UNIVERSITY *of* GLASGOW

Machine Learning

Lecture. 5.

Mark Girolami

girolami@dcs.gla.ac.uk

Department of Computing Science
University of Glasgow

Probabilistic Regression

UNIVERSITY *of* GLASGOW

- Probabilistic view of Linear Regression

Probabilistic Regression

UNIVERSITY of GLASGOW

- Probabilistic view of Linear Regression
- Likelihood Principle.

Probabilistic Regression

UNIVERSITY
of
GLASGOW

- Probabilistic view of Linear Regression
- Likelihood Principle.
- Maximum Likelihood Parameter Estimation

Probabilistic Regression

UNIVERSITY
of
GLASGOW

- Probabilistic view of Linear Regression
- Likelihood Principle.
- Maximum Likelihood Parameter Estimation
- Uncertainty in Estimates & Prediction

Probabilistic Regression

UNIVERSITY of GLASGOW

- The data model which we have explored so far is of the form

$$t = f(x; \mathbf{w}) + \epsilon$$

Probabilistic Regression

UNIVERSITY *of* GLASGOW

- The data model which we have explored so far is of the form
$$t = f(x; \mathbf{w}) + \epsilon$$
- Model based on a deterministic function of inputs, $f(x; \mathbf{w})$

Probabilistic Regression

UNIVERSITY *of* GLASGOW

- The data model which we have explored so far is of the form

$$t = f(x; \mathbf{w}) + \epsilon$$

- Model based on a deterministic function of inputs, $f(x; \mathbf{w})$
- Contaminated by noise or some error defined by ϵ

Noise Distribution

UNIVERSITY *of* GLASGOW

- Noise term can be assumed to be Normally distributed with mean zero and some variance σ i.e. $\epsilon \sim \mathcal{N}(0, \sigma)$

Noise Distribution

UNIVERSITY
of
GLASGOW

- Noise term can be assumed to be Normally distributed with mean zero and some variance σ i.e. $\epsilon \sim \mathcal{N}(0, \sigma)$
- So noise *sits on top* of, and corrupts, model output $f(x; \mathbf{w})$ to give t

Noise Distribution

UNIVERSITY *of* GLASGOW

- Noise term can be assumed to be Normally distributed with mean zero and some variance σ i.e. $\epsilon \sim \mathcal{N}(0, \sigma)$
- So noise *sits on top* of, and corrupts, model output $f(x; \mathbf{w})$ to give t
- This can be written as

$$t|x \sim \mathcal{N}(f(x; \mathbf{w}), \sigma)$$

Noise Distribution

UNIVERSITY of GLASGOW

- Noise term can be assumed to be Normally distributed with mean zero and some variance σ i.e. $\epsilon \sim \mathcal{N}(0, \sigma)$
- So noise *sits on top* of, and corrupts, model output $f(x; \mathbf{w})$ to give t
- This can be written as

$$t|x \sim \mathcal{N}(f(x; \mathbf{w}), \sigma)$$

- Likewise we can write

$$p(t|x) = \mathcal{N}(f(x; \mathbf{w}), \sigma)$$

which reads as the conditional probability distribution of t given x is Gaussian distribution with mean $f(x; \mathbf{w})$ and variance σ

Probabilistic Regression

- The question that we ask is *How likely is it that I would have observed the outputs given the inputs and model parameters*

Probabilistic Regression

UNIVERSITY *of* GLASGOW

- The question that we ask is *How likely is it that I would have observed the outputs given the inputs and model parameters*
- The likelihood of observing the data point, t, is the conditional probability of making that observation i.e. $p(t|x, \mathbf{w})$

Probabilistic Regression

UNIVERSITY of GLASGOW

- The question that we ask is *How likely is it that I would have observed the outputs given the inputs and model parameters*
- The likelihood of observing the data point, t, is the conditional probability of making that observation i.e. $p(t|x, \mathbf{w})$
- For N observations $(x_1, t_1), \cdots, (x_N, t_N) = (\mathbf{x}, \mathbf{t})$

Probabilistic Regression

UNIVERSITY of GLASGOW

- The question that we ask is *How likely is it that I would have observed the outputs given the inputs and model parameters*
- The likelihood of observing the data point, t, is the conditional probability of making that observation i.e. $p(t|x, \mathbf{w})$
- For N observations $(x_1, t_1), \cdots, (x_N, t_N) = (\mathbf{x}, \mathbf{t})$
- Want the joint probability of all the outputs conditioned on all the input values and model parameters i.e. $p(t_1, t_2, \cdots, t_N|x_1, x_2, \cdots, x_N, \mathbf{w}) = p(\mathbf{t}|\mathbf{x}, \mathbf{w})$

Probabilistic Regression

UNIVERSITY
of
GLASGOW

- The question that we ask is *How likely is it that I would have observed the outputs given the inputs and model parameters*
- The likelihood of observing the data point, t, is the conditional probability of making that observation i.e. $p(t|x, \mathbf{w})$
- For N observations $(x_1, t_1), \cdots, (x_N, t_N) = (\mathbf{x}, \mathbf{t})$
- Want the joint probability of all the outputs conditioned on all the input values and model parameters i.e. $p(t_1, t_2, \cdots, t_N | x_1, x_2, \cdots, x_N, \mathbf{w}) = p(\mathbf{t}|\mathbf{x}, \mathbf{w})$
- This joint probability is the data likelihood

Probabilistic Regression

UNIVERSITY of GLASGOW

- Assume observations made *independently* of each other. Measurement just made does not affect the following measurement to be made. Essentially assuming *statistical independence* between measurements.

Probabilistic Regression

UNIVERSITY of GLASGOW

- Assume observations made *independently* of each other. Measurement just made does not affect the following measurement to be made. Essentially assuming *statistical independence* between measurements.
- Assume noise corrupting measurements always comes from the same distribution so outputs will be *identically distributed*

Probabilistic Regression

UNIVERSITY *of* GLASGOW

- Assume observations made *independently* of each other. Measurement just made does not affect the following measurement to be made. Essentially assuming *statistical independence* between measurements.
- Assume noise corrupting measurements always comes from the same distribution so outputs will be *identically distributed*
- Assumptions can be stated as *we assume that the data is Independent and Identically Distributed* often denoted as IID

Probabilistic Regression

- With IID assumption joint probability of measurements takes factored form i.e.

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \sigma) = \prod_{n=1}^{N} p(t_n|x_n, \mathbf{w}, \sigma) = \prod_{n=1}^{N} \mathcal{N}(f(x_n; \mathbf{w}), \sigma)$$

Probabilistic Regression

- With IID assumption joint probability of measurements takes factored form i.e.

$$p(\mathbf{t}|\mathbf{x},\mathbf{w},\sigma) = \prod_{n=1}^{N} p(t_n|x_n,\mathbf{w},\sigma) = \prod_{n=1}^{N} \mathcal{N}(f(x_n;\mathbf{w}),\sigma)$$

- This is our likelihood function

Probabilistic Regression

UNIVERSITY *of* GLASGOW

- With IID assumption joint probability of measurements takes factored form i.e.

$$p(\mathbf{t}|\mathbf{x},\mathbf{w},\sigma) = \prod_{n=1}^{N} p(t_n|x_n,\mathbf{w},\sigma) = \prod_{n=1}^{N} \mathcal{N}(f(x_n;\mathbf{w}),\sigma)$$

- This is our likelihood function
- We see that the likelihood function depends on the parameters of our model

Probabilistic Regression

UNIVERSITY of GLASGOW

- With IID assumption joint probability of measurements takes factored form i.e.

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \sigma) = \prod_{n=1}^{N} p(t_n|x_n, \mathbf{w}, \sigma) = \prod_{n=1}^{N} \mathcal{N}(f(x_n; \mathbf{w}), \sigma)$$

- This is our likelihood function
- We see that the likelihood function depends on the parameters of our model
- The parameters can then be tuned to make the data more likely under the model

Maximum Likelihood

UNIVERSITY
of
GLASGOW

- Select model parameters $\mathbf{w}$ & σ which will make our observations most likely

Maximum Likelihood

UNIVERSITY *of* GLASGOW

- Select model parameters $\mathbf{w}$ & σ which will make our observations most likely
- Need to find maximum of likelihood function with respect to model parameters

Maximum Likelihood

UNIVERSITY of GLASGOW

- Select model parameters $\mathbf{w}$ & σ which will make our observations most likely
- Need to find maximum of likelihood function with respect to model parameters
- Maximise the logarithm of the likelihood function as the log-likelihood is often more convenient to work with analytically

Maximum Likelihood

UNIVERSITY *of* GLASGOW

- Select model parameters $\mathbf{w}$ & σ which will make our observations most likely
- Need to find maximum of likelihood function with respect to model parameters
- Maximise the logarithm of the likelihood function as the log-likelihood is often more convenient to work with analytically
- Need to take derivatives of the log-likelihood function

Maximum Likelihood

UNIVERSITY *of* GLASGOW

Log Likelihood $\mathcal{L} = \log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \sigma)$ can be written as

$$\begin{aligned}
&= \sum_{n=1}^{N} \log p(t_n|x_n, \mathbf{w}, \sigma) \\
&= \sum_{n=1}^{N} \log \mathcal{N}(f(x_n; \mathbf{w}), \sigma) \\
&= \sum_{n=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}|t_n - f(x_n; \mathbf{w})|^2\right) \\
&= -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}|t_n - f(x_n; \mathbf{w})|^2
\end{aligned}$$

Maximum Likelihood

UNIVERSITY *of* GLASGOW

- Stationary points with respect to $\mathbf{w}$ follows as

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{1}{\sigma^2}(\mathbf{X}^\mathsf{T}\mathbf{t} - \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}) = 0$$

Maximum Likelihood

UNIVERSITY *of* GLASGOW

- Stationary points with respect to $\mathbf{w}$ follows as

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{1}{\sigma^2}(\mathbf{X}^\mathsf{T}\mathbf{t} - \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}) = 0$$

- Matrix of second-order partial derivatives

$$\frac{\partial^2 \mathcal{L}}{\partial \mathbf{w} \partial \mathbf{w}^\mathsf{T}} = -\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}$$

which is strictly negative and so we have indeed obtained the maximum of the likelihood.

Maximum Likelihood

UNIVERSITY *of* GLASGOW

- Stationary points with respect to $\mathbf{w}$ follows as

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{1}{\sigma^2}(\mathbf{X}^\mathsf{T}\mathbf{t} - \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}) = 0$$

- Matrix of second-order partial derivatives

$$\frac{\partial^2 \mathcal{L}}{\partial \mathbf{w} \partial \mathbf{w}^\mathsf{T}} = -\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}$$

  which is strictly negative and so we have indeed obtained the maximum of the likelihood.

- *Maximum-likelihood* solution is $\widehat{\mathbf{w}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$.

Maximum Likelihood

UNIVERSITY *of* GLASGOW

- Stationary points with respect to $\mathbf{w}$ follows as

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{1}{\sigma^2}(\mathbf{X}^\mathsf{T}\mathbf{t} - \mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}) = 0$$

- Matrix of second-order partial derivatives

$$\frac{\partial^2 \mathcal{L}}{\partial \mathbf{w}\partial \mathbf{w}^\mathsf{T}} = -\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}$$

  which is strictly negative and so we have indeed obtained the maximum of the likelihood.

- *Maximum-likelihood* solution is $\widehat{\mathbf{w}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$.
- Look familiar?

Estimate Uncertainty

UNIVERSITY *of* GLASGOW

- Stationary points with respect to σ left as tutorial exercise.

Estimate Uncertainty

UNIVERSITY
of
GLASGOW

- Stationary points with respect to σ left as tutorial exercise.
- What can we say about how certain we are in our ML estimates?.

Estimate Uncertainty

UNIVERSITY *of* GLASGOW

- Stationary points with respect to σ left as tutorial exercise.
- What can we say about how certain we are in our ML estimates?.
- If $\widehat{\mathbf{w}}$ is our estimate then what variance is there around this estimate?.

Estimate Uncertainty

UNIVERSITY *of* GLASGOW

- Stationary points with respect to σ left as tutorial exercise.
- What can we say about how certain we are in our ML estimates?.
- If $\widehat{\mathbf{w}}$ is our estimate then what variance is there around this estimate?.
- The smaller the variance the more certain we are of our estimate - need expression for estimate variance.

Estimate Uncertainty

UNIVERSITY *of* GLASGOW

- ML estimate $\widehat{\mathbf{w}}$ is a vector so can we obtain covariance?

Estimate Uncertainty

UNIVERSITY
of
GLASGOW

- ML estimate $\widehat{\mathbf{w}}$ is a vector so can we obtain covariance?
- Remember that covariance of vector defined as

$$E\{(\widehat{\mathbf{w}}-E\{\widehat{\mathbf{w}}\})(\widehat{\mathbf{w}}-E\{\widehat{\mathbf{w}}\})^{\mathsf{T}}\} = E\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\mathsf{T}}\}-E\{\widehat{\mathbf{w}}\}E\{\widehat{\mathbf{w}}^{\mathsf{T}}\}$$

Estimate Uncertainty

UNIVERSITY *of* GLASGOW

- ML estimate $\widehat{\mathbf{w}}$ is a vector so can we obtain covariance?
- Remember that covariance of vector defined as

$$E\{(\widehat{\mathbf{w}}-E\{\widehat{\mathbf{w}}\})(\widehat{\mathbf{w}}-E\{\widehat{\mathbf{w}}\})^{\mathsf{T}}\} = E\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\mathsf{T}}\}-E\{\widehat{\mathbf{w}}\}E\{\widehat{\mathbf{w}}^{\mathsf{T}}\}$$

- Now ML and LS estimators unbiased so $E\{\widehat{\mathbf{w}}\} = \mathbf{w}$ true model parameters

Estimate Uncertainty

UNIVERSITY of GLASGOW

- ML estimate $\widehat{\mathbf{w}}$ is a vector so can we obtain covariance?
- Remember that covariance of vector defined as

$$E\{(\widehat{\mathbf{w}}-E\{\widehat{\mathbf{w}}\})(\widehat{\mathbf{w}}-E\{\widehat{\mathbf{w}}\})^{\mathsf{T}}\} = E\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\mathsf{T}}\}-E\{\widehat{\mathbf{w}}\}E\{\widehat{\mathbf{w}}^{\mathsf{T}}\}$$

- Now ML and LS estimators unbiased so $E\{\widehat{\mathbf{w}}\} = \mathbf{w}$ true model parameters
- So require expression for $E\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\mathsf{T}}\}$

Estimate Uncertainty

UNIVERSITY of GLASGOW

- As $\widehat{\mathbf{w}} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t}$ then the outer product of the two vectors is $\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\mathsf{T}} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t}\mathbf{t}^{\mathsf{T}}\mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}$

Estimate Uncertainty

UNIVERSITY *of* GLASGOW

- As $\widehat{\mathbf{w}} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t}$ then the outer product of the two vectors is $\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\mathsf{T}} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t}\mathbf{t}^{\mathsf{T}}\mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}$
- Take the required expectation and $E\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\mathsf{T}}\} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}E\{\mathbf{t}\mathbf{t}^{\mathsf{T}}\}\mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}$

Estimate Uncertainty

UNIVERSITY *of* GLASGOW

- As $\widehat{\mathbf{w}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$ then the outer product of the two vectors is $\widehat{\mathbf{w}}\widehat{\mathbf{w}}^\mathsf{T} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}\mathbf{t}^\mathsf{T}\mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}$
- Take the required expectation and $E\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^\mathsf{T}\} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}E\{\mathbf{t}\mathbf{t}^\mathsf{T}\}\mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}$
- Now require expression for $E\{\mathbf{t}\mathbf{t}^\mathsf{T}\}$.

Estimate Uncertainty

UNIVERSITY *of* GLASGOW

- As $\mathbf{t} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$ then

$$
\begin{aligned}
E\{\mathbf{t}\mathbf{t}^{\mathsf{T}}\} &= E\{(\mathbf{X}\mathbf{w} + \boldsymbol{\epsilon})(\mathbf{X}\mathbf{w} + \boldsymbol{\epsilon})^{\mathsf{T}}\} \\
&= E\{\mathbf{X}\mathbf{w}\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}} + 2\boldsymbol{\epsilon}\mathbf{w}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathsf{T}}\} \\
&= \mathbf{X}\mathbf{w}\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}} + 2E\{\boldsymbol{\epsilon}\}\mathbf{w}^{\mathsf{T}}\mathbf{X} + E\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathsf{T}}\} \\
&= \mathbf{X}\mathbf{w}\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}} + \sigma^2\mathbf{I}
\end{aligned}
$$

Estimate Uncertainty

UNIVERSITY *of* GLASGOW

- As $\mathbf{t} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$ then

$$
\begin{aligned}
E\{\mathbf{t}\mathbf{t}^{\mathsf{T}}\} &= E\{(\mathbf{X}\mathbf{w} + \boldsymbol{\epsilon})(\mathbf{X}\mathbf{w} + \boldsymbol{\epsilon})^{\mathsf{T}}\} \\
&= E\{\mathbf{X}\mathbf{w}\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}} + 2\boldsymbol{\epsilon}\mathbf{w}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathsf{T}}\} \\
&= \mathbf{X}\mathbf{w}\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}} + 2E\{\boldsymbol{\epsilon}\}\mathbf{w}^{\mathsf{T}}\mathbf{X} + E\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathsf{T}}\} \\
&= \mathbf{X}\mathbf{w}\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}} + \sigma^2\mathbf{I}
\end{aligned}
$$

- So

$$
\begin{aligned}
E\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\mathsf{T}}\} &= \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}E\{\mathbf{t}\mathbf{t}^{\mathsf{T}}\}\mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1} \\
&= \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}(\mathbf{X}\mathbf{w}\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}} + \sigma^2\mathbf{I})\mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1} \\
&= \mathbf{w}\mathbf{w}^{\mathsf{T}} + \sigma^2\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}
\end{aligned}
$$

Estimate Uncertainty

UNIVERSITY *of* GLASGOW

- Finally the covariance matrix for our estimates is given as

$$
\begin{aligned}
E\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\mathsf{T}}\} - E\{\widehat{\mathbf{w}}\}E\{\widehat{\mathbf{w}}^{\mathsf{T}}\} &= \mathbf{w}\mathbf{w}^{\mathsf{T}} + \sigma^2 \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1} - \mathbf{w}\mathbf{w}^{\mathsf{T}} \\
&= \sigma^2 \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}
\end{aligned}
$$

Estimate Uncertainty

- Finally the covariance matrix for our estimates is given as

$$\begin{aligned} E\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^\mathsf{T}\} - E\{\widehat{\mathbf{w}}\}E\{\widehat{\mathbf{w}}^\mathsf{T}\} &= \mathbf{w}\mathbf{w}^\mathsf{T} + \sigma^2 \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1} - \mathbf{w}\mathbf{w}^\mathsf{T} \\ &= \sigma^2 \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1} \end{aligned}$$

- Very important result as now we can assess the variance associated with our ML estimates

Estimate Uncertainty

UNIVERSITY *of* GLASGOW

- Finally the covariance matrix for our estimates is given as

$$
\begin{aligned}
E\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\mathsf{T}}\} - E\{\widehat{\mathbf{w}}\}E\{\widehat{\mathbf{w}}^{\mathsf{T}}\} &= \mathbf{w}\mathbf{w}^{\mathsf{T}} + \sigma^2 \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1} - \mathbf{w}\mathbf{w}^{\mathsf{T}} \\
&= \sigma^2 \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}
\end{aligned}
$$

- Very important result as now we can assess the variance associated with our ML estimates
- Expression for matrix of partial derivatives gives

$$
E\{\widehat{\mathbf{w}}\widehat{\mathbf{w}}^{\mathsf{T}}\} - E\{\widehat{\mathbf{w}}\}E\{\widehat{\mathbf{w}}^{\mathsf{T}}\} = -\left(\frac{\partial^2 \mathcal{L}}{\partial \mathbf{w} \partial \mathbf{w}^{\mathsf{T}}}\right)^{-1}
$$

Small curvature of likelihood $\Rightarrow$ high variance in estimate $\Rightarrow$ parameter possibly irrelevant

Estimate Uncertainty

UNIVERSITY *of* GLASGOW

- To make a *new* prediction then our maximum-likelihood estimate and the associated variance around this estimate gives $\widehat{t}_{new} \pm \sigma^2_{new}$

Estimate Uncertainty

UNIVERSITY *of* GLASGOW

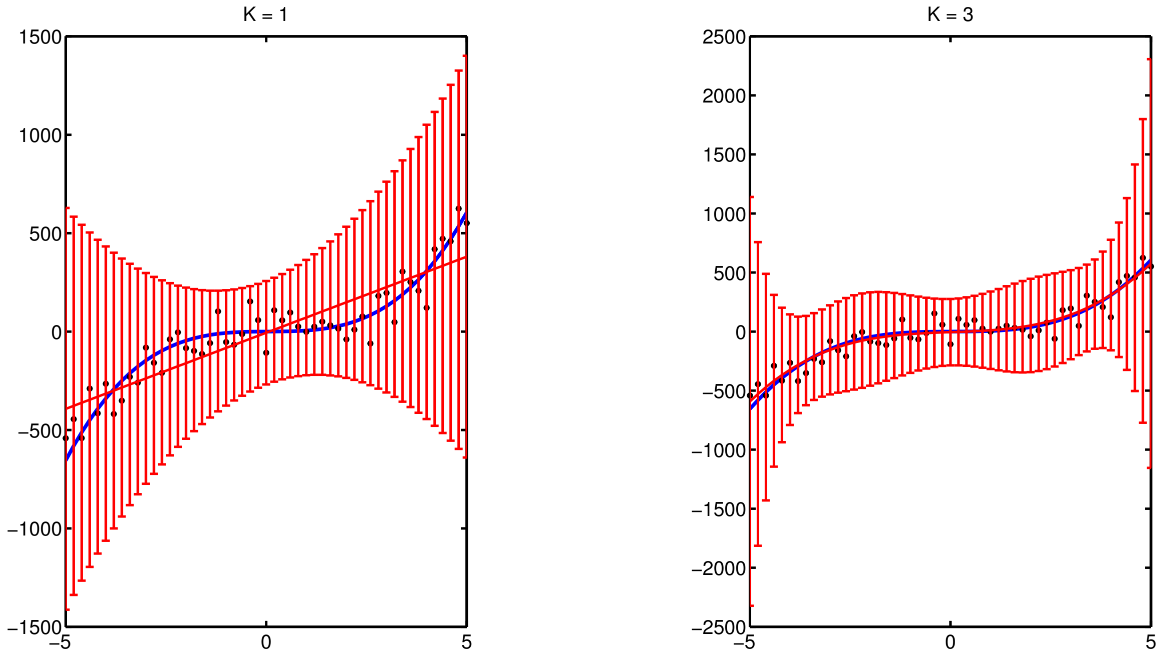
- To make a *new* prediction then our maximum-likelihood estimate and the associated variance around this estimate gives $\widehat{t}_{new} \pm \sigma^2_{new}$
- Where

$$
\begin{aligned}
\widehat{t}_{new} &= \mathbf{x}_{new}^{\mathsf{T}} \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1} \mathbf{X}^{\mathsf{T}}\mathbf{t} \\
\sigma^2_{new} &= \widehat{\sigma}^2 \mathbf{x}_{new}^{\mathsf{T}} \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1} \mathbf{x}_{new}
\end{aligned}
$$

with $\widehat{\sigma}^2 = \frac{1}{N}\left(\mathbf{t}^{\mathsf{T}}\mathbf{t} - \mathbf{t}^{\mathsf{T}}\hat{\mathbf{t}}\right)$

Estimate Uncertainty

UNIVERSITY *of* GLASGOW



(a) (b)

Figure 1: The blue solid line indicates the true noise free functions and the black dots are the actual observed noisy realisations of the data. The solid red line indicates the estimated function with the error-bars indicating the variance (uncertainty) in the estimated functional response at each of the data points ie $\widehat{t}_n \pm \sigma_n^2$.

Likelihood

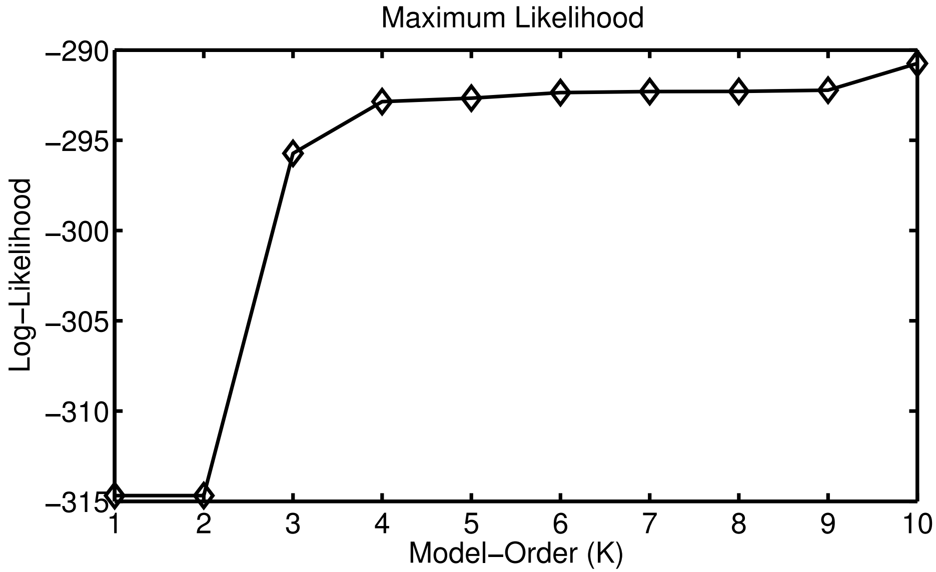


Figure 2: The Maximum Likelihood score for polynomial models from $K = 1$ to $K = 10$. Perhaps unsurprisingly the likelihood score monotonically increases with K.