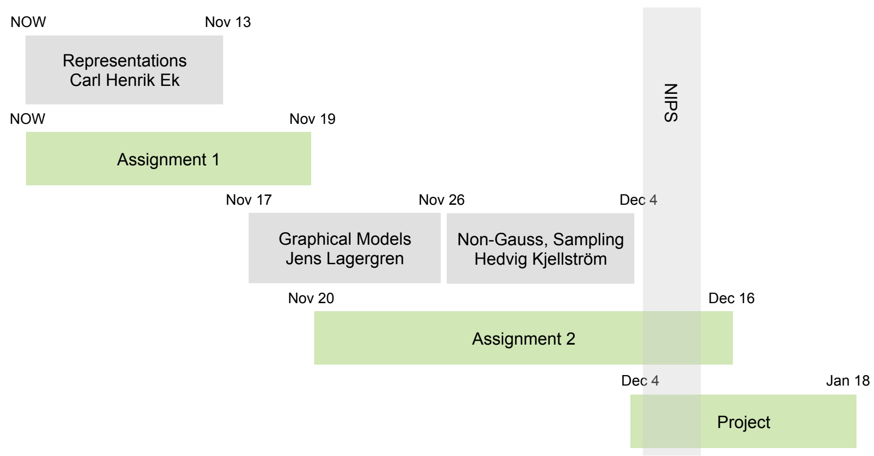
Learning outcomes

Upon completion of the course, the student should be able to

1. explain, derive, and implement a number of models for supervised, unsupervised learning,
2. explain how various models and algorithms relate to one another,
3. describe the strengths and weaknesses of various models and algorithms,
4. select an appropriate model or approach for a new machine learning task.

Course organization

Assignments, detailed schedule with reading, etc, on the homepage



The Three Teachers

Carl Henrik Ek

Assistant Professor of Computer Science at CSC / CVAP

Research area: Robotics and Machine Learning

Responsible for
- Lectures 2-5
- Practicals 1-3
- Assignment 1

Jens Lagergren

Professor of Computer Science at KTH / Science for Life Laboratory

Research area: Bio-informatics

Responsible for
- Lectures 6-9
- Practicals 4-5
- Assignment 2, first half

Hedvig Kjellström

Associate Professor of Computer Science at CSC / CVAP

Research area: Robotics and Computer Vision

Responsible for

- Entire course
- Lectures 1, 10-13
- Practical 6
- Assignment 2, second half

Hedvig Kjellström: My research

Machine learning applied to Robotics and Computer Vision:
Automatic perception of human activity in video

Object affordances, object-action complexes
“automatic understanding of how objects are used in human activities what happens to them during the activity”

Human non-verbal communication
“automatic understanding and modeling of non-verbal signals – face expressions, body motion - both conscious and unconscious”

Multi-modality and context in activity recognition
“using several modalities – vision, sound, touch etc – to better understand human activity”

More about my research:
http://www.csc.kth.se/~hedvig/research.html
Master project proposals:
http://www.csc.kth.se/~hedvig/projects.html

Introduction to Machine Learning

Bishop Chapter 1

Reference manual to Probability Theory Concepts: Bishop Chapter 2

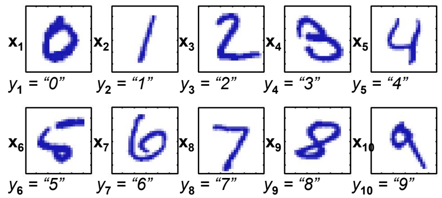
Example: Handwritten digit recognition

How would you do it?

1. Expert system where rules about connectivity, topological structure of digits etc are manually defined - for noisy data such systems are generally outperformed by:
2. Classifier (SVM, ANN, Boosting etc) trained with

Binary images $\mathbf{x}_i$

…with corresponding labels y_i



Need of course many examples from each class {“0”, …, “9”}

Supervised/Predictive Learning

Data (**training set**): $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$

features/attributes

response variable

Task: Learn mapping $\mathbf{x} \rightarrow y$

y can be discrete – classification
or continuous – regression

If observable features $\mathbf{x}$ are noisy and incomplete, the mapping incorporates decision making under uncertainty

Probability theory nice principled tool

Probabilistic formulation: Model function $y = f(\mathbf{x})$ as $p(y = 1|\mathbf{x}, \mathcal{D})$, $p(y = 2|\mathbf{x}, \mathcal{D})$, etc...

Best $y \equiv$ most probable y:

$$\hat{y} = \hat{f}(\mathbf{x}) = \arg\max_{c=1}^{C} p(y = c|\mathbf{x}, \mathcal{D})$$

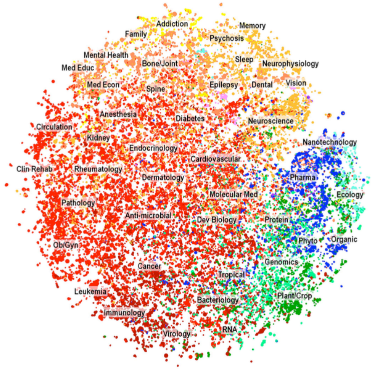
Example: Finding representative groups of biomedical articles

How would you do it?

1. Expert system where a biomedical expert defined the groups, or if the data is big and constantly changing:
2. Clustering algorithm (k-means etc) trained with

The more dimensions in the space spanned by **x**, the more training data needed.

Text documents in bag-of-word representation $\mathbf{x}_i$



Unsupervised/Descriptive Learning

Data (**training set**): $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$

Task: discover patterns in $\mathcal{D}$

Under-specified problem – what patterns? How measure error?

Probabilistic formulation: Density estimation
Models of the form $p(\mathbf{x}_i|\theta)$

Use $\mathcal{D}$ to maximize the probability $p(\mathbf{x}_i|\theta)$ of seeing each $\mathbf{x}_i$ given the model θ

Unsupervised learning is more similar to how humans and animals learn!
Practical advantage: No labeling of data required!

Reinforcement Learning

Will not be discussed in this course

Data, Models, Algorithms

Machine Learning problems are characterized by three different aspects

Data: analysis and preprocessing of data, e.g. sensor observations
Not discussed here – Computer Vision and Speech courses

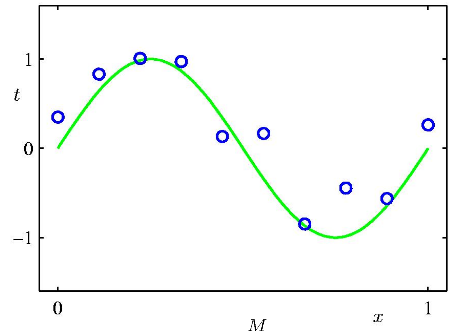
Model: Select a suitable model of the system producing the data
E.g. graphical models, see Lectures 6-8, 11

Algorithm: Algorithm for fitting the model to the data, i.e. adjusting the parameters of the model to best explain the data

Example: Polynomial Curve Fitting

Data:
Measurements $<x_i,t_i>$

Model:
Measurements are sampled from a process $y(x,\mathbf{w})$ which is a polynomial:

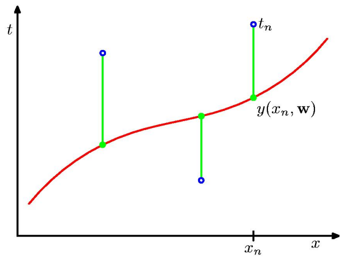


$$y(x,\mathbf{w}) = w_0 + w_1x + w_2x^2 + \ldots + w_Mx^M = \sum_{j=0}^{M} w_j x^j$$

Example: Polynomial Curve Fitting

Algorithm: Find parameters **w** by minimizing the sum of squared errors:

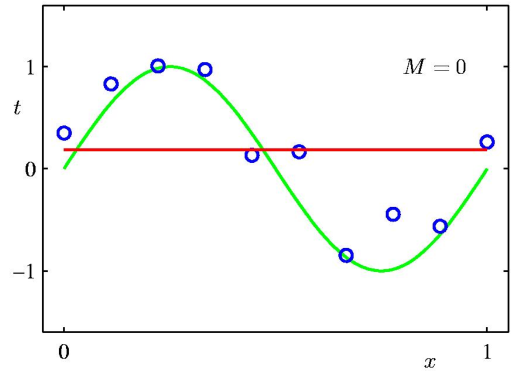
$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2$$



OK, we have now found the optimal parameters **w**, but how should we find the optimal polynomial order *M*?

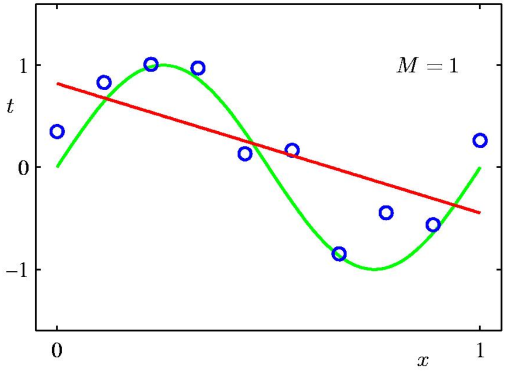
0th order polynomial

Underfitting the data: Not flexible enough to explain data



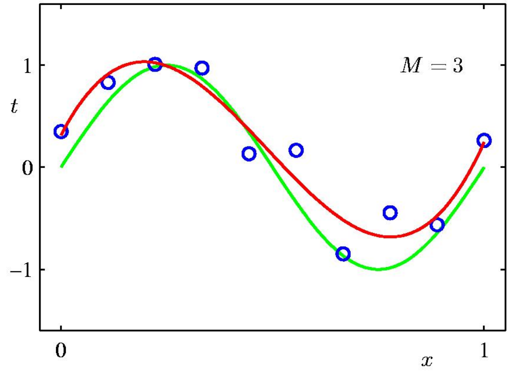
1st order polynomial

Underfitting the data: Not flexible enough to explain data



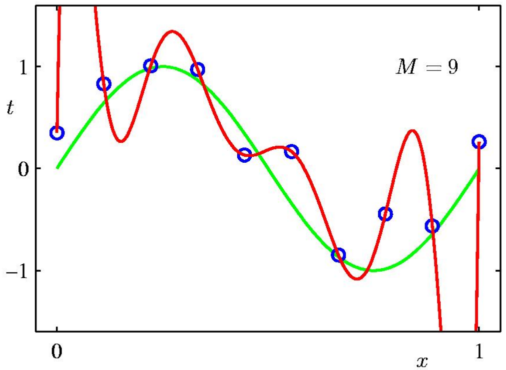
3rd order polynomial

About the right flexibility to imitate the underlying model despite spurious variations in the data

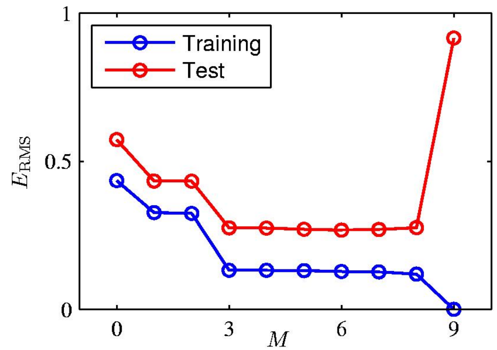


9th order polynomial

Overfitting the data: Model fits training data perfectly but not novel data – it is much more flexible than the true process



Here we saw that M=3 was better than 1 or 9 for this data.
But how find the right model flexibility in general?
Discuss with your neighbor for 5 mins



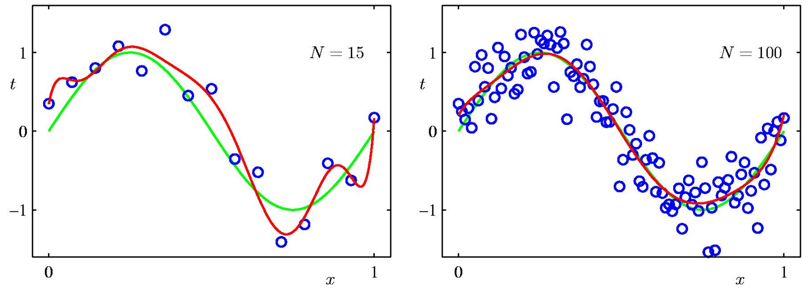
Use separate datasets for testing and training to detect overfitting

Root-Mean-Square (RMS) Error: $E_{\text{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$

Need more data!

10 data points to fit 10 parameters is not enough

9th order polynomial works well with more data



But we should use the existing data as efficiently as possible!

…so let us go back to *N* = 10
What happens to the parameters *w*?

	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

Quite unrealistic with such high coefficients in a polynomial

How could knowledge of this be included in the model? Discuss with your neighbor for 5 mins

Regularization

Basic idea:

Goodness measure should not only depend on data

Give reward to small parameter settings – known to be plausible according to Occam's razor

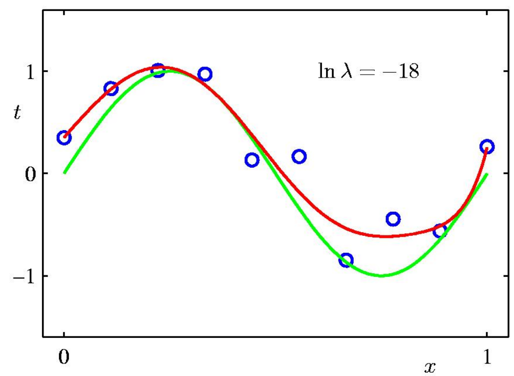
$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

Ridge regression

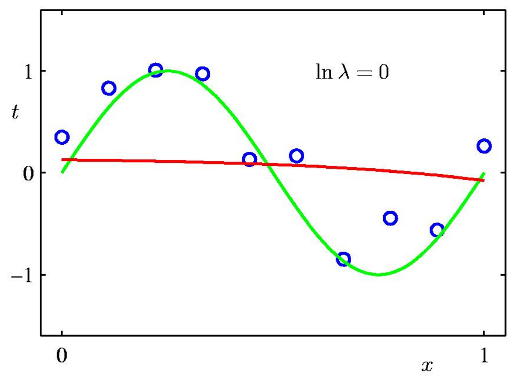
Later in the course:

Bayesian formulation – put priors on model parameters

Regularization: $\ln\lambda = -18$

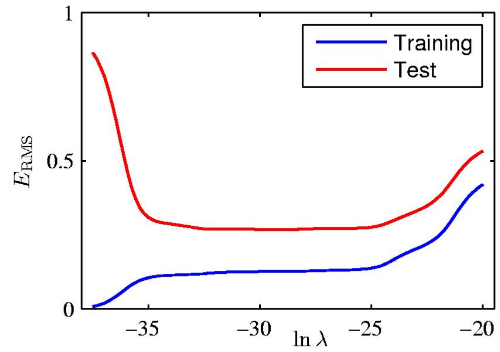


Regularization: $\ln \lambda = 0$



Back to an underfitting model!

Regularization: E_{RMS} vs $\ln \lambda$



This is also confirmed when inspecting *w*

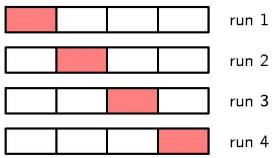
	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

Basic concept:
Model selection

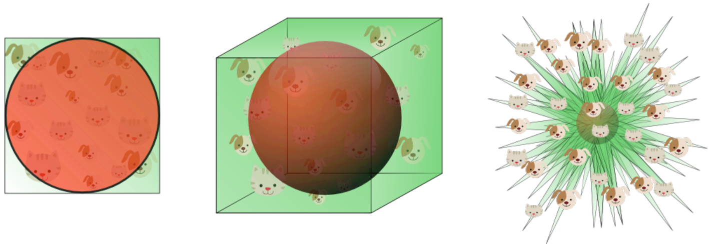
Saw that there is an optimal choice of *M*, **w**, λ in our curve-fitting example

How find the best parameters in a principled manner?
Discuss with your neighbor for 5 mins

S-fold cross-validation, S=4:



Basic concept:
Curse of Dimensionality



2D cube/sphere $= \frac{\pi}{2^2}$

3D cube/sphere $= \frac{4\pi}{2^3 * 3}$

8D cube/sphere $= \frac{\pi^4}{2^8 * 24}$

Adressed by using **lower-dimensional models** and/or **more data**

Basic concept:
No Free Lunch Theorem

Do not believe the preachers…

There is no universally best model! All models contain assumptions that work well in one domain but not in another.

What is next?

Check the homepage at least 2 times / week!
Or set it to send you emails!

We use the homepage a lot: links to video lectures, readings for lectures, lecture slides, questions answered through the News forum
https://www.kth.se/social/course/DD2434/

Next on the schedule
Wed 4 Nov 10:15-12:00 M1
Lecture 2: Regression
Carl Henrik Ek
Readings: Bishop 6.4

Assignment 1 published today, deadline November 19