Introduction to Machine Learning

TTT4185 Machine Learning for Signal Processing

Giampiero Salvi

Department of Electronic Systems
NTNU

HT2021

Examples of applications

Google self driving

IBM congestion fees

autonomous ships

Voice assitants

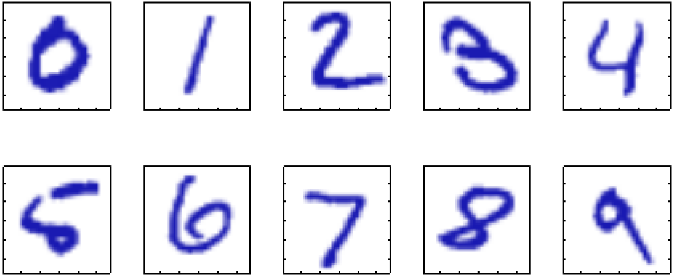
DeepMind AlphaGo

smart buildings

Machine Learning Objectives

1. automatic discovery of regularities in data through computer algorithms
    - similar to statistics
2. use knowledge acquired to take actions

Example: Written digit recognition



MNIST (figure from Bishop)

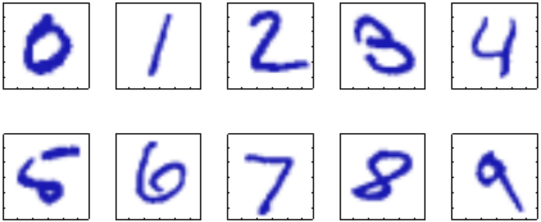
Example: Written digit recognition

Data:

- 28×28 grayscale pixels $[0, 255]$
- pre-processing: centering and normalization
- fixed length representation (784 dim)

Task

- from pixels classify one of 10 discrete digits



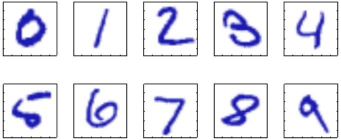
Formalization (Supervised Classification)

Training data:

- set of observations $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}, \mathbf{x}_i \in \mathbb{R}^D$
- set of target values $\{t_1, \dots, t_N\}, t_i \in \{c_1, \dots, c_K\}$

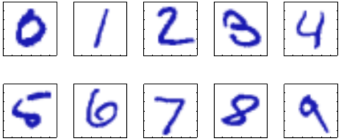
Goal

- find a function $y : \mathbb{R}^D \rightarrow \{c_1, \dots, c_K\}$ such that
- $y(\mathbf{x})$ gives correct answer for any unseen observation $\mathbf{x}$



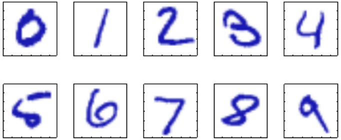
Key aspects

- Training data is incomplete
- Evaluation must be performed on unseen observations (test set)
- We need to ensure generalization
- data generation $\rightarrow$ measurements $\rightarrow$ feature extraction



Feature Extraction

1. disregard irrelevant information
2. reduce the dimensionality (complexity)



Classification vs Regression

Input $\mathbf{x}_i$ can be:

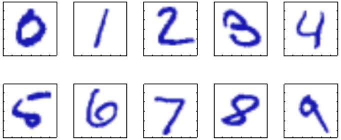
- discrete,
- continuous ($\mathbb{R}$),
- D dimensional ($\mathbb{R}^D$)

Classification:

- discrete targets: $t_i \in \{c_1, \ldots, c_K\}$

Regression:

- continuous targets $t_i \in \mathbb{R}$
- can also be multi-dimensional



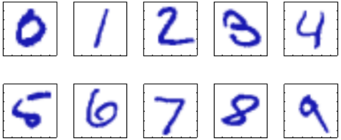
Supervised vs Unsupervised Learning

Unsupervised Learning

- we don't know the value of t_i
- data collection is cheap, but annotations are expensive
- find regularities in data

Applications

- Clustering: group data points according to distance metric
- Density estimation: find parametric model of complex distributions



Reinforcement Learning

- agent
- environment
- actions
- states
- reward

Differences from Supervised Learning

- reward not as detailed as targets
- reward can be delayed
- need to find responsibility of each actions to the reward

Other forms of Learning (Judea Pearl)

	Levels	Activities
1)	Associations	Seeing, hearing

Other forms of Learning (Judea Pearl)

	Levels	Activities
1)	Associations	Seeing, hearing
2)	Intervention	Doing
3)	Counterfactual	Imagining
		Retrospecting

In this course

- Supervised
  - Classification
  - Regression
- Unsupervised
  - Clustering
  - Density estimation
- Combined Supervised/Unsupervised
  - Example: Hidden Markov Models

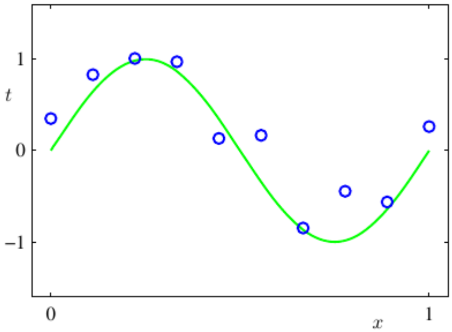
Example: polynomial fitting

Data generation:

- $t = \sin(2\pi x) + \text{noise}$
- underlying regularity $(\sin)$
- uncertainty (noise)

Model: polynomial

$$
\begin{aligned}
y(x, \omega) &= w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M \\
&= \sum_{j=0}^{M} w_j x^j
\end{aligned}
$$



- non-linear in x
- linear in w

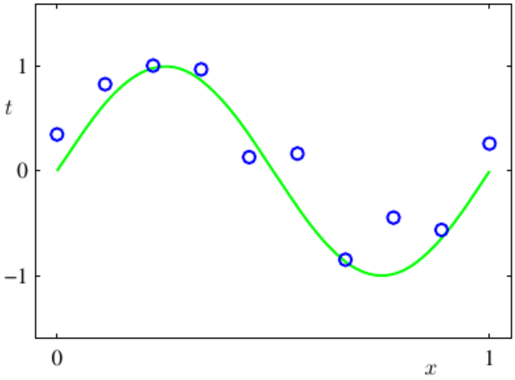
Example: polynomial fitting

Principled methods

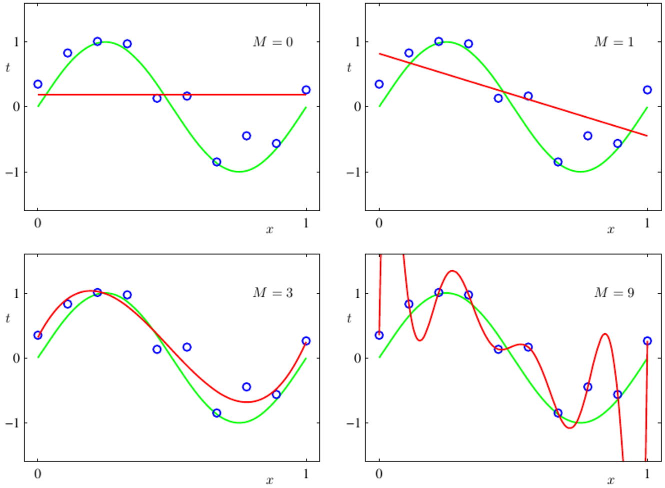
- backed up by a general theory
- in ML: probability theory

Heuristic methods

- based on common sense



Order of the polynomial (from Bishop)



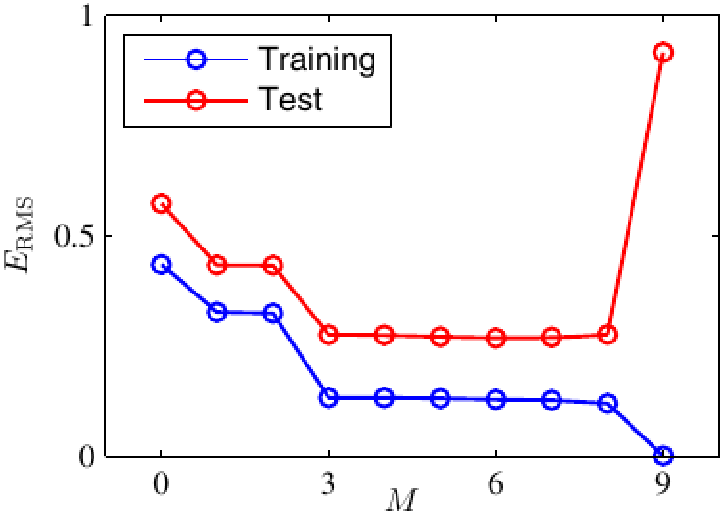
Model parameters (from Bishop)

	$M=0$	$M=1$	$M=6$	$M=9$
$w_0^\star$	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

Overfitting: Training and Test set (from Bishop)

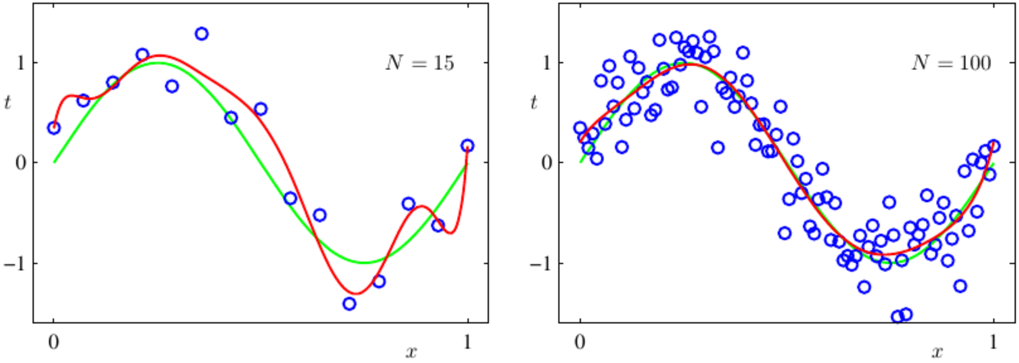
Root Mean Square Error

$$E_{\text{RMS}} = \sqrt{\frac{2E(w)}{N}}$$



Increasing training set size

$\#$ parameters $= 10$



Increasing training set size

Problems:

- annotating data is expensive
- # parameters not equal to complexity
- we would like complexity of model to correspond to complexity of underlying phenomenon

Model Selection

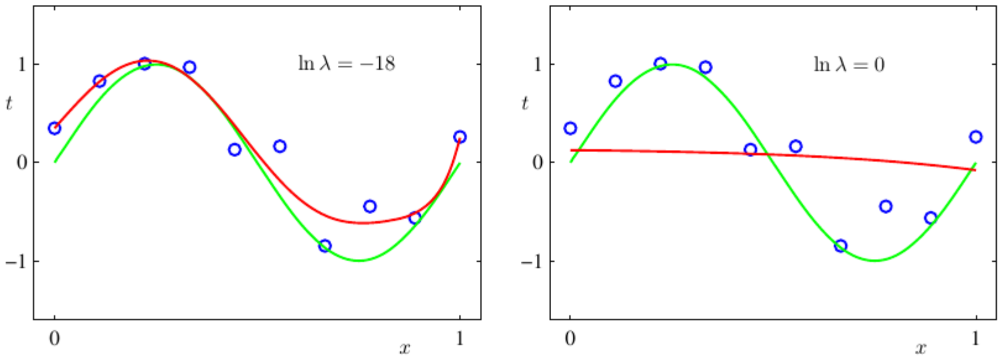
Choose the right complexity

Regularization

- Methods to reduce overfitting
- Heuristics: force model parameters to have small values
- Principled methods: use a priori information

Ridge Regression

$\#$ parameters $= \#$ data points



Ridge Regression

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	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

Ridge Regression