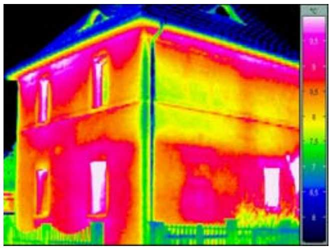
Regression
(Part I)

nanopoulos@ismll.de

The regression problem

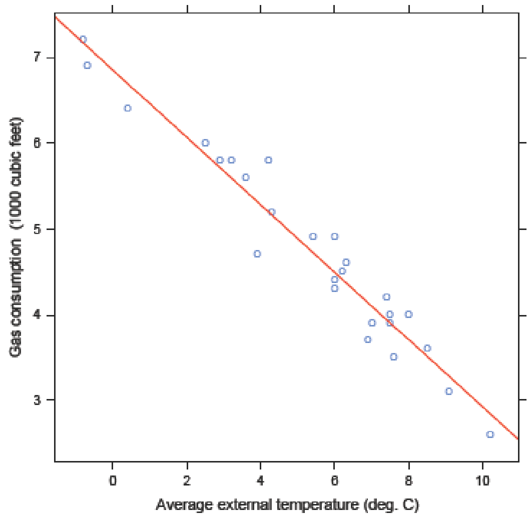
Example: how does gas consumption depend on external temperature? (Whiteside, 1960s).



weekly measurements of
- average external temperature
- total gas consumption (in 1000 cubic feet)

- How does gas consumption depend on external temperature?
- How much gas is needed for a given temperature ?

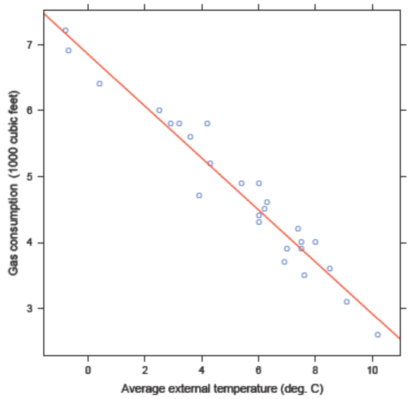
Linear model



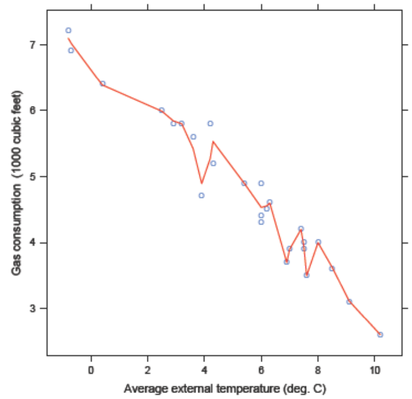
Outline

- Introduction
- Simple linear regression
- Simple polynomial regression
- Maximum likelihood estimation
- Maximum posterior estimation

Is linear the only option?



linear model



more flexible model

5

Variable types

numerical / interval-scaled / quantitative

where differences and quotients etc. are meaningful, e.g., temperature, size, weight

nominal / discrete / categorical / qualitative

where differences and quotients are not defined, usually with a finite, enumerated domain, e.g., X := {red, green, blue}

ordinal / ordered categorical

where levels are ordered, but differences and quotients are not defined, usually with a finite, enumerated domain, e.g., X := {small, medium, large}

6

Definitions: predictors and target

Let

$X_1, X_2, \ldots, X_p$ be random variables called **predictors** (or **inputs**, **covariates**).
Let $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_p$ be their domains.
We write shortly

$$X := (X_1, X_2, \ldots, X_p)$$

for the vector of random predictor variables and

$$\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times X_p$$

for its domain.

Y be a random variable called **target** (or **output**, **response**).
Let $\mathcal{Y}$ be its domain.

$\mathcal{D} \subseteq \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ be a (multi)set of instances of the unknown joint distribution $p(X, Y)$ of predictors and target called **data**.
$\mathcal{D}$ is often written as enumeration

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

7

Definitions: regression classification

The task of regression and classification is
to predict Y based on X,
i.e., to estimate

$$r(x) := E(Y \mid X = x) = \int y\, p(y|x) dx$$

based on data (called **regression function**).

If Y is numerical, the task is called **regression**.

If Y is nominal, the task is called **classification**.

8

Outline

- Introduction
- Simple linear regression
- Simple polynomial regression
- Maximum likelihood estimation
- Maximum posterior estimation

Simple linear regression

Make it simple:

- the predictor X is simple, i.e., one-dimensional ($X = X_1$).
- $r(x)$ is assumed to be linear:

$$r(x) = \beta_0 + \beta_1 x$$

- assume that the variance does not depend on x:

$$Y = \beta_0 + \beta_1 x + \epsilon, \quad E(\epsilon|x) = 0, V(\epsilon|x) = \sigma^2$$

- 3 parameters:
  β_0 **intercept** (sometimes also called bias)
  β_1 **slope**
  σ^2 **variance**

Parameters

parameter estimates

$$\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$$

fitted line

$$\hat{r}(x) := \hat{\beta}_0 + \hat{\beta}_1 x$$

predicted / fitted values

$$\hat{y}_i := \hat{r}(x_i)$$

residuals

$$\hat{\epsilon}_i := y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

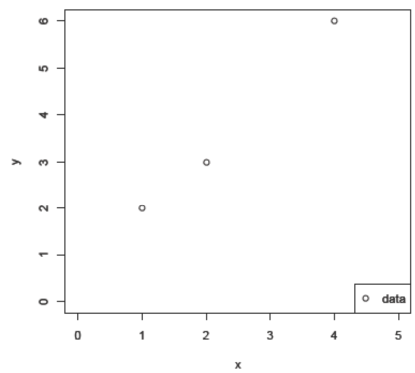
residual sums of squares (RSS)

$$\text{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

11

Parameter estimation (example)

Example:
Given the data $\mathcal{D} := \{(1,2),(2,3),(4,6)\}$, predict a value for $x = 3$.



12

Parameter estimation (example)

Example:
Given the data $\mathcal{D} := \{(1,2),(2,3),(4,6)\}$, predict a value for $x = 3$.

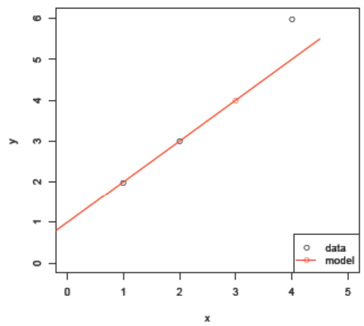
Line through first two points:

$$\hat\beta_1 = \frac{y_2 - y_1}{x_2 - x_1} = 1$$
$$\hat\beta_0 = y_1 - \hat\beta_1 x_1 = 1$$

RSS:

i	y_i	$\hat y_i$	$(y_i - \hat y_i)^2$
1	2	2	0
2	3	3	0
3	6	5	1
$\sum$			1

$\hat r(3) = 4$



Parameter estimation (example)

Example:
Given the data $\mathcal{D} := \{(1,2),(2,3),(4,6)\}$, predict a value for $x = 3$.

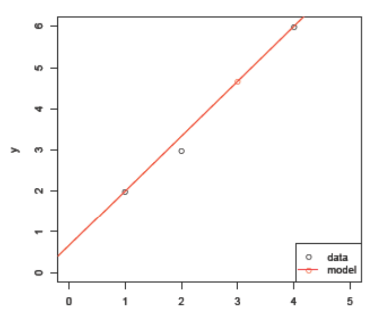
Line through first and last point:

$$\hat\beta_1 = \frac{y_3 - y_1}{x_3 - x_1} = 4/3 = 1.333$$
$$\hat\beta_0 = y_1 - \hat\beta_1 x_1 = 2/3 = 0.667$$

RSS:

i	y_i	$\hat y_i$	$(y_i - \hat y_i)^2$
1	2	2	0
2	3	3.333	0.111
3	6	6	0
$\sum$			0.111

$\hat r(3) = 4.667$



Least Squares Estimation (LSE)

In principle, there are many different methods to estimate the parameters $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\sigma}^2$ from data — depending on the properties the solution should have.

The **least squares estimates** are those parameters that minimize

$$\text{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

They can be written in closed form as follows:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

15

LSE: proof

Proof (1/2):

$$\text{RSS} = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$
$$\frac{\partial\, \text{RSS}}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} 2(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))(-1) \overset{!}{=} 0$$
$$\implies \quad n\hat{\beta}_0 = \sum_{i=1}^{n} y_i - \hat{\beta}_1 x_i$$

16

LSE: proof

Proof (2/2):

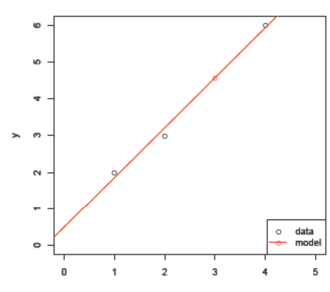
$$
\begin{aligned}
\text{RSS} &= \sum_{i=1}^{n} (y_i - (\hat\beta_0 + \hat\beta_1 x_i))^2 \\
&= \sum_{i=1}^{n} (y_i - (\bar y - \hat\beta_1 \bar x) - \hat\beta_1 x_i)^2 \\
&= \sum_{i=1}^{n} (y_i - \bar y - \hat\beta_1 (x_i - \bar x))^2 \\
\frac{\partial\, \text{RSS}}{\partial \hat\beta_1} &= \sum_{i=1}^{n} 2(y_i - \bar y - \hat\beta_1 (x_i - \bar x))(-1)(x_i - \bar x) \stackrel{!}{=} 0 \\
\implies \quad \hat\beta_1 &= \frac{\sum_{i=1}^{n} (y_i - \bar y)(x_i - \bar x)}{\sum_{i=1}^{n} (x_i - \bar x)^2}
\end{aligned}
$$

17

LSE: example

Given the data $\mathcal{D} := \{(1,2),(2,3),(4,6)\}$, predict a value for $x = 3$.
Assume simple linear model.
$\bar x = 7/3$, $\bar y = 11/3$.

i	$x_i - \bar x$	$y_i - \bar y$	$(x_i - \bar x)^2$	$(x_i - \bar x)(y_i - \bar y)$
1	$-4/3$	$-5/3$	$16/9$	$20/9$
2	$-1/3$	$-2/3$	$1/9$	$2/9$
3	$5/3$	$7/3$	$25/9$	$35/9$
$\sum$			$42/9$	$57/9$



$$
\hat\beta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar x)(y_i - \bar y)}{\sum_{i=1}^{n} (x_i - \bar x)^2} = 57/42 = 1.357
$$

$$
\hat\beta_0 = \bar y - \hat\beta_1 \bar x = \frac{11}{3} - \frac{57}{42} \cdot \frac{7}{3} = \frac{63}{126} = 0.5
$$

RSS:

i	y_i	$\hat y_i$	$(y_i - \hat y_i)^2$
1	2	1.857	0.020
2	3	3.214	0.046
3	6	5.929	0.005
$\sum$			0.071

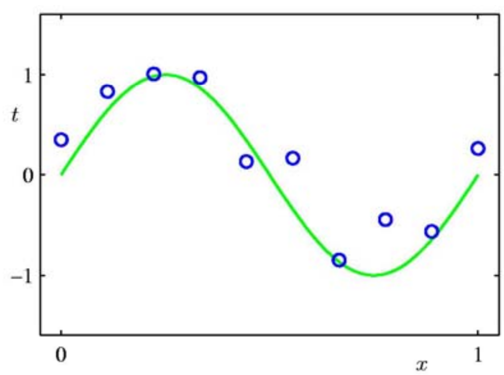
$\hat r(3) = 4.571$

18

Outline

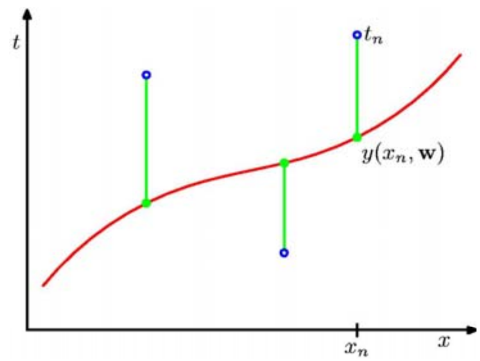
- Introduction
- Simple linear regression
- Simple polynomial regression
- Maximum likelihood estimation
- Maximum posterior estimation

Simple polynomial regression



$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

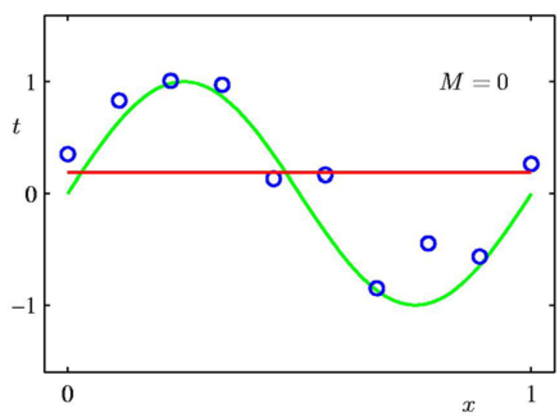
Sum-of-Squares Error Function



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

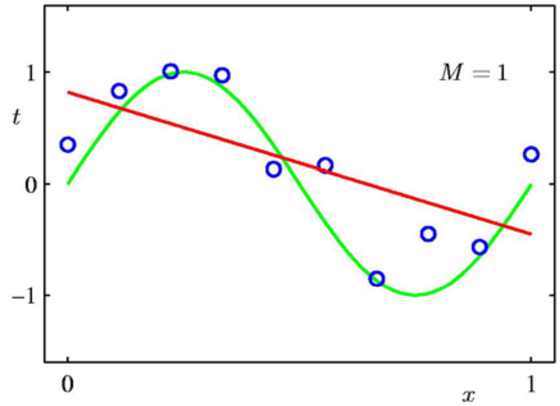
21

0th Order Polynomial

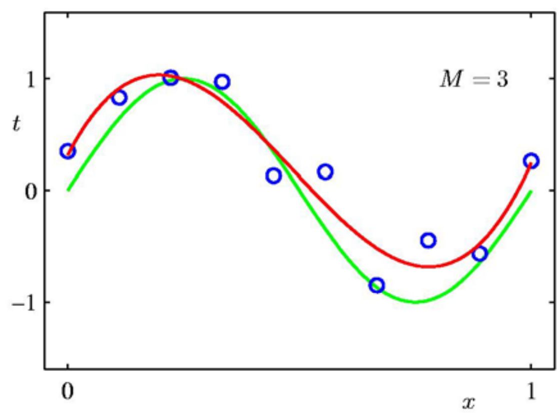


22

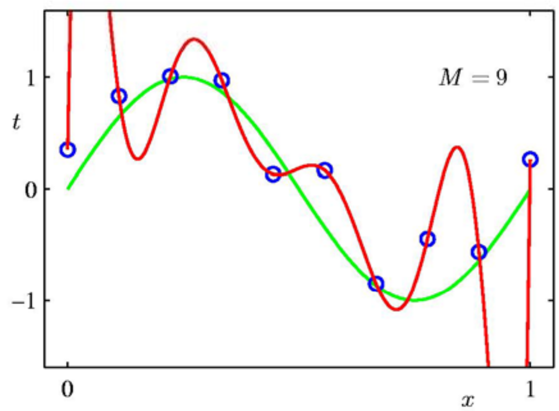
1^{st} Order Polynomial



3^{rd} Order Polynomial



9^{th} Order Polynomial

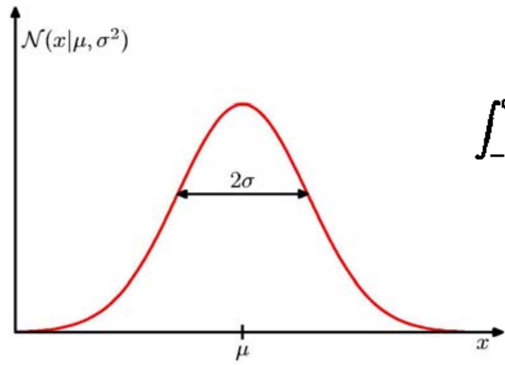


Outline

- Introduction
- Simple linear regression
- Simple regression with polynomials
- Maximum likelihood estimation
- Maximum posterior estimation

The Gaussian Distribution

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) \mathrm{d}x = 1$$

27

Gaussian Mean and Variance

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

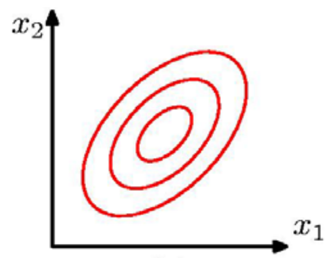
$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2$$

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

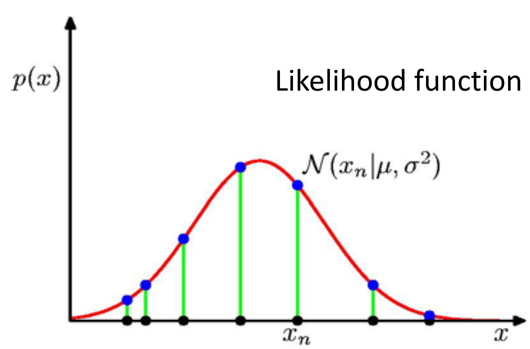
28

The Multivariate Gaussian

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$



Gaussian Parameter Estimation



$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu, \sigma^2\right)$$

Maximum (Log) Likelihood

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

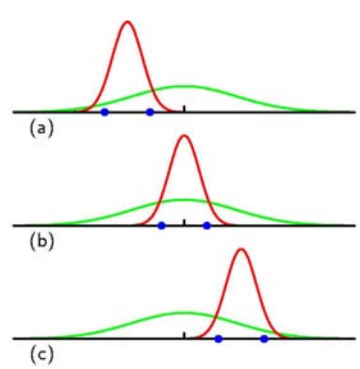
$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

Properties of μ_{ML} and σ^2_{ML}

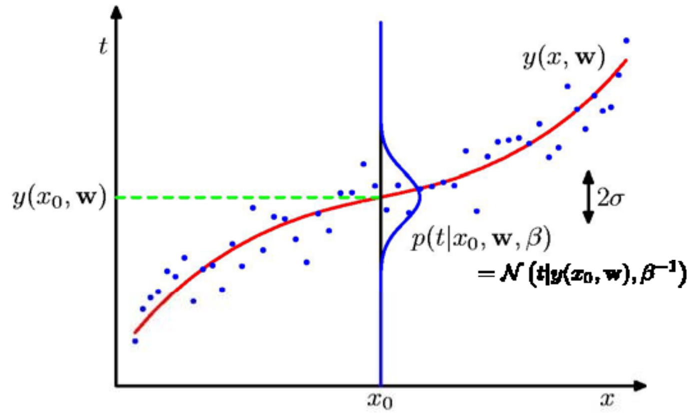
$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu$$

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \left(\frac{N-1}{N}\right)\sigma^2$$

$$\begin{aligned}\widetilde{\sigma}^2 &= \frac{N}{N-1}\sigma^2_{\mathrm{ML}} \\ &= \frac{1}{N-1}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2\end{aligned}$$



Polynomial regression re-visited



Maximum Likelihood

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n | y(x_n, \mathbf{w}), \beta^{-1}\right)$$

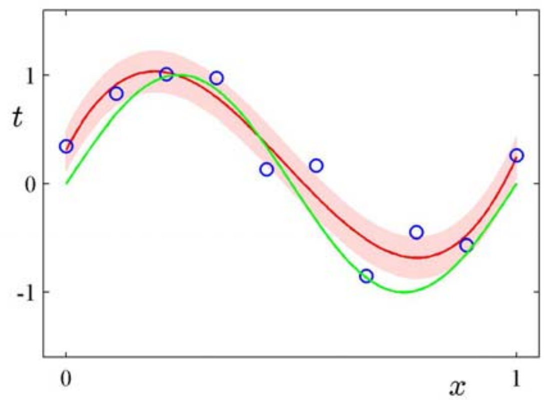
$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \underbrace{-\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2}_{\beta E(\mathbf{w})} + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$$

Determine $\mathbf{w}_{\mathrm{ML}}$ by minimizing sum-of-squares error, $E(\mathbf{w})$.

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

Predictive Distribution

$$p(t|x, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}\left(t|y(x, \mathbf{w}_{\mathrm{ML}}), \beta_{\mathrm{ML}}^{-1}\right)$$



Outline

- Introduction
- Simple linear regression
- Simple regression with polynomials
- Maximum likelihood estimation
- Maximum posterior estimation

Maximum posterior estimation (MAP)

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}\,0, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\alpha)$$

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

Determine $\mathbf{w}_{\mathbf{MAP}}$ by minimizing regularized sum-of-squares error, $\widetilde{E}(\mathbf{w})$.