Introduction to Statistical Machine Learning

Christfried Webers

Statistical Machine Learning Group
NICTA
and
College of Engineering and Computer Science
The Australian National University

Canberra
February – June 2013

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

Outlines

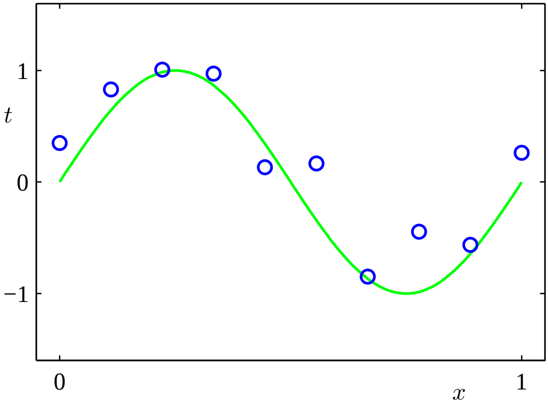
Part II

Introduction

Polynomial Curve Fitting

- some artificial data created from the function

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

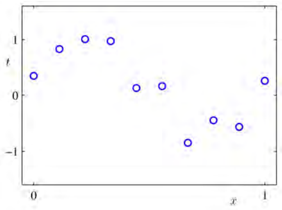


Polynomial Curve Fitting - Input Specification

$$N = 10$$

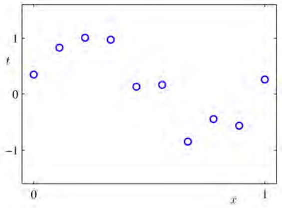
$$\mathbf{x} \equiv (x_1, \ldots, x_N)^T$$

$$\mathbf{t} \equiv (t_1, \ldots, t_N)^T$$



Polynomial Curve Fitting - Input Specification

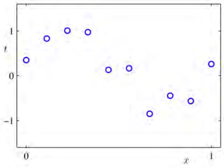
$$
\begin{aligned}
N &= 10 \\
\mathbf{x} &\equiv (x_1, \dots, x_N)^T \\
\mathbf{t} &\equiv (t_1, \dots, t_N)^T \\
x_i &\in \mathbb{R} \quad i = 1, \dots, N \\
t_i &\in \mathbb{R} \quad i = 1, \dots, N
\end{aligned}
$$



Polynomial Curve Fitting - Model Specification

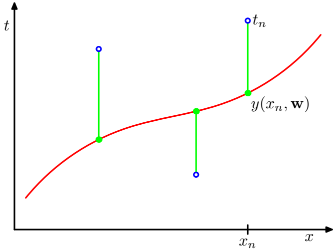
M : order of polynomial

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M$$
$$= \sum_{m=0}^{M} w_m x^m$$

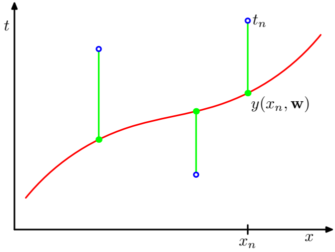


- nonlinear function of x
- linear function of the unknown model parameter $\mathbf{w}$
- How can we find good parameters $\mathbf{w} = (w_1, \ldots, w_M)^T$?

Learning is Improving Performance



Learning is Improving Performance



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

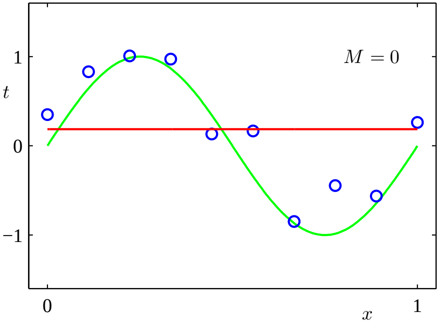
$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \left(y(x_n, \mathbf{w}) - t_n\right)^2$$

- unique minimum of $E(\mathbf{w})$ for argument $\mathbf{w}^{\star}$

Model Comparison or Model Selection

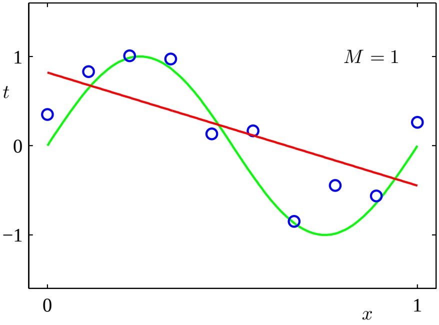
$$y(x, \mathbf{w}) = \sum_{m=0}^{M} w_m x^m \quad \Big|_{M=0}$$

$$= w_0$$



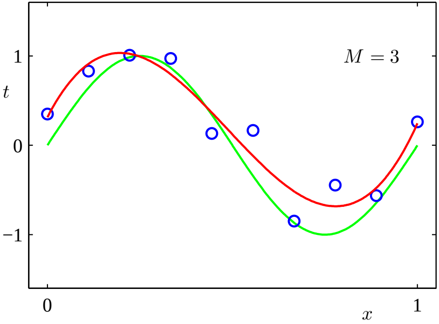
Model Comparison or Model Selection

$$y(x, \mathbf{w}) = \sum_{m=0}^{M} w_m x^m \quad \Bigg|_{M=1}$$
$$= w_0 + w_1 x$$



Model Comparison or Model Selection

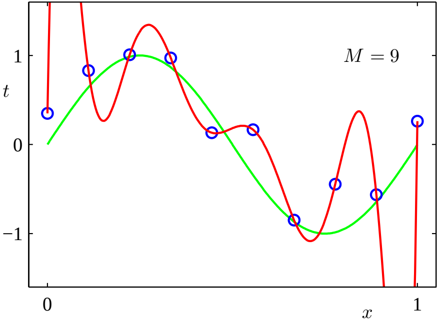
$$y(x, \mathbf{w}) = \sum_{m=0}^{M} w_m x^m \Bigg|_{M=3}$$
$$= w_0 + w_1 x + w_2 x^2 + w_3 x^3$$



Model Comparison or Model Selection

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

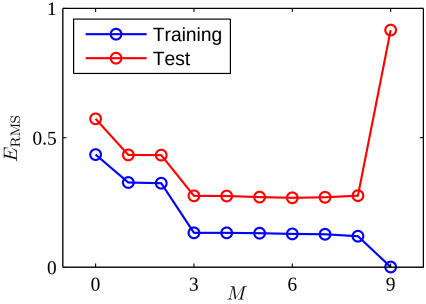
- overfitting



Testing the Model

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

$$E_{\text{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$$



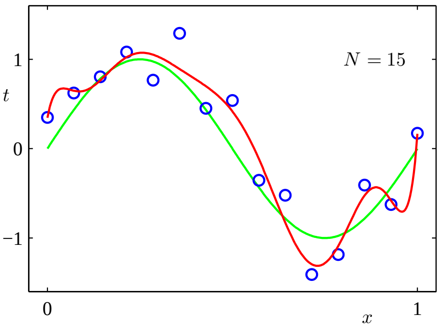
Testing the Model

	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
$w_2^\star$			-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

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

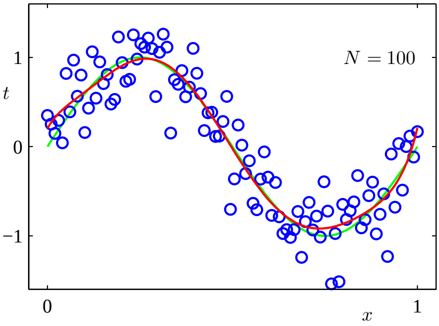
More Data

- $N = 15$



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 !
- later: Bayesian approach



Regularisation

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

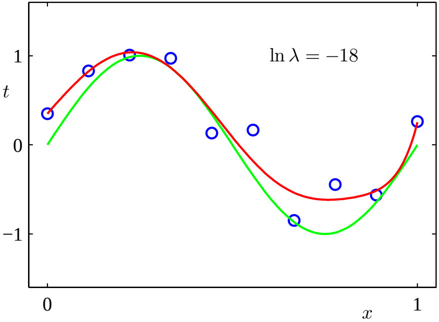
$$\widetilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \left(y(x_n, \mathbf{w}) - t_n \right)^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

- Squared norm of the parameter vector $\mathbf{w}$

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

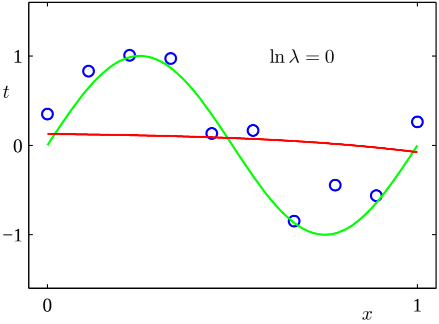
Regularisation

- $M = 9$



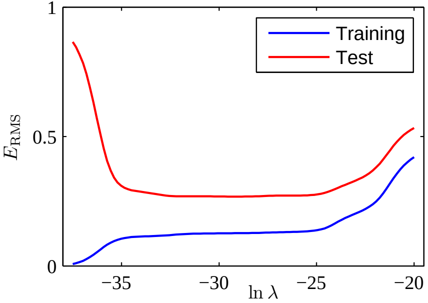
Regularisation

- $M = 9$

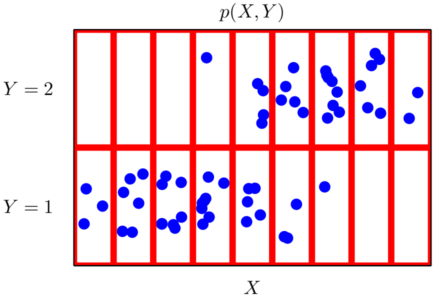


Regularisation

- $M = 9$

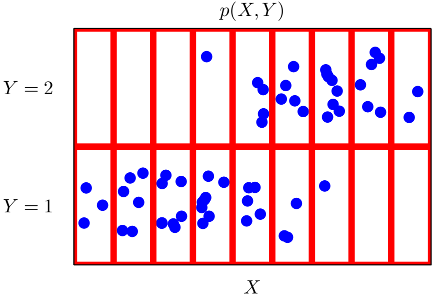


Probability Theory



Probability Theory

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

$$p(X = d, Y = 1) = 8/60$$

$$\begin{aligned} p(X = d) &= p(X = d, Y = 2) + p(X = d, Y = 1) \\ &= 1/60 + 8/60 \end{aligned}$$

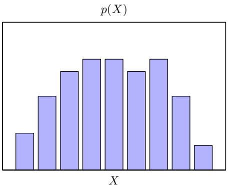
$$p(X = d) = \sum_{Y} p(X = d, Y)$$

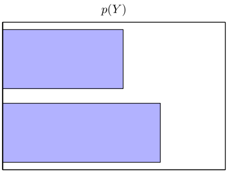
$$p(X) = \sum_{Y} p(X, Y)$$

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

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





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

$$p(X = d \mid Y = 1) = 8/34$$

Calculate $p(Y = 1)$:

$$p(Y = 1) = \sum_X p(X, Y = 1) = 34/60$$

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

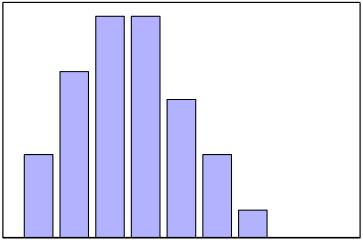
$$p(X, Y) = p(X \mid Y)\, p(Y)$$

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

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

$p(X|Y = 1)$



Sum Rule and Product Rule

- Sum Rule

$$p(X) = \sum_{Y} p(X, Y)$$

- Product Rule

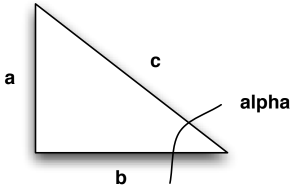
$$p(X, Y) = p(X \mid Y)\, p(Y)$$

Why not using Fractions?

- Why not using pairs of numbers (s, t) such that $p(X, Y) = s/t$ (e.g. $s = 8$, $t = 60$)?

Why not using Fractions?

- Why not using pairs of numbers (s, t) such that $p(X, Y) = s/t$ (e.g. $s = 8$, $t = 60$)?
- Why not using pairs of numbers (a, c) instead of $\sin(\text{alpha}) = a/c$?



Bayes Theorem

Use product rule

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

Bayes Theorem

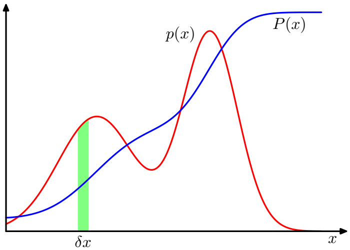
$$p(Y \mid X) = \frac{p(X \mid Y)\, p(Y)}{p(X)}$$

and

$$\begin{aligned} p(X) &= \sum_Y p(X, Y) && \text{(sum rule)} \\ &= \sum_Y p(X \mid Y)\, p(Y) && \text{(product rule)} \end{aligned}$$

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.
- $$p(x \in (a, b)) = \int_a^b p(x)\, \mathrm{d}x.$$



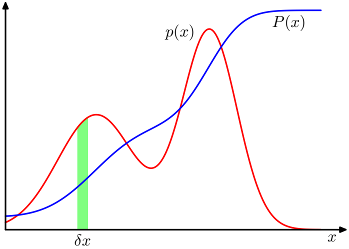
Constraints on $p(x)$

- Nonnegative

$$p(x) \geq 0$$

- Normalisation

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

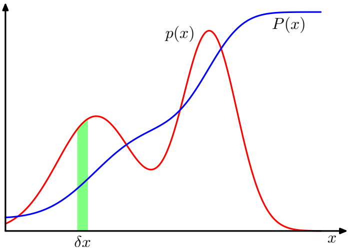


Cumulative distribution function $P(x)$

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

or

$$\frac{d}{dx} P(x) = p(x)$$



Multivariate Probability Density

- Vector $\mathbf{x} \equiv (x_1, \ldots, x_D)^T = \begin{bmatrix} x_1 \\ \vdots \\ x_D \end{bmatrix}$
- Nonnegative
$$p(\mathbf{x}) \geq 0$$
- Normalisation
$$\int_{-\infty}^{\infty} p(\mathbf{x}) \, \mathrm{d}\mathbf{x} = 1.$$
- This means
$$\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} p(\mathbf{x}) \, \mathrm{d}x_1 \ldots \, \mathrm{d}x_D = 1.$$

Sum and Product Rule for Probability Densities

- Sum Rule

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

- Product Rule

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

Expectations

- Weighted average of a function f(x) under the probability distribution $p(x)$

$$\mathbb{E}\left[f\right] = \sum_x p(x) f(x) \qquad \text{discrete distribution } p(x)$$

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

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

- Given a finite number N of points x_n drawn from the probability distribution $p(x)$.
- Approximate the expectation by a finite sum:

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

- How to draw points from a probability distribution $p(x)$? Lecture coming about "Sampling"

Expection of a function of several variables

- arbitrary function $f(x, y)$

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

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

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

Conditional Expectation

- arbitrary function $f(x)$

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

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

- Note that $\mathbb{E}_x\left[f \mid y\right]$ is a function of y.
- Other notation used in the literature : $\mathbb{E}_{x|y}\left[f\right]$.
- What is $\mathbb{E}\left[\mathbb{E}\left[f(x) \mid y\right]\right]$? Can we simplify it?
- This must mean $\mathbb{E}_y\left[\mathbb{E}_x\left[f(x) \mid y\right]\right]$. (Why?)

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

ISML 2013

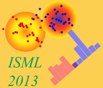
Variance

- arbitrary function $f(x)$

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

- Special case: $f(x) = x$

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



Covariance

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

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

- With $\mathbb{E}[x] = a$ and $\mathbb{E}[y] = b$

$$\begin{aligned}\operatorname{cov}[x, y] &= \mathbb{E}_{x,y}\left[(x - a)(y - b)\right] \\ &= \mathbb{E}_{x,y}[x\, y] - \mathbb{E}_{x,y}[x\, b] - \mathbb{E}_{x,y}[a\, y] + \mathbb{E}_{x,y}[a\, b] \\ &= \mathbb{E}_{x,y}[x\, y] - b\, \underbrace{\mathbb{E}_{x,y}[x]}_{=\mathbb{E}_x[x]} - a\, \underbrace{\mathbb{E}_{x,y}[y]}_{=\mathbb{E}_y[y]} + a\, b\, \underbrace{\mathbb{E}_{x,y}[1]}_{=1} \\ &= \mathbb{E}_{x,y}[x\, y] - a\, b - a\, b + a\, b = \mathbb{E}_{x,y}[x\, y] - a\, b \\ &= \mathbb{E}_{x,y}[x\, y] - \mathbb{E}[x]\,\mathbb{E}[y]\end{aligned}$$

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

Covariance for Vector Valued Variables

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

$$\begin{aligned}\operatorname{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{y}^{\mathbf{T}} - \mathbb{E}\left[\mathbf{y}^{\mathbf{T}}\right])\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\mathbf{x}\,\mathbf{y}^{\mathbf{T}}\right] - \mathbb{E}[\mathbf{x}]\,\mathbb{E}\left[\mathbf{y}^{\mathbf{T}}\right]\end{aligned}$$