Machine Learning

Lecture. 3.

Mark Girolami

girolami@dcs.gla.ac.uk

Department of Computing Science
University of Glasgow

Generalisation

UNIVERSITY
of
GLASGOW

- The important observations made in Laboratory One.

Generalisation

UNIVERSITY
of
GLASGOW

- The important observations made in Laboratory One.
- Increasing model complexity (polynomial order) yields monotonic decrease in MSE on *training* data.

Generalisation

UNIVERSITY
of
GLASGOW

- The important observations made in Laboratory One.
- Increasing model complexity (polynomial order) yields monotonic decrease in MSE on *training* data.
- Increasing model complexity does not necessarily yield monotonic decrease in *testing error*

Generalisation

UNIVERSITY
of
GLASGOW

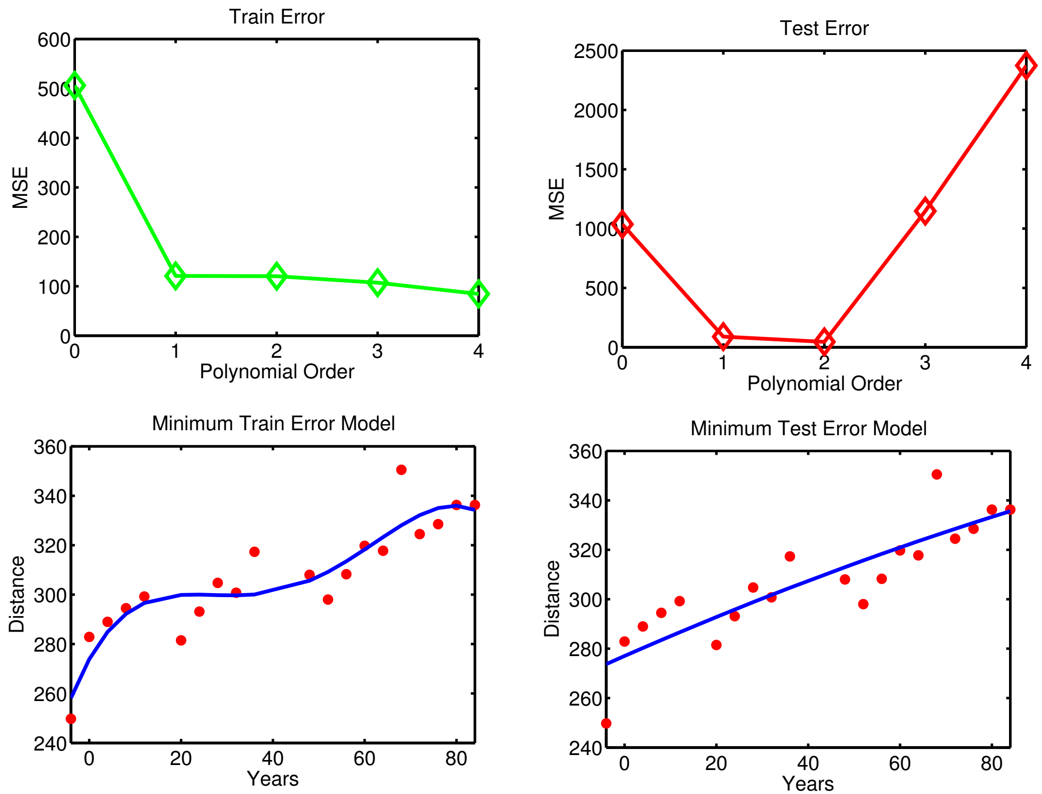

Figure 1: Results from Laboratory 1, designing polynomial order regression model to predict long jump distance in last five Olympic Games (1988 - 2004) given results from all previous games.

Generalisation

UNIVERSITY *of* GLASGOW

- Employing too simple a model then poor predictions will be made but if too complex a model employed the quality of predictions also adversely affected.

Generalisation

UNIVERSITY of GLASGOW

- Employing too simple a model then poor predictions will be made but if too complex a model employed the quality of predictions also adversely affected.
- This week looking at underlying mechanisms which cause this phenomenon and we will be introduced to methods which allow us to estimate what our model predictive performance or test error will be.

Generalisation

UNIVERSITY
of
GLASGOW

- Employing too simple a model then poor predictions will be made but if too complex a model employed the quality of predictions also adversely affected.
- This week looking at underlying mechanisms which cause this phenomenon and we will be introduced to methods which allow us to estimate what our model predictive performance or test error will be.
- What is important is developing a model that can *generalise* its performance beyond the available examples used for *training*.

Generalisation

UNIVERSITY
of
GLASGOW

- Consider again our averaged Loss-Function defined as

$$\frac{1}{N}\sum_{n=1}^{N}\mathcal{L}(t_n, f(x_n;\mathbf{w}))$$

Generalisation

- Consider again our averaged Loss-Function defined as

$$\frac{1}{N}\sum_{n=1}^{N}\mathcal{L}(t_n, f(x_n;\mathbf{w}))$$

- Each *input-output* pair (x_n, t_n) can be assumed to follow a natural distribution which makes it more likely to observe certain *input-output* pairs than others.

Generalisation

UNIVERSITY *of* GLASGOW

- Consider again our averaged Loss-Function defined as

$$\frac{1}{N}\sum_{n=1}^{N}\mathcal{L}(t_n, f(x_n;\mathbf{w}))$$

- Each *input-output* pair (x_n, t_n) can be assumed to follow a natural distribution which makes it more likely to observe certain *input-output* pairs than others.
- We can say that there is a *Probability Distribution* $p(x,t)$ which characterizes how likely it is to observe any particular pair (x,t)

Generalisation

UNIVERSITY
of
GLASGOW

- Ideally what we would like to be able to do would be to minimise the loss over all the possible *input-output* pairs that could possibly be observed.

Generalisation

UNIVERSITY *of* GLASGOW

- Ideally what we would like to be able to do would be to minimise the loss over all the possible *input-output* pairs that could possibly be observed.
- In other words we want to minimise the **Expected Loss**.

Generalisation

UNIVERSITY *of* GLASGOW

- Ideally what we would like to be able to do would be to minimise the loss over all the possible *input-output* pairs that could possibly be observed.
- In other words we want to minimise the **Expected Loss**.
- The Expectation operator is defined as the population average of a function which for a continuous (real) random variable X which takes on values $x \in \mathbb{R}$ with probability density $p(x)$ is defined as $E\{f(X)\} = \int f(x)p(x)dx$. For example the expected value or population average of X is $E\{X\} = \int xp(x)dx$. If X takes on a number of K discrete values $(X = x_k)$ then $E\{X\} = \sum_{k=1}^{K} x_k P(x_k)$

Generalisation

UNIVERSITY
of
GLASGOW

- **Expected Loss** then defined as

$$E\{\mathcal{L}\} = \int\int \mathcal{L}(t, f(x; \mathbf{w}))p(x, t)dxdt$$

Generalisation

- **Expected Loss** then defined as

$$E\{\mathcal{L}\} = \int\int \mathcal{L}(t, f(x; \mathbf{w}))p(x, t)dxdt$$

- As we have N examples drawn from $p(x, t)$ we can estimate the expected loss with the sample average

Generalisation

UNIVERSITY *of* GLASGOW

- **Expected Loss** then defined as

$$E\{\mathcal{L}\} = \int\int \mathcal{L}(t, f(x; \mathbf{w}))p(x, t)dxdt$$

- As we have N examples drawn from $p(x, t)$ we can estimate the expected loss with the sample average

$$\frac{1}{N}\sum_{n=1}^{N} \mathcal{L}(t_n, f(x_n; \mathbf{w}))$$

Bias-Variance Decomposition

UNIVERSITY
of
GLASGOW

- The expected squared error loss can be rewritten so that we can gain insight regarding the source of our modeling errors

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- The expected squared error loss can be rewritten so that we can gain insight regarding the source of our modeling errors
- We assume that the *true* model for our data is linear i.e. $w_0 + w_1 x$. Let us also assume that we had an infinite amount of data i.e. $N \rightarrow \infty$ then the MSE, which is based on a sample of data drawn from $p(x, t)$, will tend to the expected loss.

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- The expected squared error loss can be rewritten so that we can gain insight regarding the source of our modeling errors
- We assume that the *true* model for our data is linear i.e. $w_0 + w_1 x$. Let us also assume that we had an infinite amount of data i.e. $N \to \infty$ then the MSE, which is based on a sample of data drawn from $p(x, t)$, will tend to the expected loss.
- We denote $[1 \;\; x]^{\mathsf{T}}$ as $\mathbf{x}$ in what follows.

Bias-Variance Decomposition

UNIVERSITY
of
GLASGOW

- For MSE loss

Bias-Variance Decomposition

UNIVERSITY of GLASGOW

- For MSE loss

$$
\begin{aligned}
& \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} |t_n - f(x_n; \mathbf{w})|^2 \\
= & \int\int |t - f(x; \mathbf{w})|^2 p(x,t) dx dt \\
= & \int\int |t - \mathbf{w}^\mathsf{T}\mathbf{x}|^2 p(t|x) p(x) dx dt
\end{aligned}
$$

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- Now if we differentiate the expected loss with respect to the parameters $\mathbf{w} = [w_0 \ \ w_1]^\mathsf{T}$ and solve for $\mathbf{w}$ then we obtain

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- Now if we differentiate the expected loss with respect to the parameters $\mathbf{w} = [w_0 \;\; w_1]^{\mathsf{T}}$ and solve for $\mathbf{w}$ then we obtain

$$2 \int \int (t\mathbf{x} - \mathbf{x}\mathbf{x}^{\mathsf{T}}\mathbf{w}) p(t|x) p(x) dx dt = 0$$

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- Now if we differentiate the expected loss with respect to the parameters $\mathbf{w} = [w_0 \;\; w_1]^\mathsf{T}$ and solve for $\mathbf{w}$ then we obtain

$$2\int\int (t\mathbf{x} - \mathbf{x}\mathbf{x}^\mathsf{T}\mathbf{w})p(t|x)p(x)dxdt = 0$$

- Now $\int\int t\mathbf{x}p(t|x)p(x)dxdt$ is expected value of the cross term $t\mathbf{x}$ under $p(x,t)$. Gives description of how *inputs* x and *outputs* t are *correlated*. It is a measure of their *cross-covariance* denoted by $E\{TX\}$, where the upper case is used to denote that these are random variables as opposed to the values which they may take on i.e. t & x.

Bias-Variance Decomposition

UNIVERSITY
of
GLASGOW

- The right hand term is defined as

Bias-Variance Decomposition

UNIVERSITY of GLASGOW

- The right hand term is defined as

$$
\begin{aligned}
\int\int \mathbf{x}\mathbf{x}^{\mathsf{T}}\mathbf{w}p(t|x)p(x)dxdt &= \int p(t|x)dt \int \mathbf{x}\mathbf{x}^{\mathsf{T}}\mathbf{w}p(x)dx \\
&= 1 \times \int \mathbf{x}\mathbf{x}^{\mathsf{T}}p(x)dx\ \mathbf{w} \\
&= \int \begin{bmatrix} 1 & x \\ x & x^2 \end{bmatrix} p(x)dx\ \mathbf{w} \\
&= \begin{bmatrix} 1 & E\{X\} \\ E\{X\} & E\{X^2\} \end{bmatrix} \mathbf{w} \\
&= E\{XX^{\mathsf{T}}\}\ \mathbf{w}
\end{aligned}
$$

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- For infinite amount of data the *true* model parameters are obtained from

$$\mathbf{w} = \left(E\{XX^{\mathsf{T}}\}\right)^{-1} E\{TX\}$$

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- For infinite amount of data the *true* model parameters are obtained from

$$\mathbf{w} = \left(E\{XX^{\mathsf{T}}\}\right)^{-1} E\{TX\}$$

Comparing with the Least-Squares estimate we can see how $\widehat{\mathbf{w}}$ is an estimate of $\mathbf{w}$ based on the sample of data available.

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- For infinite amount of data the *true* model parameters are obtained from

$$\mathbf{w} = \left(E\{XX^\mathsf{T}\}\right)^{-1} E\{TX\}$$

  Comparing with the Least-Squares estimate we can see how $\widehat{\mathbf{w}}$ is an estimate of $\mathbf{w}$ based on the sample of data available.

- We would then expect to apportion some of the error observed to the sample based approximations to the expectations appearing in the above equation.

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- Consider the error made at a particular point x_*

$$\int |t - f(x_*; \mathbf{w})|^2 p(t|x_*) dt$$

Bias-Variance Decomposition

UNIVERSITY of GLASGOW

- Consider the error made at a particular point x_*

$$\int |t - f(x_*; \mathbf{w})|^2 p(t|x_*) dt$$

Differentiating with respect to $f(x_*; \mathbf{w})$ and setting to zero we find that

$$f(x_*; \mathbf{w}) \int p(t|x_*) dt = f(x_*; \mathbf{w}) = \int t p(t|x_*) dt = E\{T|x_*\}$$

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- Consider the error made at a particular point x_*

$$\int |t - f(x_*; \mathbf{w})|^2 p(t|x_*) dt$$

Differentiating with respect to $f(x_*; \mathbf{w})$ and setting to zero we find that

$$f(x_*; \mathbf{w}) \int p(t|x_*) dt = f(x_*; \mathbf{w}) = \int t p(t|x_*) dt = E\{T|x_*\}$$

- The best function estimate at a point x_* is the conditional expectation $E\{T|x_*\}$ in other words the expected value of t given that the *input* equals x_*. This is the best that we can hope to do.

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- Expected loss, $\int\int |t - f(x;\mathbf{w})|^2 p(t|x)p(x)dxdt$, can be written as

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- Expected loss, $\int\int |t - f(x;\mathbf{w})|^2 p(t|x)p(x)dxdt$, can be written as

$$\int\int |t + E\{T|x\} - E\{T|x\} - f(x;\mathbf{w})|^2 p(t|x)p(x)dxdt =$$

$$\int\int |t - E\{T|x\}|^2 p(t|x)p(x)dxdt +$$

$$\int\int |E\{T|x\} - f(x;\mathbf{w})|^2 p(t|x)p(x)dxdt -$$

$$2\int\int |E\{T|x\} - f(x;\mathbf{w})||t - E\{T|x\}|p(t|x)p(x)dxdt$$

Bias-Variance Decomposition

UNIVERSITY
of
GLASGOW

- It is straightforward to see that the third term above equals zero as

Bias-Variance Decomposition

UNIVERSITY
of
GLASGOW

- It is straightforward to see that the third term above equals zero as

$$2 \int \int |E\{T|x\} - f(x; \mathbf{w})||t - E\{T|x\}|p(t|x)p(x)dxdt =$$

$$2 \int \int |t - E\{T|x\}|p(t|x)dt|E\{T|x\} - f(x; \mathbf{w})|p(x)dx =$$

$$2 \int |E\{T|x\} - E\{T|x\}||E\{T|x\} - f(x; \mathbf{w})|p(x)dx = 0$$

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- Likewise the first term can be written as

Bias-Variance Decomposition

UNIVERSITY of GLASGOW

- Likewise the first term can be written as

$$\int\int |t - E\{T|x\}|^2 p(t|x)p(x)dxdt =$$

$$\int\int \left(t^2 + E^2\{T|x\} - 2tE\{T|x\}\right) p(t|x)p(x)dxdt =$$

$$\int \left(E\{T^2|x\} + E^2\{T|x\} - 2E^2\{T|x\}\right) p(x)dx =$$

$$\int \left(E\{T^2|x\} - E^2\{T|x\}\right) p(x)dx$$

Bias-Variance Decomposition

UNIVERSITY of GLASGOW

- Likewise the first term can be written as

$$\int\int |t - E\{T|x\}|^2 p(t|x)p(x)dxdt =$$

$$\int\int \left(t^2 + E^2\{T|x\} - 2tE\{T|x\}\right) p(t|x)p(x)dxdt =$$

$$\int \left(E\{T^2|x\} + E^2\{T|x\} - 2E^2\{T|x\}\right) p(x)dx =$$

$$\int \left(E\{T^2|x\} - E^2\{T|x\}\right) p(x)dx$$

- This gives the variance of the output (target) around the conditional mean value (which is the best estimate of the target value), characterizes the data noise and so the uncertainty in the target value estimates.

Bias-Variance Decomposition

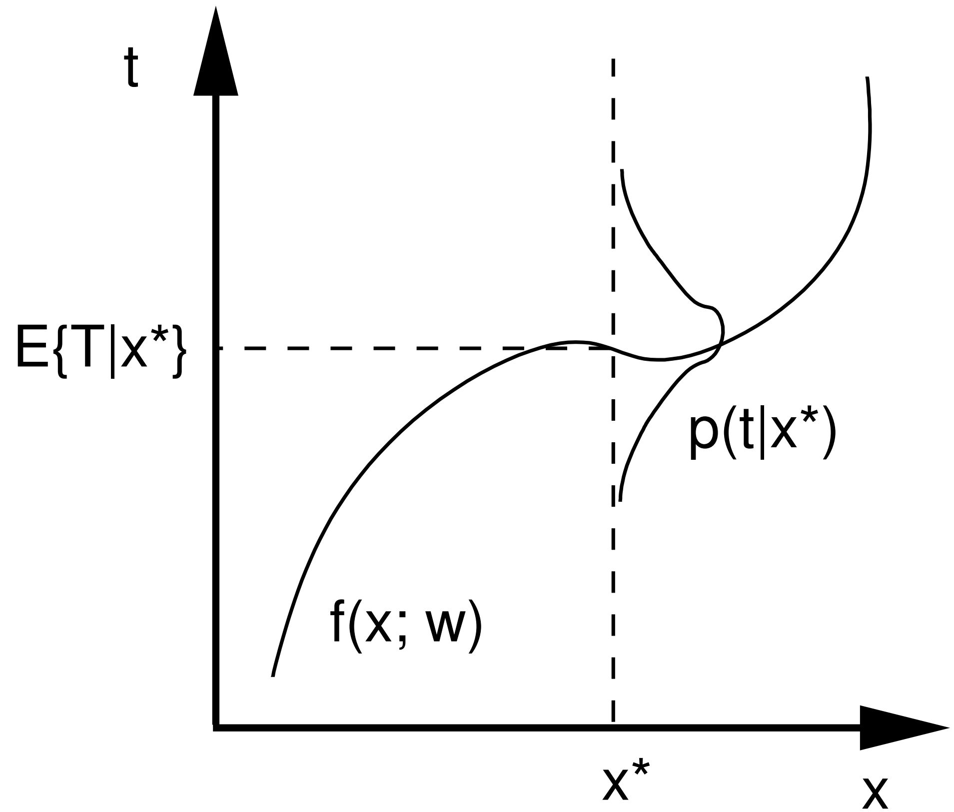

Figure 2: Diagram illustrating the irreducible component of error. The true function to be estimated is $f(x;\mathbf{w})$ and the best estimate in the mean square sense is the conditional mean $E\{T|x^*\}$ however we also see that the conditional distribution $p(t|X^*)$ will have a finite variance $E\{T^2|x^*\} - E^2\{T|x^*\}$ which contributes to the overall error.

Bias-Variance Decomposition

- Second term, $\int \int |E\{T|x\} - f(x; \mathbf{w})|^2 p(t|x)p(x)dxdt$

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- Second term, $\int \int |E\{T|x\} - f(x; \mathbf{w})|^2 p(t|x)p(x)dxdt$
- Is an *approximation* error measuring mismatch between our model parameters identified with an infinite amount of data and the parameters estimated from a finite sample.

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- Second term, $\int \int |E\{T|x\} - f(x;\mathbf{w})|^2 p(t|x)p(x)dxdt$
- Is an *approximation* error measuring mismatch between our model parameters identified with an infinite amount of data and the parameters estimated from a finite sample.
- Parameters of model $f(x;\mathbf{w})$ are estimated from a particular data set $\mathcal{D} = (x_n, t_n)_{n=1,\cdots,N}$.

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- Second term, $\int\int |E\{T|x\} - f(x;\mathbf{w})|^2 p(t|x)p(x)dxdt$
- Is an *approximation* error measuring mismatch between our model parameters identified with an infinite amount of data and the parameters estimated from a finite sample.
- Parameters of model $f(x;\mathbf{w})$ are estimated from a particular data set $\mathcal{D} = (x_n, t_n)_{n=1,\cdots,N}$.
- Repeat experiment and obtain another data set $\mathcal{D}'$ then our function estimate would differ somewhat from that obtained from data set $\mathcal{D}$.

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- If there were a sampling distribution for our data sets $P(\mathcal{D})$ then the expected value of our estimated function would be the model of choice i.e. $\int f(x;\mathbf{w})P(\mathcal{D})d\mathcal{D} = E_{P(\mathcal{D})}\{f(x;\mathbf{w})\}$.

Bias-Variance Decomposition

UNIVERSITY of GLASGOW

- If there were a sampling distribution for our data sets $P(\mathcal{D})$ then the expected value of our estimated function would be the model of choice i.e. $\int f(x;\mathbf{w})P(\mathcal{D})d\mathcal{D} = E_{P(\mathcal{D})}\{f(x;\mathbf{w})\}$.
- Recap here and note that each $f(x;\mathbf{w})$ is estimated from a data set $\mathcal{D}$ via the least squares estimator.

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- If there were a sampling distribution for our data sets $P(\mathcal{D})$ then the expected value of our estimated function would be the model of choice i.e. $\int f(x;\mathbf{w})P(\mathcal{D})d\mathcal{D} = E_{P(\mathcal{D})}\{f(x;\mathbf{w})\}$.
- Recap here and note that each $f(x;\mathbf{w})$ is estimated from a data set $\mathcal{D}$ via the least squares estimator.
- Therefore averaging our models over multiple data sets ensures that we have, on average over data sets, a mean-square optimal model.

Bias-Variance Decomposition

- So back to the second term in our error criterion, we can employ the same trick as previous and so

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- So back to the second term in our error criterion, we can employ the same trick as previous and so

$$\int\int |E\{T|x\} - f(x;\mathbf{w})|^2 p(t|x)p(x)dxdt =$$

$$\int\int |E\{T|x\} - E_{P(\mathcal{D})}\{f(x;\mathbf{w})\} + E_{P(\mathcal{D})}\{f(x;\mathbf{w})\} - f(x;\mathbf{w})|^2 p(t|x)p(x)dxdt =$$

$$\int\int |E\{T|x\} - E_{P(\mathcal{D})}\{f(x;\mathbf{w})\}|^2 p(t|x)p(x)dxdt +$$

$$\int\int |E_{P(\mathcal{D})}\{f(x;\mathbf{w})\} - f(x;\mathbf{w})|^2 p(t|x)p(x)dxdt -$$

$$2\int\int |E\{T|x\} - E_{P(\mathcal{D})}\{f(x;\mathbf{w})\}||E_{P(\mathcal{D})}\{f(x;\mathbf{w})\} - f(x;\mathbf{w})|p(t|x)p(x)dxdt$$

Bias-Variance Decomposition

UNIVERSITY
of
GLASGOW

- Now we average this over all possible data sets and we find that, as before the third term is zero

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- Now we average this over all possible data sets and we find that, as before the third term is zero
- All that remains is

$$\int |E\{T|x\} - E_{P(\mathcal{D})}\{f(x;\mathbf{w})\}|^2 p(x)dx +$$

$$\int E_{P(\mathcal{D})}\left\{|E_{P(\mathcal{D})}\{f(x;\mathbf{w})\} - f(x;\mathbf{w})|^2\right\} p(x)dx$$

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- Now we average this over all possible data sets and we find that, as before the third term is zero
- All that remains is

$$\int |E\{T|x\} - E_{P(\mathcal{D})}\{f(x;\mathbf{w})\}|^2 p(x)dx +$$
$$\int E_{P(\mathcal{D})}\left\{|E_{P(\mathcal{D})}\{f(x;\mathbf{w})\} - f(x;\mathbf{w})|^2\right\} p(x)dx$$

- The expectation does not appear in 1st term as it is independent of data set, as both terms independent of target values $\int p(t|x)dt = 1$ so integral with respect to t drops out

Bias-Variance Decomposition

- At long and weary last we can look at the overall expression for the expected loss and here we also take expectations with respect to the data sets.

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- At long and weary last we can look at the overall expression for the expected loss and here we also take expectations with respect to the data sets.

$$\int \int E_{P(\mathcal{D})}\{|t - f(x;\mathbf{w})|^2\}p(t|x)p(x)dxdt =$$

$$\int \left(E\{T^2|x\} - E^2\{T|x\}\right) p(x)dx + \qquad (1)$$

$$\int |E\{T|x\} - E_{P(\mathcal{D})}\{f(x;\mathbf{w})\}|^2 p(x)dx + \qquad (2)$$

$$\int E_{P(\mathcal{D})} \left\{|E_{P(\mathcal{D})}\{f(x;\mathbf{w})\} - f(x;\mathbf{w})|^2\right\} p(x)dx \qquad (3)$$

Bias-Variance Decomposition

- The first term, $\int \left(E\{T^2|x\} - E^2\{T|x\} \right) p(x)dx$, defines the irreducible error, irrespective of model, caused by noise in the observations.

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- The first term, $\int \left(E\{T^2|x\} - E^2\{T|x\}\right) p(x)dx$, defines the irreducible error, irrespective of model, caused by noise in the observations.
- The second term, $\int |E\{T|x\} - E_{P(\mathcal{D})}\{f(x;\mathbf{w})\}|^2 p(x)dx$, is the bias squared, a measure of structural miss-match between model and underlying data generating function.

Bias-Variance Decomposition

UNIVERSITY of GLASGOW

- The first term, $\int \left(E\{T^2|x\} - E^2\{T|x\}\right) p(x)dx$, defines the irreducible error, irrespective of model, caused by noise in the observations.
- The second term, $\int |E\{T|x\} - E_{P(\mathcal{D})}\{f(x;\mathbf{w})\}|^2 p(x)dx$, is the bias squared, a measure of structural miss-match between model and underlying data generating function.
- Adopting too simple a functional class for model, insufficiently flexible, then averaged estimate $E_{P(\mathcal{D})}\{f(x;\mathbf{w})\}$ is biased away from the conditional-mean $E\{T|x\}$. Model bias can be reduced by employing appropriately expressive functional classes.

Bias-Variance Decomposition

UNIVERSITY
of
GLASGOW

- The third term, $\int E_{P(\mathcal{D})}\left\{|E_{P(\mathcal{D})}\{f(x;\mathbf{w})\} - f(x;\mathbf{w})|^2\right\}p(x)dx$, is referred to as the variance giving a measure of how much predictions between training data sets will vary.

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- The third term, $\int E_{P(\mathcal{D})}\left\{|E_{P(\mathcal{D})}\{f(x;\mathbf{w})\} - f(x;\mathbf{w})|^2\right\} p(x)dx$, is referred to as the variance giving a measure of how much predictions between training data sets will vary.
- Model variance is something which we must control carefully as highly variable predictions will be unreliable.

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

- The third term, $\int E_{P(\mathcal{D})}\left\{|E_{P(\mathcal{D})}\{f(x;\mathbf{w})\} - f(x;\mathbf{w})|^2\right\} p(x)dx$, is referred to as the variance giving a measure of how much predictions between training data sets will vary.
- Model variance is something which we must control carefully as highly variable predictions will be unreliable.
- Whilst a more complex model will reduce the bias there may be a corresponding increase in the variance and it is this trade-off between the two competing criteria that is the focus of much attention in devising predictive models for real applications

Bias-Variance Decomposition

UNIVERSITY *of* GLASGOW

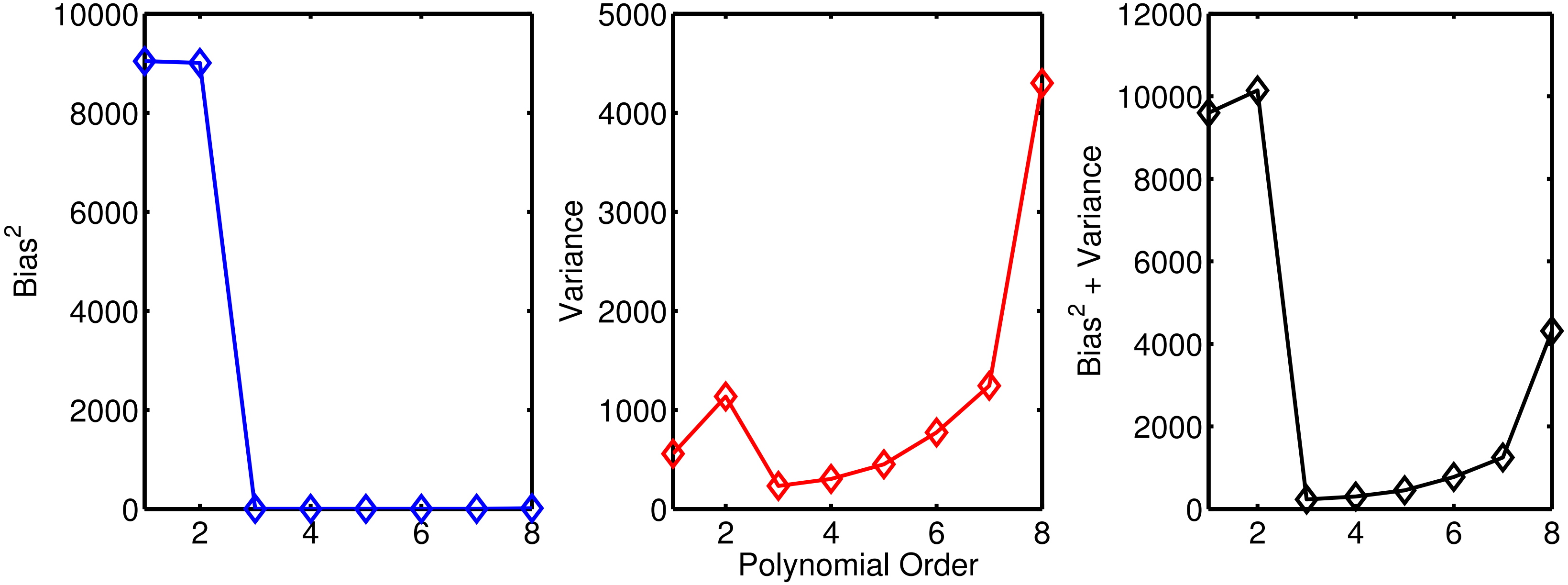

Figure 3: The leftmost plot shows the estimated $bias^2$ for a polynomial model, the middle plot shows the corresponding estimated variance, the rightmost plot gives the cumulative effect of both $bias^2$ + variance. As complexity of the model increases $bias^2$ continually decreases providing an increasingly superior fit to the data. Whilst variance may increase with model complexity with the net effect being that the minimum of $bias^2$ + variance (the expected loss minus the constant term) is achieved at $K = 3$ which is the correct complexity for the function being approximated.

Bias-Variance

UNIVERSITY *of* GLASGOW

- The bias-variance decomposition demonstrates that despite more complex models being able to better describe the available data the variation, in terms of generalisation capability, will increase.

Bias-Variance

UNIVERSITY *of* GLASGOW

- The bias-variance decomposition demonstrates that despite more complex models being able to better describe the available data the variation, in terms of generalisation capability, will increase.
- In many real modeling situations *true* model will not be part of functional class selected, driving the model bias as low as possible is clearly an unwise strategy to follow.

Bias-Variance

UNIVERSITY
of
GLASGOW

- The bias-variance decomposition demonstrates that despite more complex models being able to better describe the available data the variation, in terms of generalisation capability, will increase.
- In many real modeling situations *true* model will not be part of functional class selected, driving the model bias as low as possible is clearly an unwise strategy to follow.
- The Least-Squares estimator happens to be an unbiased estimator.

Bias-Variance

UNIVERSITY *of* GLASGOW

- The bias-variance decomposition demonstrates that despite more complex models being able to better describe the available data the variation, in terms of generalisation capability, will increase.
- In many real modeling situations *true* model will not be part of functional class selected, driving the model bias as low as possible is clearly an unwise strategy to follow.
- The Least-Squares estimator happens to be an unbiased estimator.
- Unbiased estimator may not be most appropriate in many applications.

Cross-Validation

- Require measure of *expected loss* to provide indication of the generalisation ability of predictive models

Cross-Validation

UNIVERSITY *of* GLASGOW

- Require measure of *expected loss* to provide indication of the generalisation ability of predictive models
- From the bias-variance decomposition increasing model complexity reduces model bias reflected in a lower training error.

Cross-Validation

UNIVERSITY *of* GLASGOW

- Require measure of *expected loss* to provide indication of the generalisation ability of predictive models
- From the bias-variance decomposition increasing model complexity reduces model bias reflected in a lower training error.
- Training error obtained from same data used for parameter estimation so provides optimistic estimate of the achievable test error

Cross-Validation

UNIVERSITY *of* GLASGOW

- Require measure of *expected loss* to provide indication of the generalisation ability of predictive models
- From the bias-variance decomposition increasing model complexity reduces model bias reflected in a lower training error.
- Training error obtained from same data used for parameter estimation so provides optimistic estimate of the achievable test error
- Cross-validation directly estimates generalisation (test) error simply by holding out a fraction of training data and using this to obtain a prediction error.

Cross-Validation

UNIVERSITY
of
GLASGOW

- Given a data set $\mathcal{D} = (x_1, t_1), \cdots, (x_N, t_N)$, remove one input and target pair, say (x_i, t_i), so creating the data sample $\mathcal{D}_{-i}$

Cross-Validation

UNIVERSITY *of* GLASGOW

- Given a data set $\mathcal{D} = (x_1, t_1), \cdots, (x_N, t_N)$, remove one input and target pair, say (x_i, t_i), so creating the data sample $\mathcal{D}_{-i}$
- Use $\mathcal{D}_{-i}$ to induce our learning machine, e.g.

$$\widehat{\mathbf{w}}_{-i} = \left(\mathbf{X}_{-i}^{\mathsf{T}}\mathbf{X}_{-i}\right)^{-1}\mathbf{X}_{-i}^{\mathsf{T}}\mathbf{t}_{-i}$$

Cross-Validation

UNIVERSITY of GLASGOW

- Given a data set $\mathcal{D} = (x_1, t_1), \cdots, (x_N, t_N)$, remove one input and target pair, say (x_i, t_i), so creating the data sample $\mathcal{D}_{-i}$
- Use $\mathcal{D}_{-i}$ to induce our learning machine, e.g.

$$\widehat{\mathbf{w}}_{-i} = \left(\mathbf{X}_{-i}^{\mathsf{T}}\mathbf{X}_{-i}\right)^{-1}\mathbf{X}_{-i}^{\mathsf{T}}\mathbf{t}_{-i}$$

The $(N-1) \times (K+1)$ matrix with ith row removed is $\mathbf{X}_{-i}$, the $(N-1) \times 1$ vector with ith element removed is $\mathbf{t}_{-i}$ & $\widehat{\mathbf{w}}_{-i}$ is least-squares estimate based on $\mathcal{D}_{-i}$

Cross-Validation

UNIVERSITY *of* GLASGOW

- Given a data set $\mathcal{D} = (x_1, t_1), \cdots, (x_N, t_N)$, remove one input and target pair, say (x_i, t_i), so creating the data sample $\mathcal{D}_{-i}$
- Use $\mathcal{D}_{-i}$ to induce our learning machine, e.g.

$$\widehat{\mathbf{w}}_{-i} = \left(\mathbf{X}_{-i}^{\mathsf{T}}\mathbf{X}_{-i}\right)^{-1}\mathbf{X}_{-i}^{\mathsf{T}}\mathbf{t}_{-i}$$

  The $(N-1) \times (K+1)$ matrix with ith row removed is $\mathbf{X}_{-i}$, the $(N-1) \times 1$ vector with ith element removed is $\mathbf{t}_{-i}$ & $\widehat{\mathbf{w}}_{-i}$ is least-squares estimate based on $\mathcal{D}_{-i}$
- For the held-out *input-target* pair (x_i, t_i) we can compute the corresponding loss $\mathcal{L}(t_i, f(x_i; \widehat{\mathbf{w}}_{-i}))$, e.g $|t_i - \widehat{\mathbf{w}}_{-i}^{\mathsf{T}}\mathbf{x}_i|^2$ where $\mathbf{x}_i$ is the ith row of $\mathbf{X}$

Cross-Validation

UNIVERSITY *of* GLASGOW

- Perform this procedure N times cycling through all the data and leaving each one out in turn and so our Leave-One-Out estimate of the generalisation error or expected loss will simply be

Cross-Validation

- Perform this procedure N times cycling through all the data and leaving each one out in turn and so our Leave-One-Out estimate of the generalisation error or expected loss will simply be

$$
\begin{aligned}
\mathcal{L}_{cv} &= \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(t_i, f(x_i; \widehat{\mathbf{w}}_{-i})) \\
&= \frac{1}{N}\sum_{i=1}^{N}|t_i - \widehat{\mathbf{w}}_{-i}^{\mathsf{T}}\mathbf{x}_i|^2
\end{aligned}
$$

Cross-Validation

UNIVERSITY of GLASGOW

- Perform this procedure N times cycling through all the data and leaving each one out in turn and so our Leave-One-Out estimate of the generalisation error or expected loss will simply be

$$\begin{aligned}\mathcal{L}_{cv} &= \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(t_i, f(x_i; \widehat{\mathbf{w}}_{-i})) \\ &= \frac{1}{N}\sum_{i=1}^{N}|t_i - \widehat{\mathbf{w}}_{-i}^{\mathsf{T}}\mathbf{x}_i|^2\end{aligned}$$

- Cross-Validation is entirely general with regard to the loss function for which it can estimate the expectation.

Cross-Validation

UNIVERSITY *of* GLASGOW

- Fifty input-target pairs from a noisy third-order polynomial function are sampled and these are used to learn a polynomial regression function.

Cross-Validation

UNIVERSITY
of
GLASGOW

- Fifty input-target pairs from a noisy third-order polynomial function are sampled and these are used to learn a polynomial regression function.
- A further 1000 input-target pairs are used as an independent test set with which to compute the overall test error.

Cross-Validation

UNIVERSITY
of
GLASGOW

- Fifty input-target pairs from a noisy third-order polynomial function are sampled and these are used to learn a polynomial regression function.
- A further 1000 input-target pairs are used as an independent test set with which to compute the overall test error.
- In addition we use the LOOCV estimator as described above to estimate the expected test-error

Cross-Validation

UNIVERSITY of GLASGOW

- Fifty input-target pairs from a noisy third-order polynomial function are sampled and these are used to learn a polynomial regression function.
- A further 1000 input-target pairs are used as an independent test set with which to compute the overall test error.
- In addition we use the LOOCV estimator as described above to estimate the expected test-error
- A range of polynomial orders are considered from order 1 (linear model) up to 10th order (highly flexible model and for each model-order the training error, test error and LOOCV error are computed.

Cross-Validation

UNIVERSITY
of
GLASGOW

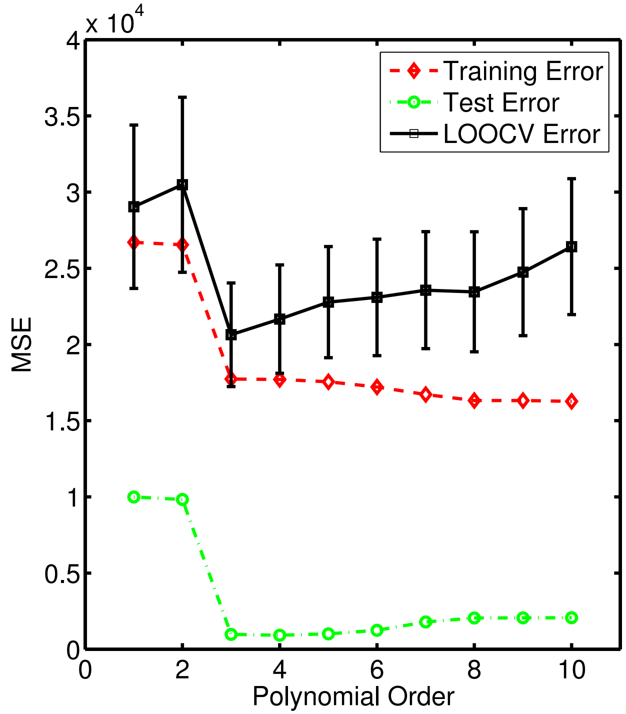

Figure 4: The Training, Testing and Leave-One-Out error curves obtained for a noisy cubic function where a sample size of 50 is available for training and LOOCV estimation. The test error is computed using 1000 independent samples.

CV Scaling

UNIVERSITY *of* GLASGOW

- We are looping N times and within the loop we have to perform our *training* method which in this case is obtaining the Least-Squares solution

CV Scaling

UNIVERSITY
of
GLASGOW

- We are looping N times and within the loop we have to perform our *training* method which in this case is obtaining the Least-Squares solution
- Least-Squares solution requires a matrix inversion that scales as $\mathcal{O}((K+1)^3)$ where $K+1$ is the dimension of the matrix being inverted

CV Scaling

UNIVERSITY *of* GLASGOW

- We are looping N times and within the loop we have to perform our *training* method which in this case is obtaining the Least-Squares solution
- Least-Squares solution requires a matrix inversion that scales as $\mathcal{O}((K+1)^3)$ where $K+1$ is the dimension of the matrix being inverted
- Matrix multiplications will contribute $\mathcal{O}(N(K+1)^2 + 2N(K+1)^3)$ scaling

CV Scaling

- We are looping N times and within the loop we have to perform our *training* method which in this case is obtaining the Least-Squares solution
- Least-Squares solution requires a matrix inversion that scales as $\mathcal{O}((K+1)^3)$ where $K+1$ is the dimension of the matrix being inverted
- Matrix multiplications will contribute $\mathcal{O}(N(K+1)^2 + 2N(K+1)^3)$ scaling
- Overall dominant scaling for LOOCV is $\mathcal{O}(N^2(K+1)^3)$. As either K or N become large we can see that LOOCV can become rather expensive computationally