Confidence in parameter estimates

Story so far

Predictions

Prediction

Likelihood for model selection

Summary

Machine Learning
Lecture 6 - Modelling the noise

Devdatt Dubhashi
dubhashi@chalmers.se

Dept. of Computer Science and Engg.
Chalmers University

January 30, 2017

Optimum parameters

- We have point estimates of our parameters.
- How confident should we be in them?
  - If we changed them a little bit, would the model still be good?

Confidence in parameter estimates

- Imagine there are **true** parameters, **w** and σ^2.

Confidence in parameter estimates

- Imagine there are **true** parameters, $\mathbf{w}$ and σ^2.
- How good our our estimates $\widehat{\mathbf{w}}$ and $\widehat{\sigma^2}$?
  - Are they correct (on average)?
  - If we could keep adding data, would we converge on the true value?

Confidence in parameter estimates
Story so far
Predictions
Prediction
Likelihood for model selection
Summary

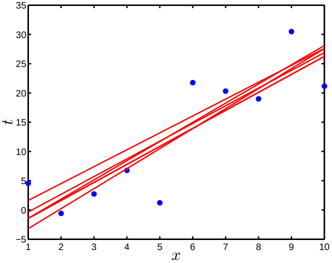
Confidence in parameter estimates

- Imagine there are **true** parameters, $\mathbf{w}$ and σ^2.
- How good our our estimates $\widehat{\mathbf{w}}$ and $\widehat{\sigma^2}$?
  - Are they correct (on average)?
  - If we could keep adding data, would we converge on the true value?
- How confident should we be in our estimates?
  - Could we change parameters a little bit and still have a good model?

Confidence in parameter estimates
Story so far
Predictions
Prediction
Likelihood for model selection
Summary

Confidence in parameter estimates

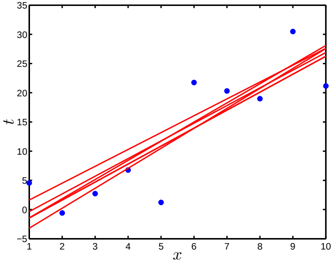
- Imagine there are **true** parameters, $\mathbf{w}$ and σ^2.
- How good our our estimates $\widehat{\mathbf{w}}$ and $\widehat{\sigma^2}$?
  - Are they correct (on average)?
  - If we could keep adding data, would we converge on the true value?
- How confident should we be in our estimates?
  - Could we change parameters a little bit and still have a good model?

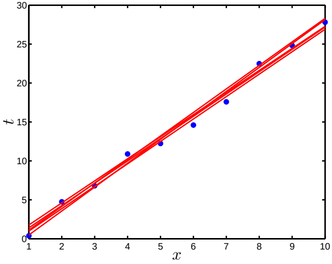


Confidence in parameter estimates
Story so far
Predictions
Prediction
Likelihood for model selection
Summary

Confidence in parameter estimates

- Imagine there are **true** parameters, $\mathbf{w}$ and σ^2.
- How good our our estimates $\widehat{\mathbf{w}}$ and $\widehat{\sigma^2}$?
  - Are they correct (on average)?
  - If we could keep adding data, would we converge on the true value?
- How confident should we be in our estimates?
  - Could we change parameters a little bit and still have a good model?





Back to the model...

- Parameter estimates:

$$\begin{aligned}\widehat{\mathbf{w}} &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t} \\ \widehat{\sigma^2} &= \frac{1}{N}(\mathbf{t} - \mathbf{X}\mathbf{w})^\mathsf{T}(\mathbf{t} - \mathbf{X}\mathbf{w})\end{aligned}$$

Back to the model...

- Parameter estimates:

$$\begin{aligned} \widehat{\mathbf{w}} &= (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t} \\ \widehat{\sigma^2} &= \frac{1}{N}(\mathbf{t} - \mathbf{X}\mathbf{w})^{\mathsf{T}}(\mathbf{t} - \mathbf{X}\mathbf{w}) \end{aligned}$$

- **True** values: $\mathbf{w}$, σ^2

Back to the model...

- Parameter estimates:

$$\widehat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$$
$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t} - \mathbf{X}\mathbf{w})^\mathsf{T}(\mathbf{t} - \mathbf{X}\mathbf{w})$$

- **True** values: $\mathbf{w}$, σ^2
- Our model:

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I})$$

Back to the model...

- Parameter estimates:

$$\begin{aligned} \widehat{\mathbf{w}} &= (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t} \\ \widehat{\sigma^2} &= \frac{1}{N}(\mathbf{t}-\mathbf{X}\mathbf{w})^{\mathsf{T}}(\mathbf{t}-\mathbf{X}\mathbf{w}) \end{aligned}$$

- **True** values: $\mathbf{w}$, σ^2
- Our model:

$$p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w},\sigma^2\mathbf{I})$$

- What's $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$?

Confidence in parameter estimates
Story so far
Predictions
Prediction
Likelihood for model selection
Summary

Back to the model...

- Parameter estimates:

$$\widehat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$$
$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}-\mathbf{X}\mathbf{w})^\mathsf{T}(\mathbf{t}-\mathbf{X}\mathbf{w})$$

- **True** values: $\mathbf{w}$, σ^2
- Our model:

$$p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w},\sigma^2\mathbf{I})$$

- What's $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$?
  - What do we expect our parameter estimate to be?

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$

We'll try and find $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$ in terms of the true value $\mathbf{w}$:

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$

We'll try and find $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$ in terms of the true value $\mathbf{w}$:

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} = \int \widehat{\mathbf{w}} p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)\, d\mathbf{t}$$

Confidence in parameter estimates

Story so far

Predictions

Prediction

Likelihood for model selection

Summary

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$

We'll try and find $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$ in terms of the true value $\mathbf{w}$:

$$\begin{aligned}
\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} &= \int \widehat{\mathbf{w}} p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)\, d\mathbf{t} \\
&= \int (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}\, p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)\, d\mathbf{t} \\
&= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\right\}
\end{aligned}$$

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$

We'll try and find $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$ in terms of the true value $\mathbf{w}$:

$$\begin{aligned}
\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} &= \int \widehat{\mathbf{w}} p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)\, d\mathbf{t} \\
&= \int (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t} p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)\, d\mathbf{t} \\
&= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\right\} \\
&= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} \\
\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} &= \mathbf{I}\mathbf{w} = \mathbf{w}
\end{aligned}$$

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$

We'll try and find $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$ in terms of the true value $\mathbf{w}$:

$$
\begin{aligned}
\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} &= \int \widehat{\mathbf{w}} p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)\, d\mathbf{t} \\
&= \int (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t} p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)\, d\mathbf{t} \\
&= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\right\} \\
&= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} \\
\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} &= \mathbf{I}\mathbf{w} = \mathbf{w}
\end{aligned}
$$

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$

Confidence in parameter estimates

Story so far

Predictions

Prediction

Likelihood for model selection

Summary

We'll try and find $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}$ in terms of the true value $\mathbf{w}$:

$$
\begin{aligned}
\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} &= \int \widehat{\mathbf{w}} p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)\; d\mathbf{t} \\
&= \int (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t} p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)\; d\mathbf{t} \\
&= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\right\} \\
&= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} \\
\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} &= \mathbf{I}\mathbf{w} = \mathbf{w}
\end{aligned}
$$

$\widehat{\mathbf{w}}$ is unbiased

On average, we expect our estimate to equal the true value!

$\mathrm{cov}\{\widehat{\mathbf{w}}\}$

- What does $\mathrm{cov}\{\widehat{\mathbf{w}}\}$ tell us?

$\mathrm{cov}\{\widehat{\mathbf{w}}\}$

- What does $\mathrm{cov}\{\widehat{\mathbf{w}}\}$ tell us?
- Recall the linear model, $\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$

$\text{cov}\{\widehat{\mathbf{w}}\}$

- What does $\text{cov}\{\widehat{\mathbf{w}}\}$ tell us?
- Recall the linear model, $\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$

$$\text{cov}\{\widehat{\mathbf{w}}\} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

$\mathrm{cov}\{\widehat{\mathbf{w}}\}$

- What does $\mathrm{cov}\{\widehat{\mathbf{w}}\}$ tell us?
- Recall the linear model, $\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$

$$\mathrm{cov}\{\widehat{\mathbf{w}}\} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

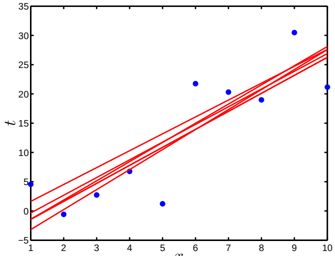
- Tells us how well defined the parameters are by the data. How much can the parameters vary and still give a **good** model.
  - a and c – how much can we change w_0 and w_1. b – how the values should be changed together.

$\text{cov}\{\widehat{\mathbf{w}}\}$

- What does $\text{cov}\{\widehat{\mathbf{w}}\}$ tell us?
- Recall the linear model, $\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$

$$\text{cov}\{\widehat{\mathbf{w}}\} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

- Tells us how well defined the parameters are by the data. How much can the parameters vary and still give a **good** model.
  - a and c – how much can we change w_0 and w_1. b – how the values should be changed together.

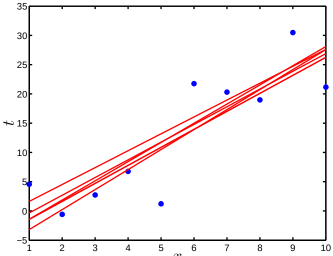


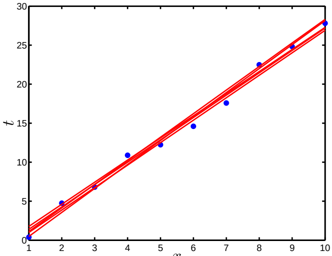
$\text{cov}\{\widehat{\mathbf{w}}\}$

- What does $\text{cov}\{\widehat{\mathbf{w}}\}$ tell us?
- Recall the linear model, $\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$

$$\text{cov}\{\widehat{\mathbf{w}}\} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

- Tells us how well defined the parameters are by the data. How much can the parameters vary and still give a **good** model.
  - a and c – how much can we change w_0 and w_1. b – how the values should be changed together.





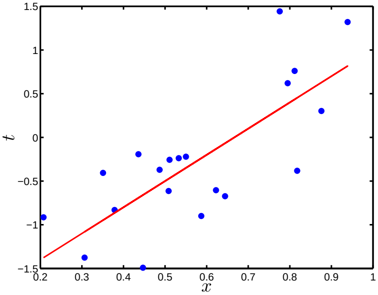
$\text{cov}\left\{\widehat{\mathbf{w}}\right\}$

$$\begin{aligned}\text{cov}\left\{\widehat{\mathbf{w}}\right\} &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\mathsf{T}}\right\} \\ &\quad - \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}^{\mathsf{T}}\end{aligned}$$

$\mathrm{cov}\{\widehat{\mathbf{w}}\}$

$$
\begin{aligned}
\mathrm{cov}\{\widehat{\mathbf{w}}\} &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\mathsf{T}}\right\} \\
&\quad -\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\}^{\mathsf{T}} \\
&= \mathbf{E}\left\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\mathsf{T}}\right\} - \mathbf{w}\mathbf{w}^{\mathsf{T}} \\
&= \vdots \\
\mathrm{cov}\{\widehat{\mathbf{w}}\} &= \sigma^2(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}
\end{aligned}
$$

Example

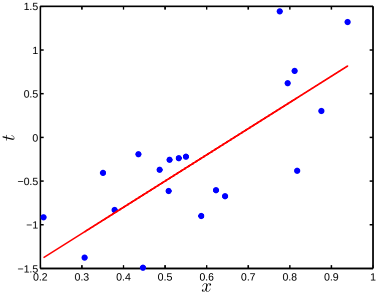


$$t_n = -2 + 3x_n + \epsilon_n$$

$$p(\epsilon_n) = \mathcal{N}(0, \sigma^2)$$

$$\sigma^2 = 0.5^2$$

Example



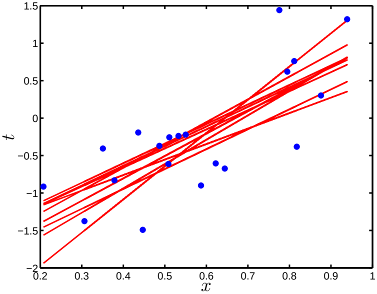
$$t_n = -2 + 3x_n + \epsilon_n$$

$$p(\epsilon_n) = \mathcal{N}(0, \sigma^2)$$

$$\sigma^2 = 0.5^2$$

$$\widehat{\mathbf{w}} = \begin{bmatrix} -1.95 \\ 2.94 \end{bmatrix}, \ \mathrm{cov}\{\widehat{\mathbf{w}}\} = \begin{bmatrix} 0.1195 & -0.1847 \\ -0.1847 & 0.3190 \end{bmatrix}$$

Example



$$t_n = -2 + 3x_n + \epsilon_n$$

$$p(\epsilon_n) = \mathcal{N}(0, \sigma^2)$$

$$\sigma^2 = 0.5^2$$

$$\widehat{\mathbf{w}} = \begin{bmatrix} -1.95 \\ 2.94 \end{bmatrix}, \quad \text{cov}\{\widehat{\mathbf{w}}\} = \begin{bmatrix} 0.1195 & -0.1847 \\ -0.1847 & 0.3190 \end{bmatrix}$$

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\}$

We saw that $\widehat{\mathbf{w}}$ was unbiased, what about $\widehat{\sigma^2}$?

$$\begin{aligned}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} &= \frac{1}{N}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{(\mathbf{t}-\mathbf{X}\widehat{\mathbf{w}})^{\mathsf{T}}(\mathbf{t}-\mathbf{X}\widehat{\mathbf{w}})\right\} \\ &= \sigma^2\left(1-\frac{D}{N}\right).\end{aligned}$$

Useful identity

$$\begin{aligned}p(\mathbf{t}) &= \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \\ \mathbf{E}_{p(\mathbf{t})}\left\{\mathbf{t}^{\mathsf{T}}\mathbf{A}\mathbf{t}\right\} &= \mathrm{Tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^{\mathsf{T}}\mathbf{A}\boldsymbol{\mu} \\ \mathrm{Tr}(\mathbf{A}) &= \sum_i A_{ii}\end{aligned}$$

Another useful identity

$$\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$$

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\}$

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} = \sigma^2\left(1 - \frac{D}{N}\right)$$

- In general $D < N$.
- So $1 - D/N < 1$.
- So $\widehat{\sigma^2} < \sigma^2$

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\}$

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} = \sigma^2\left(1 - \frac{D}{N}\right)$$

- In general $D < N$.
- So $1 - D/N < 1$.
- So $\widehat{\sigma^2} < \sigma^2$
- $\widehat{\sigma^2}$ is biased and will generally be too low.

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\}$

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} = \sigma^2\left(1 - \frac{D}{N}\right)$$

- In general $D < N$.
- So $1 - D/N < 1$.
- So $\widehat{\sigma^2} < \sigma^2$
- $\widehat{\sigma^2}$ is biased and will generally be too low.
- Why?
  - Because it is based on $\widehat{\mathbf{w}}$ which will, in general, be closer to the data than $\mathbf{w}$.

Confidence in parameter estimates
Story so far
Predictions
Prediction
Likelihood for model selection
Summary

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\}$

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} = \sigma^2\left(1 - \frac{D}{N}\right)$$

- In general $D < N$.
- So $1 - D/N < 1$.
- So $\widehat{\sigma^2} < \sigma^2$
- $\widehat{\sigma^2}$ is biased and will generally be too low.
- Why?
  - Because it is based on $\widehat{\mathbf{w}}$ which will, in general, be closer to the data than $\mathbf{w}$.
- As N increases, $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} \rightarrow \sigma^2$

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\}$

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} = \sigma^2\left(1 - \frac{D}{N}\right)$$

- In general $D < N$.
- So $1 - D/N < 1$.
- So $\widehat{\sigma^2} < \sigma^2$
- $\widehat{\sigma^2}$ is biased and will generally be too low.
- Why?
  - Because it is based on $\widehat{\mathbf{w}}$ which will, in general, be closer to the data than $\mathbf{w}$.
- As N increases, $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} \to \sigma^2$
- To think about – what if $D = N$ or $D > N$?

Example

Generate 100 datasets from the following model:

$$t_n = w_0 + w_1 x_n + \epsilon_n, \;\; p(\epsilon_n) = \mathcal{N}(0, 0.25)$$

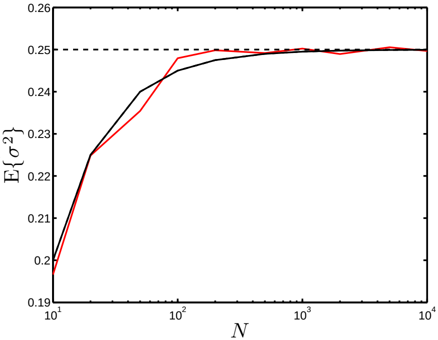
For $N = [10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10000]$

Example

Generate 100 datasets from the following model:

$$t_n = w_0 + w_1 x_n + \epsilon_n, \;\; p(\epsilon_n) = \mathcal{N}(0, 0.25)$$

For $N = [10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10000]$



Red curve – average $\widehat{\sigma^2}$ over 100 datasets. Black curve – theoretical value. Dashed line – true value.

Summary

- Recapped expectations.

Summary

- Recapped expectations.
- Computed $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} = \mathbf{w}$
  - $\widehat{\mathbf{w}}$ is **unbiased**.

Summary

- Recapped expectations.
- Computed $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} = \mathbf{w}$
  - $\widehat{\mathbf{w}}$ is **unbiased**.
- Computed $\text{cov}\{\widehat{\mathbf{w}}\} = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})$
  - Tells us how much slack there is in our parameters.

Summary

- Recapped expectations.
- Computed $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} = \mathbf{w}$
  - $\widehat{\mathbf{w}}$ is **unbiased**.
- Computed $\mathrm{cov}\{\widehat{\mathbf{w}}\} = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})$
  - Tells us how much slack there is in our parameters.
- Computed $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} = \sigma^2(1 - D/N)$
  - $\widehat{\sigma^2}$ is **biased**.
  - Gets better and better as we get more data.

Confidence in parameter estimates

Story so far

Predictions

Prediction

Likelihood for model selection

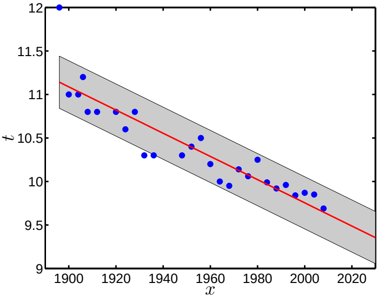
Summary

Predictions

- Our aim is to make predictions (e.g. London 2012)

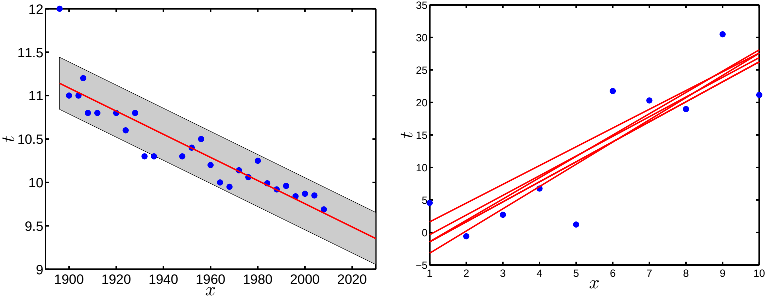
Predictions

- Our aim is to make predictions (e.g. London 2012)
- The noise in our data tells us that we can't predict exactly.



Predictions

- Our aim is to make predictions (e.g. London 2012)
- The noise in our data tells us that we can't predict exactly.
- The uncertainty in the parameters $\text{cov}\{\widehat{\mathbf{w}}\}$ should make them even less certain.



Predictions

- Our model is defined as:

$$t = \mathbf{w}^{\mathsf{T}}\mathbf{x} + \epsilon$$

- Given our estimate of the parameters, $\widehat{\mathbf{w}}$ and a new input, $\mathbf{x}_{\text{new}}$, if we had to predict a single value:

$$t_{\text{new}} = \widehat{\mathbf{w}}^{\mathsf{T}}\mathbf{x}_{\text{new}}$$

- Is this sensible?

Confidence in parameter estimates

Story so far

Predictions

Prediction

Likelihood for model selection

Summary

Predictions

- Our model is defined as:

$$t = \mathbf{w}^{\mathsf{T}}\mathbf{x} + \epsilon$$

- Given our estimate of the parameters, $\widehat{\mathbf{w}}$ and a new input, $\mathbf{x}_{\text{new}}$, if we had to predict a single value:

$$t_{\text{new}} = \widehat{\mathbf{w}}^{\mathsf{T}}\mathbf{x}_{\text{new}}$$

- Is this sensible? What is $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{t_{\text{new}}\right\}$?

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{t_{\text{new}}\right\} = \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}^{\mathsf{T}}\mathbf{x}_{\text{new}}\right\} = \mathbf{w}^{\mathsf{T}}\mathbf{x}_{\text{new}}$$

- which is a good thing!

Predictions

- What about $\text{var}\{t_{\text{new}}\}$?

$$\text{var}\{t_{\text{new}}\} = \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{t_{\text{new}}^2\right\} - \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{t_{\text{new}}\right\}^2$$

Confidence in parameter estimates

Story so far

Predictions

Prediction

Likelihood for model selection

Summary

Predictions

- What about $\mathrm{var}\{t_{\mathrm{new}}\}$?

$$
\begin{aligned}
\mathrm{var}\{t_{\mathrm{new}}\} &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{t_{\mathrm{new}}^2\right\} - \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{t_{\mathrm{new}}\right\}^2 \\
&= \mathbf{E}\left\{(\widehat{\mathbf{w}}^{\mathsf{T}}\mathbf{x}_{\mathrm{new}})^2\right\} - (\mathbf{w}^{\mathsf{T}}\mathbf{x}_{\mathrm{new}})^2 \\
&= \mathbf{x}_{\mathrm{new}}^{\mathsf{T}}\mathbf{E}\left\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\mathsf{T}}\right\}\mathbf{x}_{\mathrm{new}} - \mathbf{x}_{\mathrm{new}}^{\mathsf{T}}\mathbf{w}\mathbf{w}^{\mathsf{T}}\mathbf{x}_{\mathrm{new}} \\
&= \vdots \\
\mathrm{var}\{t_{\mathrm{new}}\} &= \sigma^2\mathbf{x}_{\mathrm{new}}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{x}_{\mathrm{new}}
\end{aligned}
$$

Prediction and variance

$$
\begin{aligned}
t_{\text{new}} &= \widehat{\mathbf{w}}^{\mathsf{T}} \mathbf{x}_{\text{new}} \\
\text{var}\{t_{\text{new}}\} &= \sigma^2 \mathbf{x}_{\text{new}}^{\mathsf{T}} (\mathbf{X}^{\mathsf{T}} \mathbf{X})^{-1} \mathbf{x}_{\text{new}}
\end{aligned}
$$

Prediction and variance

$$\begin{aligned} t_{\text{new}} &= \widehat{\mathbf{w}}^{\mathsf{T}} \mathbf{x}_{\text{new}} \\ \text{var}\{t_{\text{new}}\} &= \sigma^2 \mathbf{x}_{\text{new}}^{\mathsf{T}} (\mathbf{X}^{\mathsf{T}} \mathbf{X})^{-1} \mathbf{x}_{\text{new}} \end{aligned}$$

- Recall the expression for the covariance of the parameter estimate:

$$\text{cov}\{\widehat{\mathbf{w}}\} = \sigma^2 (\mathbf{X}^{\mathsf{T}} \mathbf{X})^{-1}$$

Prediction and variance

$$\begin{aligned} t_{\text{new}} &= \widehat{\mathbf{w}}^{\mathsf{T}} \mathbf{x}_{\text{new}} \\ \text{var}\{t_{\text{new}}\} &= \sigma^2 \mathbf{x}_{\text{new}}^{\mathsf{T}} (\mathbf{X}^{\mathsf{T}} \mathbf{X})^{-1} \mathbf{x}_{\text{new}} \end{aligned}$$

- Recall the expression for the covariance of the parameter estimate:

$$\text{cov}\{\widehat{\mathbf{w}}\} = \sigma^2 (\mathbf{X}^{\mathsf{T}} \mathbf{X})^{-1}$$

- Appears in the variance of the prediction:

$$\text{var}\{t_{\text{new}}\} = \mathbf{x}_{\text{new}}^{\mathsf{T}} \text{cov}\{\widehat{\mathbf{w}}\} \mathbf{x}_{\text{new}}$$

Confidence in parameter estimates
Story so far
Predictions
Prediction
Likelihood for model selection
Summary

Prediction and variance

Confidence in parameter estimates
Story so far
Predictions
Prediction
Likelihood for model selection
Summary

$$\begin{aligned} t_{\text{new}} &= \widehat{\mathbf{w}}^{\mathsf{T}}\mathbf{x}_{\text{new}} \\ \text{var}\{t_{\text{new}}\} &= \sigma^2 \mathbf{x}_{\text{new}}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{x}_{\text{new}} \end{aligned}$$

- Recall the expression for the covariance of the parameter estimate:

$$\text{cov}\{\widehat{\mathbf{w}}\} = \sigma^2(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$$

- Appears in the variance of the prediction:

$$\text{var}\{t_{\text{new}}\} = \mathbf{x}_{\text{new}}^{\mathsf{T}}\text{cov}\{\widehat{\mathbf{w}}\}\mathbf{x}_{\text{new}}$$

- If the variance in the parameters is high, so is the variance in the predictions.

Example

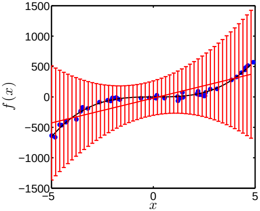
Data sampled from a 3rd order polynomial function:

$$t = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \epsilon$$

Example

Data sampled from a 3rd order polynomial function:

$$t = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \epsilon$$



Linear

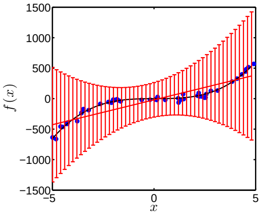
Plots show $t_{\text{new}} \pm \text{var}\{t_{\text{new}}\}$. (Black line is truth).

Confidence in parameter estimates
Story so far
Predictions
Prediction
Likelihood for model selection
Summary

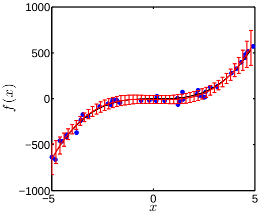
Example

Data sampled from a 3rd order polynomial function:

$$t = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \epsilon$$



Linear



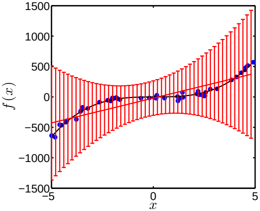
Cubic

Plots show $t_{\text{new}} \pm \text{var}\{t_{\text{new}}\}$. (Black line is truth).

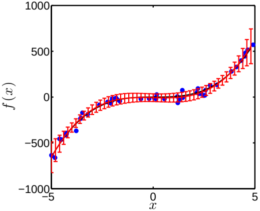
Example

Data sampled from a 3rd order polynomial function:

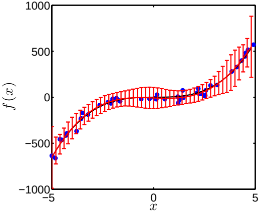
$$t = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \epsilon$$



Linear



Cubic



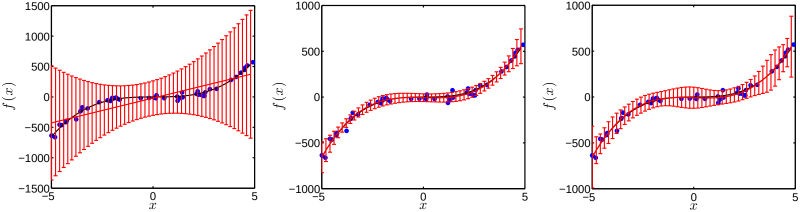
6th order

Plots show $t_{\text{new}} \pm \text{var}\{t_{\text{new}}\}$. (Black line is truth).

Example

Data sampled from a 3rd order polynomial function:

$$t = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \epsilon$$



Linear Cubic 6th order

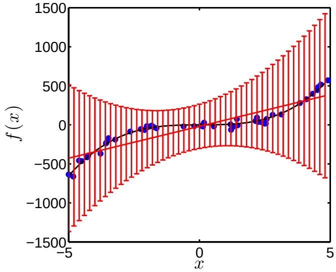
Plots show $t_{\text{new}} \pm \text{var}\{t_{\text{new}}\}$. (Black line is truth).

Why does the predictive variance increase above and below the correct order?

Not complex enough model – more 'noise'

In practice we don't know σ^2 so substitute $\widehat{\sigma^2}$:

$$\mathrm{var}\{t_{\mathrm{new}}\} = \widehat{\sigma^2}\mathbf{x}_{\mathrm{new}}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{x}_{\mathrm{new}}$$

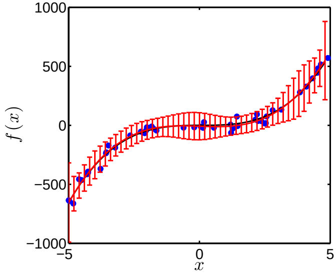


- The model is too simple.
- Some true variability can only be modelled noise.
- $\widehat{\sigma^2}$ is significantly over-estimated.
- Results in high $\mathrm{var}\{t_{\mathrm{new}}\}$.

Too complex model – parameters not well defined

Similarly, we substitute $\widehat{\sigma^2}$ into expression for $\text{cov}\{\widehat{\mathbf{w}}\}$:

$$\text{cov}\{\widehat{\mathbf{w}}\} = \widehat{\sigma^2}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$$



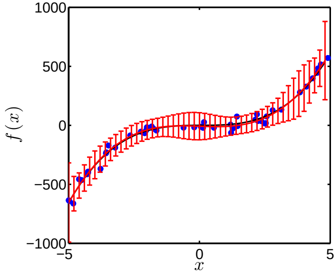
- 6th order model is too flexible.
- Many sets of parameters lead to a good model.
- Means that $\text{cov}\{\widehat{\mathbf{w}}\}$ is high.

Confidence in parameter estimates
Story so far
Predictions
Prediction
Likelihood for model selection
Summary

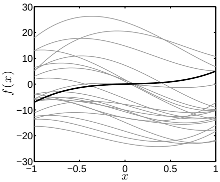
Too complex model – parameters not well defined

Similarly, we substitute $\widehat{\sigma^2}$ into expression for $\text{cov}\{\widehat{\mathbf{w}}\}$:

$$\text{cov}\{\widehat{\mathbf{w}}\} = \widehat{\sigma^2}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$$



- 6th order model is too flexible.
- Many sets of parameters lead to a good model.
- Means that $\text{cov}\{\widehat{\mathbf{w}}\}$ is high.

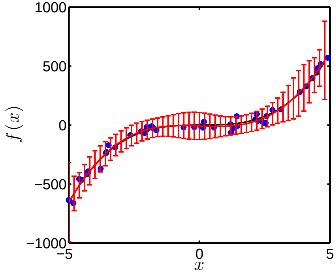


- ‘good’ 6th order models.

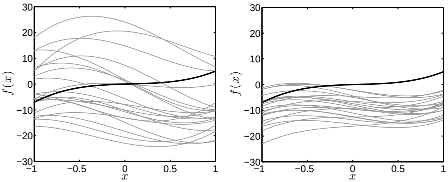
Too complex model – parameters not well defined

Similarly, we substitute $\widehat{\sigma^2}$ into expression for $\text{cov}\{\widehat{\mathbf{w}}\}$:

$$\text{cov}\{\widehat{\mathbf{w}}\} = \widehat{\sigma^2}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$$



- 6th order model is too flexible.
- Many sets of parameters lead to a good model.
- Means that $\text{cov}\{\widehat{\mathbf{w}}\}$ is high.



- 'good' 6th order models.
- 'good' 3rd order models.

Olympic prediction

Confidence in parameter estimates

Story so far

Predictions

Prediction

Likelihood for model selection

Summary

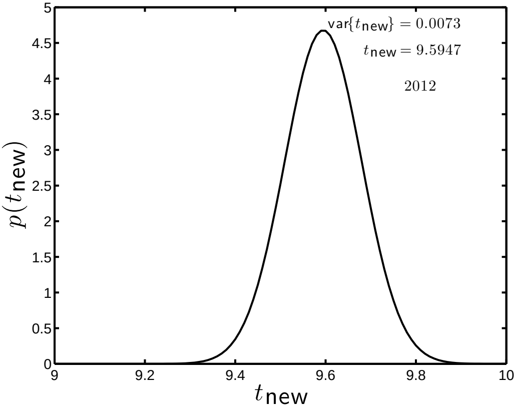
Linear model:

$$t = w_0 + w_1 x + \epsilon$$

Olympic prediction

Linear model:

$$t = w_0 + w_1 x + \epsilon$$



Confidence in parameter estimates

Story so far

Predictions

Prediction

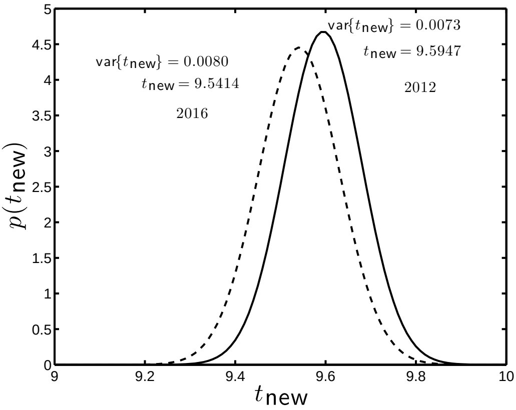
Likelihood for model selection

Summary

Olympic prediction

Linear model:

$$t = w_0 + w_1 x + \epsilon$$



Predictive variance increases as we get further from the training data.

Introduction

D. Dubhashi

Confidence in parameter estimates

Story so far

Predictions

Prediction

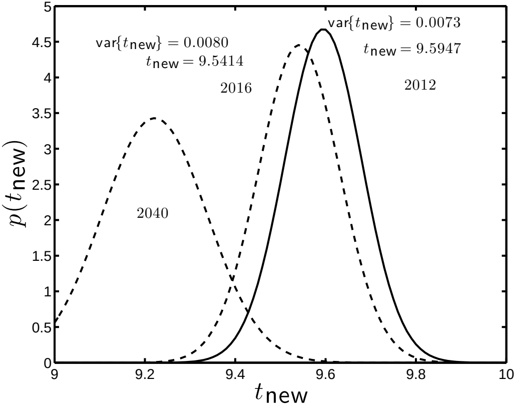
Likelihood for model selection

Summary

Olympic prediction

Linear model:

$$t = w_0 + w_1 x + \epsilon$$



Predictive variance increases as we get further from the training data.

Introduction
D. Dubhashi
Confidence in parameter estimates
Story so far
Predictions
Prediction
Likelihood for model selection
Summary

Summary

- Decided to model the noise.
- Introduced likelihood and maximised it to find $\widehat{\mathbf{w}}$ and $\widehat{\sigma^2}$.
- What did it buy us?

Confidence in parameter estimates

Story so far

Predictions

Prediction

Likelihood for model selection

Summary

Summary

- Decided to model the noise.
- Introduced likelihood and maximised it to find $\widehat{\mathbf{w}}$ and $\widehat{\sigma^2}$.
- What did it buy us?
- We can now:
  - Quantify the uncertainty in our parameters.
  - Quantify the uncertainty in our predictions.
  - This is very important in all applications....

Summary

- Decided to model the noise.
- Introduced likelihood and maximised it to find $\widehat{\mathbf{w}}$ and $\widehat{\sigma^2}$.
- What did it buy us?
- We can now:
  - Quantify the uncertainty in our parameters.
  - Quantify the uncertainty in our predictions.
  - This is very important in all applications....
- What next?
  - Going Bayesian.
  - Got to forget about single parameter values - parameters are random variables too.

Aside - from one model to many

- All of our efforts so far have been to find the 'best' model:
  - The one that minimises the loss.
  - The one that maximises the likelihood.
- Given the uncertainty, maybe we shouldn't trust one on its own?
- Consider the following RV:

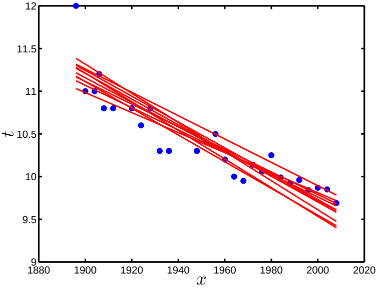
$$p(\mathbf{q}) = \mathcal{N}(\widehat{\mathbf{w}}, \text{cov}\{\widehat{\mathbf{w}}\})$$

- Samples of this RV $\mathbf{q}_s$ are **models** (assume $\widehat{\sigma^2}$ is fixed)
- We can generate lots of good models...

- Sample lots of $\mathbf{q}$ from:

$$p(\mathbf{q}) = \mathcal{N}(\widehat{\mathbf{w}}, \mathrm{cov}\{\widehat{\mathbf{w}}\})$$

- Each corresponds to a model.

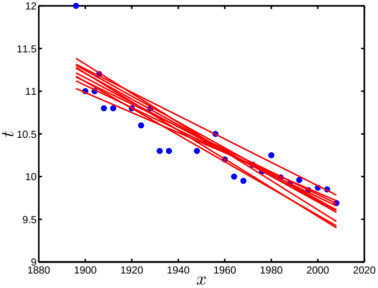


- Sample lots of $\mathbf{q}$ from:

$$p(\mathbf{q}) = \mathcal{N}(\widehat{\mathbf{w}}, \mathrm{cov}\{\widehat{\mathbf{w}}\})$$

- Each corresponds to a model.
- Compute a prediction from each one:

$$t_s = \mathbf{q}_s^{\mathsf{T}} \mathbf{x}_{\text{new}}$$



Confidence in parameter estimates

Story so far

Predictions

Prediction

Likelihood for model selection

Summary

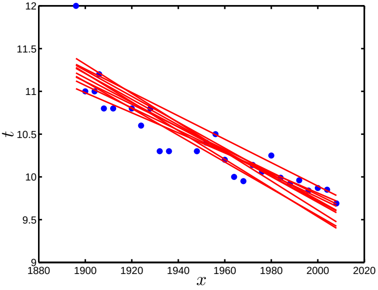
- Sample lots of $\mathbf{q}$ from:

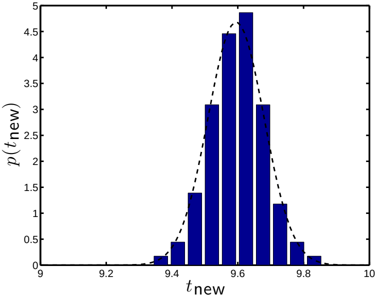
$$p(\mathbf{q}) = \mathcal{N}(\widehat{\mathbf{w}}, \text{cov}\{\widehat{\mathbf{w}}\})$$

- Each corresponds to a model.
- Compute a prediction from each one:

$$t_s = \mathbf{q}_s^{\mathsf{T}} \mathbf{x}_{\text{new}}$$

- Look at the distribution of predictions:





Do we need to take samples at all?

- Take an expectation...

$$\mathbf{E}_{p(\mathbf{q})}\left\{t_{\text{new}}\right\} = \int t_{\text{new}}\mathcal{N}(\widehat{\mathbf{w}}, \text{cov}\{\widehat{\mathbf{w}}\})\ dt_{\text{new}}$$

- We'll see more of this in the next lecture....