Statistical Machine Learning

Christian Walder

College of Engineering and Computer Science
The Australian National University

Canberra
Semester One, 2020.

(Many figures from C. M. Bishop, "Pattern Recognition and Machine Learning")

Outlines

Part II

Introduction to Statistical Machine Learning

Flavour of this course

- Formalise intuitions about problems
- Us
- Ge
- Pro
- Design and analysis of algorithms
- Numerical algorithms in python
- Understand the choices when designing mac
  methods

What is Machine Learning?

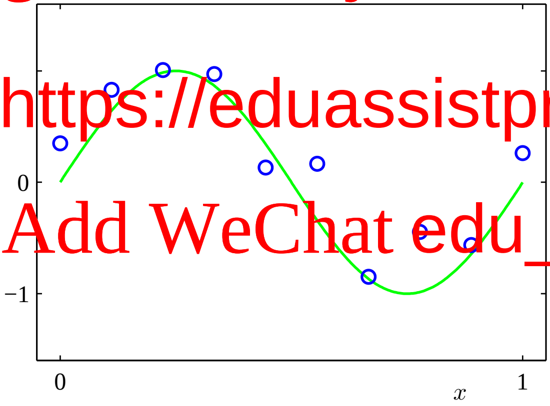
Definitio

A comput
respect t
if its perfor
with experience E.

Polynomial Curve Fitting

- some artificial data created from the function

$$\sin(2\pi x) + \text{random noise} \qquad x = 0, \dots, 1$$



Polynomial Curve Fitting - Input Specification

$\equiv ({}_1, \dots, {}_N)$

N

$\mathbf{x}$

$\mathbf{t}$

$x_i \in \mathbb{R} \quad i = 1, \ldots, N$

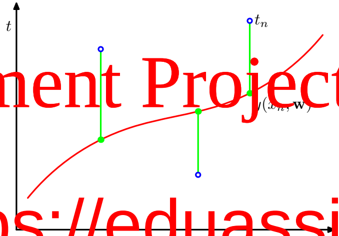
$t_i \in \mathbb{R} \quad i = 1, \ldots, N$

Polynomial Curve Fitting - Model Specification

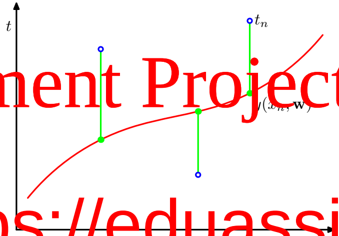
M : order of polynomial

$y(x, \mathbf{w}) =$ ^{2} ^{M}

- nonlinear function of x
- linear function of the unknown model paramet
- How can we find good parameters $\mathbf{w} =$


t
t_n
$y(x_n, \mathbf{w})$

Learning is Improving Performance



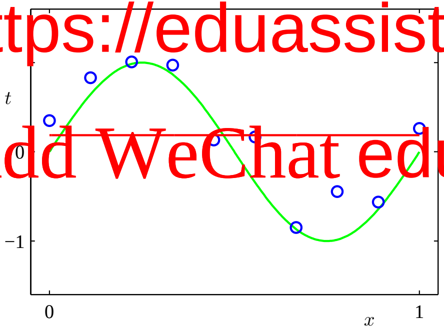
- Performance measure : Error between target prediction of the model for the training data

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

- unique minimum of $E(\mathbf{w})$ for argument $\mathbf{w}^\star$ under certain conditions (what are they?)

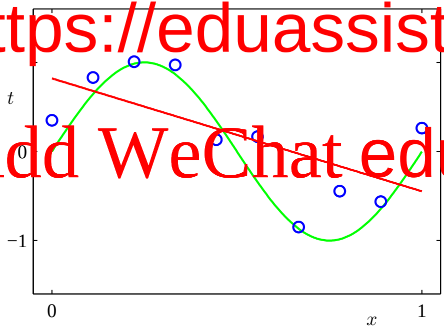
Model Comparison or Model Selection

$$y(x, \mathbf{w}) = \sum_{m=0}^{M} w_m x^m \Bigg|_{M=0}$$



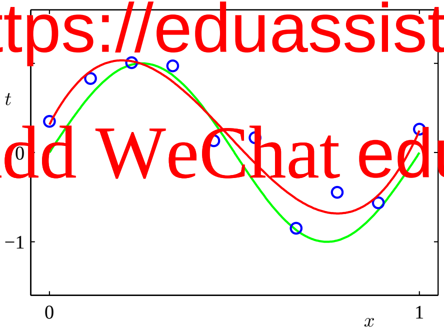
Model Comparison or Model Selection

$$y(x, \mathbf{w}) = \sum_{m=0}^{M} w_m x^m \Bigg|_{M=1}$$



$$y(x, \mathbf{w}) = \sum_{m=0}^{M} w_m x^m \Bigg|_{M=3}$$

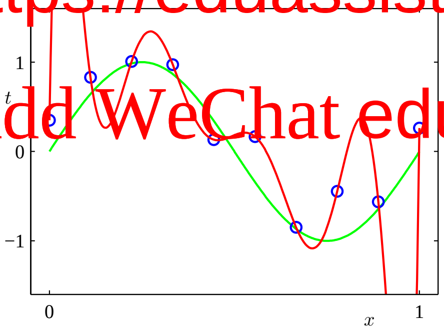
2 3



Model Comparison or Model Selection

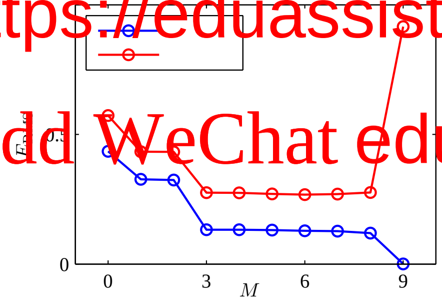
$$
\begin{aligned}
y(x, \mathbf{w}) &= \sum_{m=0}^{M} w_m \, x^m \Bigg|_{M=9} \\
&= w_0 + w_1 \, x + \cdots + w_8 \, x^8 + w_9 \, x^9
\end{aligned}
$$

- ove



Testing the Model

- Train the model and get $\mathbf{w}^\star$
- Get 100 new data points
- Root-mean-square (RMS) error



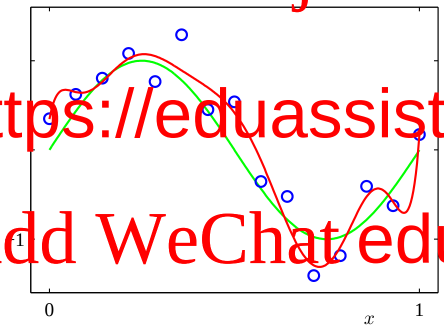
Testing the Model

	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
$w_6^\star$				-1
$w_7^\star$				1
$w_8^\star$				-
$w_9^\star$				

Table: Coefficients $\mathbf{w}^\star$ for polynomials of various order.

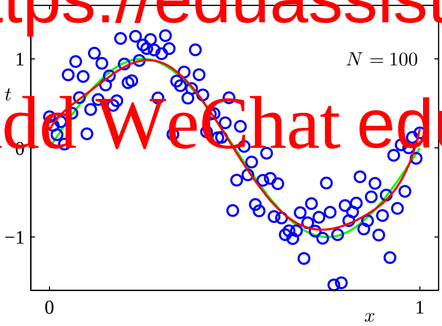
More Data

- $N = 15$



More Data

- $N = 100$
- heuristics : have no less than 5 to 10 times as many data points than parameters
- but number of parameters is not necessarily the most appropriate measure of model complexity !
- late



Regularisation

- How to constrain the growing of the coefficients $\mathbf{w}$?
- Add a regularisation term to the error function

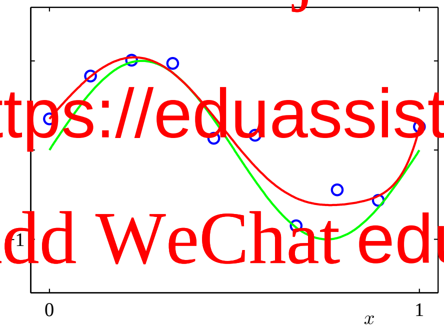
- Squared norm of the parameter vector

$$\|\mathbf{w}\|^2 \equiv \mathbf{w}^T \mathbf{w} = w_0^2 + w_1^2 + \cdots$$

- unique minimum of $E(\mathbf{w})$ for argument
  conditions (what are they for $\lambda = 0$? for $\lambda > 0$?)

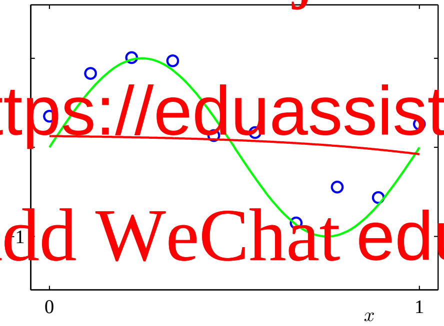
Regularisation

- $M = 9$



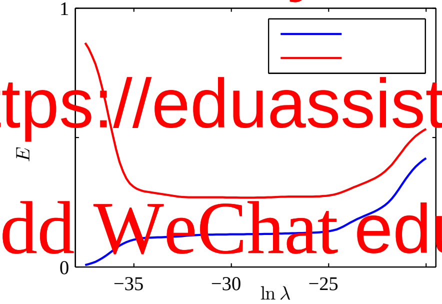
Regularisation

- $M = 9$



Regularisation

- $M = 9$

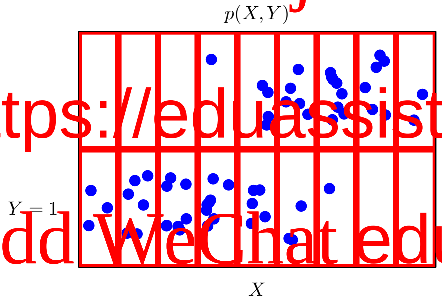


What is Machine Learning?

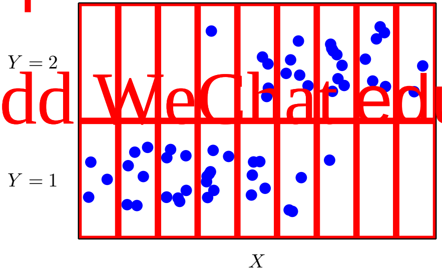
Definition (Mitchell, 1998)

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its perfor
with expe

- Ta
- Experience: $\mathbf{x}$ input examples, $\mathbf{t}$ output l
- Performance: squared error
- Model choice
- Regularisation
- **do not train on the test set!**


$p(X, Y)$
$Y = 1$
X

Y vs. X	a	b	c	d	e	f	g	h	i	sum
2	0	0	0	1	4	5	8	6	2	26
1	3	6	8	8	5	3	1	0	0	34
sum	3	6	8	9	9	8	9	6	2	60



Sum Rule

Y vs. X	a	b	c	d	e	f	g	h	i	sum
2	0	0	0	1	4	5	8	6	2	26
1	3	6	8	8	5	3	1	0	0	34
sum	3	6	8	9	9	8	9	6	2	60

$p($

$$= 1/60 + 8/60$$

$$p(X = d) = \sum_Y p(X = d,$$

$$p(X) = \sum_Y p(X, Y)$$

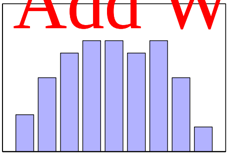
Sum Rule

Y vs. X	a	b	c	d	e	f	g	h	i	sum
2	0	0	0	1	4	5	8	6	2	26
1	3	6	8	8	5	3	1	0	0	34
sum	3	6	8	9	9	8	9	6	2	60

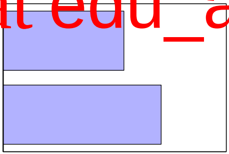
p

Y X

$p(X)$



X



Product Rule

Y vs. X	a	b	c	d	e	f	g	h	i	sum
2	0	0	0	1	4	5	8	6	2	26
1	3	6	8	8	5	3	1	0	0	34
sum	3	6	8	9	9	8	9	6	2	60

Conditional Probability

Calculat

$$p(Y = 1) = \sum_X p(X, Y = 1) =$$

$$p(X = d, Y = 1) = p(X = d \mid Y =$$

$$p(X, Y) = p(X \mid Y)\, p(Y)$$

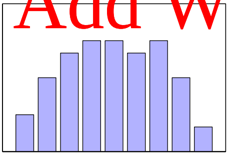
Another intuitive view is renormalisation of relative frequencies:

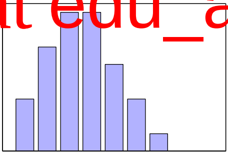
$$p(X \mid Y) = \frac{p(X, Y)}{p(Y)}$$

Y vs. X	a	b	c	d	e	f	g	h	i	sum
2	0	0	0	1	4	5	8	6	2	26
1	3	6	8	8	5	3	1	0	0	34
sum	3	6	8	9	9	8	9	6	2	60

p ... Y

$p(X)$





Sum Rule and Product Rule

- Sum Rule

- Pro

These rules form the basis of Bayesian machine lear
this course!

Bayes Theorem

Use product rule

$$p(X, Y) = p(X \mid Y)\, p(Y) = p(Y \mid X)\, p(X)$$

Bayes Theorem

$$p(Y \mid X) = \frac{p(X \mid Y)\, p(Y)}{p(X)}$$

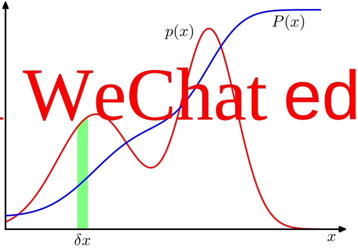
and

$$p(X) = \sum_Y p(X, Y)$$

$$= \sum_Y p(X \mid Y)\, p(Y) \qquad \text{(product rule)}$$

Probability Densities

- Real valued variable $x \in \mathbb{R}$
- Probability of x to fall in the interval $(x, x + \delta x)$ is given by $p(x)\delta x$ for infinitesimal small δx.
-

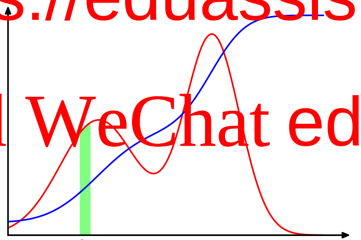


Constraints on $p(x)$

- Nonnegative

$$p(x) \geq 0$$

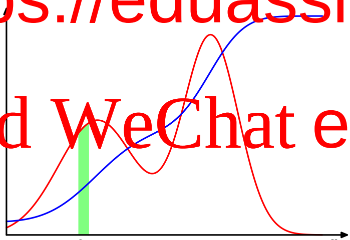
- Normalisation



Cumulative distribution function $P(x)$

$$P(x) = \int_{-\infty}^{x} p(z)\, \mathrm{d}z$$

or



Multivariate Probability Density

- Vector $\mathbf{x} \equiv (x_1, \dots, x_D)^T = \begin{bmatrix} x_1 \\ \vdots \\ x_D \end{bmatrix}$
- No
- Nor

$$\int_{-\infty}^{\infty} p(\mathbf{x})\,\mathrm{d}\mathbf{x} = 1.$$

- This means

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(\mathbf{x})\,\mathrm{d}x_1 \dots \mathrm{d}x_D = 1.$$

- Su

- Pro

$$p(x, y) = p(y \mid x)\, p(x$$

Expectations

- Weighted average of a function f(x) under the probability dist

$$\mathbb{E}\left[f\right] = \int p(x)\, f(x)\, \mathrm{d}x$$ proba

How to approximate $\mathbb{E}[f]$

- Given a finite number N of points x_n drawn from the probability distribution $p(x)$.
- Ap

$$\simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

- How to draw points from a probability distribution? Lecture coming about "Sampling"

Expection of a function of several variables

- arbitrary function $f(x, y)$

$\mathbb{E}_x$

$$\mathbb{E}_x [f(x, y)] = \quad p(x) f(x, y)\, \mathrm{d}x$$ probability density $p(x)$

- Note that $\mathbb{E}_x [f(x, y)]$ is a function of y.

Conditional Expectation

- arbitrary function $f(x)$

$$\mathbb{E}_x[f \mid y] = \sum_x p(x \mid y)\, f(x) \qquad \text{discrete distribution } p(x)$$

$$\mathbb{E}_x[f \mid y] = \quad p(x \mid y)\, f(x)\, \mathrm{d}x \qquad \text{probability density } p(x)$$

- Not
- Oth
- Wh $[\;[\;(\;)\;|\;]]$

- This must mean $\mathbb{E}_y[\mathbb{E}_x[f(x) \mid y]]$. (Why?)

$$\begin{aligned}\mathbb{E}_y[\mathbb{E}_x[f(x) \mid y]] &= \sum_y p(y)\, \mathbb{E}_x[f \mid y] = \\ &= \sum_{x,y} f(x)\, p(x,y) = \sum_x f(x)\, p(x) \\ &= \mathbb{E}_x[f(x)]\end{aligned}$$

Variance

- arbitrary function $f(x)$

- Sp

$$\operatorname{var}[x] = \mathbb{E}\left[(x - \mathbb{E}[x])^2\right] = \mathbb{E}\left[\right.$$

Covariance

- Two random variables $x \in \mathbb{R}$ and $y \in \mathbb{R}$

$$\operatorname{cov}[x, y] = \mathbb{E}_{x,y}\left[(x - \mathbb{E}[x])(y - \mathbb{E}[y])\right]$$
$$= \mathbb{E}_{x,y}[x\,y] - \mathbb{E}[x]\,\mathbb{E}[y]$$

- Wit

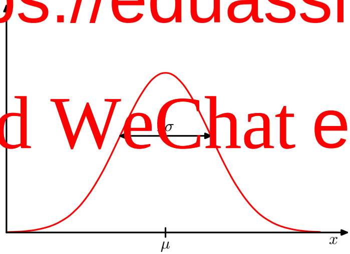
$$= {}_{x,y}[\quad] - {}_{x,y}[\quad] - {}_{x,y}[\quad] + {}_{x,y}[\quad]$$
$$= \mathbb{E}_{x,y}[x\,y] - b\,\underbrace{\mathbb{E}_{x,y}[x]}_{=\mathbb{E}_x[x]} - a\,\mathbb{E}$$
$$= \mathbb{E}_{x,y}[x\,y] - a\,b - a\,b + a\,b \qquad {}_{x,y}$$
$$= \mathbb{E}_{x,y}[x\,y] - \mathbb{E}[x]\,\mathbb{E}[y]$$

- Expresses how strongly x and y vary together. If x and y are independent, their covariance vanishes.

- Tw

$$= \mathbb{E}_{\mathbf{x},\mathbf{y}} \quad \mathbf{x}\,\mathbf{y} \quad - \mathbb{E}\left[\mathbf{x}\right]\mathbb{E} \quad \mathbf{y}$$

- $x \in \mathbb{R}$
- Gaussian Distribution with mean μ and variance σ^2



The Gaussian Distribution

- $\mathcal{N}(x \,|\, \mu, \sigma^2) > 0$
- $\int_{-\infty}^{\infty} \mathcal{N}(x \,|\, \mu, \sigma^2) \,\mathrm{d}x = 1$
- Expectation over x

- Exp

$$\mathbb{E}\left[x^2\right] = \int_{-\infty}^{\infty} \mathcal{N}(x \,|\, \mu, \sigma^2)\, x^2 \,\mathrm{d}x$$

- Variance of x

$$\operatorname{var}[x] = \mathbb{E}\left[x^2\right] - \mathbb{E}\left[x\right]^2 = \sigma^2$$

Strategy in this course

- Estimate best predictor = training = learning
  Giv

  1.
  2.
  3.
  4. Calculate the optimal parameter ($\mathbf{w}$)
  5. Model uncertainty using the Bayesian appro
  6. Implement and compute (the algorithm in py
  7. Interpret and diagnose results