ISTA 421 + INFO 521 Introduction to Machine Learning

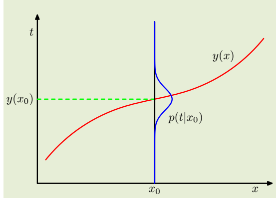
Lecture 10: Maximum Likelihood Uncertainty 2

Clay Morrison
claytonm@email.arizona.edu
Harvill 437A
Phone 621-6609

25 September 2017

SISTA 1

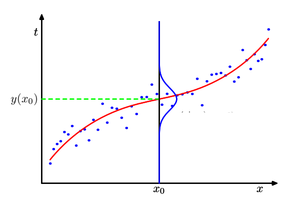
Review



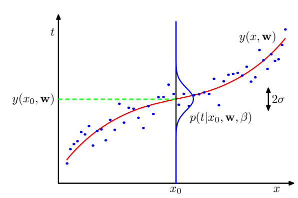
The generating process…

$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n \;\;;\;\; \epsilon \sim \mathcal{N}(0, \sigma^2)$

$$L = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^\top \mathbf{x}_n, \sigma^2)$$



… generates data …



… that we fit a model to

$$p(\hat{\mathbf{t}}|\mathbf{X}, \hat{\mathbf{w}}, \hat{\sigma^2}) = \prod_{n=1}^{N} p(\hat{t}_n|\mathbf{x}_n, \hat{\mathbf{w}}, \hat{\sigma^2}) = \prod_{n=1}^{N} \mathcal{N}(\hat{\mathbf{w}}^\top \mathbf{x}_n, \hat{\sigma^2})$$

prediction

estimated parameters

Maximum Likelihood Estimates of Params

$\hat{\mathbf{w}}$ $\quad \widehat{\mathbf{w}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{t}$

$\hat{\sigma}^2$ $\quad \widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^\top\mathbf{t} - \mathbf{t}^\top\mathbf{X}\widehat{\mathbf{w}})$

The MLE is unique

$\frac{\partial^2 \log L}{\partial \mathbf{w} \partial \mathbf{w}^\top} = -\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}$

Analysis of Uncertainty in Param Estimates via Expectation

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

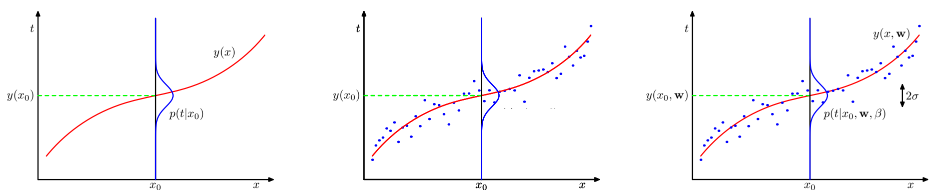
The Fisher Information $\mathcal{I} = \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}$

$\text{cov}\{\widehat{\mathbf{w}}\} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} = -\left(\frac{\partial^2 \log L}{\partial \mathbf{w} \partial \mathbf{w}^\top}\right)^{-1}$

Predictions:

SISTA 2

Back to: The Generative Picture



The generating process… ... generates data ... ... that we fit a model to

$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n \;\; ; \;\; \epsilon \sim \mathcal{N}(0, \sigma^2)$

How about modeling the uncertainty in our predictions?

Want model of region of values in which our prediction might fall

estimated parameters

$$p(\hat{\mathbf{t}}|\mathbf{X}, \hat{\mathbf{w}}, \hat{\sigma^2}) = \prod_{n=1}^{N} p(\hat{t}_n|\mathbf{x}_n, \hat{\mathbf{w}}, \hat{\sigma^2})$$
$$= \prod_{n=1}^{N} \mathcal{N}(\hat{\mathbf{w}}^\top \mathbf{x}_n, \hat{\sigma^2})$$

predictive distribution

SISTA 3

Variability in Predictions

- We are predicting 2 values:

$$t_{new} \;,\; \sigma^2_{new}$$

SISTA 4

Variability in Predictions

- We are predicting 2 values:

$$t_{new} \ , \ \sigma^2_{new}$$

$t_{new} = \hat{\mathbf{w}}^\top \mathbf{x}_{new}$ Same solution as minimizing mean squared loss

SISTA 5

Variability in Predictions

- We are predicting 2 values:

$$t_{new} \ , \ \sigma^2_{new}$$

$t_{new} = \hat{\mathbf{w}}^\top \mathbf{x}_{new}$ Same solution as minimizing mean squared loss

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

SISTA 6

Variability in Predictions

- We are predicting 2 values:

$$t_{new} \;,\; \sigma^2_{new}$$

$t_{new} = \hat{\mathbf{w}}^\top \mathbf{x}_{new}$ Same solution as minimizing mean squared loss

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

7

Variability in Predictions

- We are predicting 2 values:

$$t_{new} \;,\; \sigma^2_{new}$$

$t_{new} = \hat{\mathbf{w}}^\top \mathbf{x}_{new}$ Same solution as minimizing mean squared loss

$$\begin{aligned}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{t_{new}\right\} &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\hat{\mathbf{w}}\right\}^\top \mathbf{x}_{new} \\ &= \mathbf{w}^\top \mathbf{x}_{new}\end{aligned}$$

8

Variability in Predictions

- We are predicting 2 values:

$$t_{new} \ , \ \sigma^2_{new}$$

$$t_{new} = \hat{\mathbf{w}}^\top \mathbf{x}_{new}$$

Same solution as minimizing mean squared loss

$$\begin{aligned}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{t_{new}\right\} &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\hat{\mathbf{w}}\right\}^\top \mathbf{x}_{new} \\ &= \mathbf{w}^\top \mathbf{x}_{new}\end{aligned}$$

The ***expected value*** of our prediction is the new data attribute multiplied by the ***true w***

SISTA 9

Predicting the Variance of t_{new}

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

SISTA 10

Predicting the Variance of t_{new}

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

Substitute $t_{new} = \hat{\mathbf{w}}^\top \mathbf{x}_{new}$

11

Predicting the Variance of t_{new}

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

Substitute $t_{new} = \hat{\mathbf{w}}^\top \mathbf{x}_{new}$

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

12

Predicting the Variance of t_{new}

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

Substitute $t_{new} = \hat{\mathbf{w}}^\top \mathbf{x}_{new}$

$$\begin{aligned}\text{var}\{t_{\text{new}}\} &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{(\widehat{\mathbf{w}}^\top\mathbf{x}_{\text{new}})^2\right\} - (\mathbf{w}^\top\mathbf{x}_{\text{new}})^2 \\ &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{x}_{\text{new}}^\top\widehat{\mathbf{w}}\widehat{\mathbf{w}}^\top\mathbf{x}_{\text{new}}\right\} - \mathbf{x}_{\text{new}}^\top\mathbf{w}\mathbf{w}^\top\mathbf{x}_{\text{new}}.\end{aligned}$$

Substitute $\widehat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t}$

SISTA 13

Predicting the Variance of t_{new}

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

Substitute $t_{new} = \hat{\mathbf{w}}^\top \mathbf{x}_{new}$

$$\begin{aligned}\text{var}\{t_{\text{new}}\} &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{(\widehat{\mathbf{w}}^\top\mathbf{x}_{\text{new}})^2\right\} - (\mathbf{w}^\top\mathbf{x}_{\text{new}})^2 \\ &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{x}_{\text{new}}^\top\widehat{\mathbf{w}}\widehat{\mathbf{w}}^\top\mathbf{x}_{\text{new}}\right\} - \mathbf{x}_{\text{new}}^\top\mathbf{w}\mathbf{w}^\top\mathbf{x}_{\text{new}}.\end{aligned}$$

Substitute $\widehat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t}$

$$\text{var}\{t_{\text{new}}\} = \mathbf{x}_{\text{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\mathbf{t}^\top\right\}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{\text{new}} - \mathbf{x}_{\text{new}}^\top\mathbf{w}\mathbf{w}^\top\mathbf{x}_{\text{new}}$$

SISTA 14

Predicting the Variance of t_{new}

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

Substitute $t_{new} = \hat{\mathbf{w}}^\top \mathbf{x}_{new}$

$$\begin{aligned}\text{var}\{t_{\text{new}}\} &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{(\widehat{\mathbf{w}}^\top \mathbf{x}_{\text{new}})^2\right\} - (\mathbf{w}^\top \mathbf{x}_{\text{new}})^2 \\ &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{x}_{\text{new}}^\top \widehat{\mathbf{w}}\widehat{\mathbf{w}}^\top \mathbf{x}_{\text{new}}\right\} - \mathbf{x}_{\text{new}}^\top \mathbf{w}\mathbf{w}^\top \mathbf{x}_{\text{new}}.\end{aligned}$$

Substitute $\widehat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t}$

$$\text{var}\{t_{\text{new}}\} = \mathbf{x}_{\text{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\mathbf{t}^\top\right\}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{\text{new}} - \mathbf{x}_{\text{new}}^\top\mathbf{w}\mathbf{w}^\top\mathbf{x}_{\text{new}}$$

On slide 23 of Lec 9, in the derivation of the covariance of $\hat{\boldsymbol{w}}$, we identified

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\mathbf{t}\mathbf{t}^\top\} = \mathbf{X}\mathbf{w}\mathbf{w}^\top\mathbf{X}^\top - \sigma^2\mathbf{I}$$

SISTA 15

Predicting the Variance of t_{new}

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

Substitute $t_{new} = \hat{\mathbf{w}}^\top \mathbf{x}_{new}$

$$\begin{aligned}\text{var}\{t_{\text{new}}\} &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{(\widehat{\mathbf{w}}^\top \mathbf{x}_{\text{new}})^2\right\} - (\mathbf{w}^\top \mathbf{x}_{\text{new}})^2 \\ &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{x}_{\text{new}}^\top \widehat{\mathbf{w}}\widehat{\mathbf{w}}^\top \mathbf{x}_{\text{new}}\right\} - \mathbf{x}_{\text{new}}^\top \mathbf{w}\mathbf{w}^\top \mathbf{x}_{\text{new}}.\end{aligned}$$

Substitute $\widehat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t}$

$$\text{var}\{t_{\text{new}}\} = \mathbf{x}_{\text{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\mathbf{t}^\top\right\}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{\text{new}} - \mathbf{x}_{\text{new}}^\top\mathbf{w}\mathbf{w}^\top\mathbf{x}_{\text{new}}$$

On slide 23 of Lec 9, in the derivation of the covariance of $\hat{\boldsymbol{w}}$, we identified

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\mathbf{t}\mathbf{t}^\top\} = \mathbf{X}\mathbf{w}\mathbf{w}^\top\mathbf{X}^\top - \sigma^2\mathbf{I}$$

$$\begin{aligned}\text{var}\{t_{\text{new}}\} &= \mathbf{x}_{\text{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\sigma^2\mathbf{I} + \mathbf{X}\mathbf{w}\mathbf{w}^\top\mathbf{X}^\top)\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{\text{new}} - \mathbf{x}_{\text{new}}^\top\mathbf{w}\mathbf{w}^\top\mathbf{x}_{\text{new}} \\ &= \sigma^2\mathbf{x}_{\text{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{\text{new}} + \mathbf{x}_{\text{new}}^\top\mathbf{w}\mathbf{w}^\top\mathbf{x}_{\text{new}} - \mathbf{x}_{\text{new}}^\top\mathbf{w}\mathbf{w}^\top\mathbf{x}_{\text{new}} \\ &= \sigma^2\mathbf{x}_{\text{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{\text{new}}.\end{aligned}$$

SISTA 16

Predicting the Variance of t_{new}

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

Substitute $t_{new} = \hat{\mathbf{w}}^\top \mathbf{x}_{new}$

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

Substitute $\widehat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t}$

$$\mathrm{var}\{t_{\mathrm{new}}\} = \mathbf{x}_{\mathrm{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\mathbf{t}^\top\right\}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{\mathrm{new}} - \mathbf{x}_{\mathrm{new}}^\top\mathbf{w}\mathbf{w}^\top\mathbf{x}_{\mathrm{new}}$$

On slide 23 of Lec 9, in the derivation of the covariance of $\hat{\boldsymbol{w}}$, we identified

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\mathbf{t}\mathbf{t}^\top\} = \mathbf{X}\mathbf{w}\mathbf{w}^\top\mathbf{X}^\top - \sigma^2\mathbf{I}$$

$$\begin{aligned}\mathrm{var}\{t_{\mathrm{new}}\} &= \mathbf{x}_{\mathrm{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\sigma^2\mathbf{I} + \mathbf{X}\mathbf{w}\mathbf{w}^\top\mathbf{X}^\top)\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{\mathrm{new}} - \mathbf{x}_{\mathrm{new}}^\top\mathbf{w}\mathbf{w}^\top\mathbf{x}_{\mathrm{new}} \\ &= \sigma^2\mathbf{x}_{\mathrm{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{\mathrm{new}} + \mathbf{x}_{\mathrm{new}}^\top\mathbf{w}\mathbf{w}^\top\mathbf{x}_{\mathrm{new}} - \mathbf{x}_{\mathrm{new}}^\top\mathbf{w}\mathbf{w}^\top\mathbf{x}_{\mathrm{new}} \\ &= \sigma^2\mathbf{x}_{\mathrm{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{\mathrm{new}}.\end{aligned}$$

Recall: $\mathrm{cov}\{\widehat{\mathbf{w}}\} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} = -\left(\frac{\partial^2 \log L}{\partial\mathbf{w}\partial\mathbf{w}^\top}\right)^{-1}$

SISTA 17

Predicting the Variance of t_{new}

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

Substitute $t_{new} = \hat{\mathbf{w}}^\top \mathbf{x}_{new}$

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

Substitute $\widehat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t}$

$$\mathrm{var}\{t_{\mathrm{new}}\} = \mathbf{x}_{\mathrm{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\mathbf{t}^\top\right\}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{\mathrm{new}} - \mathbf{x}_{\mathrm{new}}^\top\mathbf{w}\mathbf{w}^\top\mathbf{x}_{\mathrm{new}}$$

On slide 23 of Lec 9, in the derivation of the covariance of $\hat{\boldsymbol{w}}$, we identified

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\{\mathbf{t}\mathbf{t}^\top\} = \mathbf{X}\mathbf{w}\mathbf{w}^\top\mathbf{X}^\top - \sigma^2\mathbf{I}$$

$$\begin{aligned}\mathrm{var}\{t_{\mathrm{new}}\} &= \mathbf{x}_{\mathrm{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\sigma^2\mathbf{I} + \mathbf{X}\mathbf{w}\mathbf{w}^\top\mathbf{X}^\top)\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{\mathrm{new}} - \mathbf{x}_{\mathrm{new}}^\top\mathbf{w}\mathbf{w}^\top\mathbf{x}_{\mathrm{new}} \\ &= \sigma^2\mathbf{x}_{\mathrm{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{\mathrm{new}} + \mathbf{x}_{\mathrm{new}}^\top\mathbf{w}\mathbf{w}^\top\mathbf{x}_{\mathrm{new}} - \mathbf{x}_{\mathrm{new}}^\top\mathbf{w}\mathbf{w}^\top\mathbf{x}_{\mathrm{new}} \\ &= \sigma^2\mathbf{x}_{\mathrm{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{\mathrm{new}}.\end{aligned}$$

Recall: $\mathrm{cov}\{\widehat{\mathbf{w}}\} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} = -\left(\frac{\partial^2 \log L}{\partial\mathbf{w}\partial\mathbf{w}^\top}\right)^{-1}$

So, could be written

$$\sigma^2_{\mathrm{new}} = \mathbf{x}_{\mathrm{new}}^\top \mathrm{cov}\{\widehat{\mathbf{w}}\}\mathbf{x}_{\mathrm{new}}$$

SISTA 18

In Summary

$$t_{\text{new}} = \mathbf{x}_{\text{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t} = \mathbf{x}_{\text{new}}^\top\widehat{\mathbf{w}}$$

$$\sigma^2_{\text{new}} = \sigma^2\mathbf{x}_{\text{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{\text{new}}$$

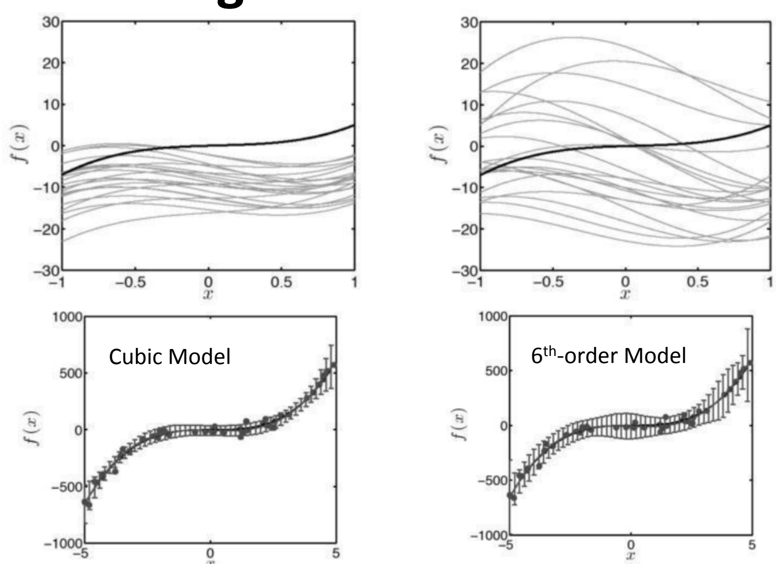
We estimate this from the data: $\hat{\sigma}^2$

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

19

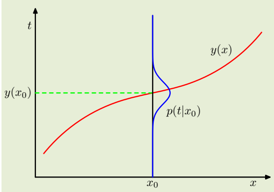
$t_{\text{new}} \pm \sigma^2_{\text{new}}$

Plotting Predictive Error Bars



20

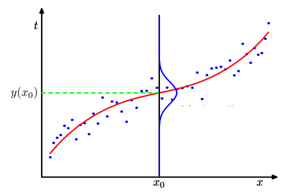
Review



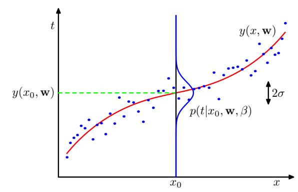
The generating process…

$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n \;\; ; \;\; \epsilon \sim \mathcal{N}(0, \sigma^2)$

$$L = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^\top \mathbf{x}_n, \sigma^2)$$



… generates data …



… that we fit a model to

$$p(\hat{\mathbf{t}}|\mathbf{X}, \hat{\mathbf{w}}, \hat{\sigma^2}) = \prod_{n=1}^{N} p(\hat{t}_n|\mathbf{x}_n, \hat{\mathbf{w}}, \hat{\sigma^2}) = \prod_{n=1}^{N} \mathcal{N}(\hat{\mathbf{w}}^\top \mathbf{x}_n, \hat{\sigma^2})$$

prediction

estimated parameters

Maximum Likelihood Estimates of Params

The MLE is unique

Analysis of Uncertainty in Param Estimates via Expectation

The Fisher Information $\mathcal{I} = \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}$

$\hat{\mathbf{w}}$: $\widehat{\mathbf{w}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{t}$ $\quad \frac{\partial^2 \log L}{\partial \mathbf{w} \partial \mathbf{w}^\top} = -\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}$ $\quad \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\ \mathbf{w},\sigma^2)}\left\{\widehat{\mathbf{w}}\right\} = \mathbf{w}$ $\quad \mathrm{cov}\{\widehat{\mathbf{w}}\} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} = -\left(\frac{\partial^2 \log L}{\partial \mathbf{w} \partial \mathbf{w}^\top}\right)^{-1}$

$\hat{\sigma}^2$: $\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^\top\mathbf{t} - \mathbf{t}^\top\mathbf{X}\widehat{\mathbf{w}})$

Predictions: $t_{new} = \hat{\mathbf{w}}^\top \mathbf{x}_{new}$ $\quad \sigma^2_{\text{new}} = \sigma^2 \mathbf{x}_{\text{new}}^\top (\mathbf{X}^\top\mathbf{X})^{-1} \mathbf{x}_{\text{new}}$ $\quad \sigma^2_{\text{new}} = \mathbf{x}_{\text{new}}^\top \mathrm{cov}\{\widehat{\mathbf{w}}\} \mathbf{x}_{\text{new}}$

SISTA 21

Quantifying the Uncertainty in our Estimate of $\hat{\sigma}^2$

$\mathbf{t} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

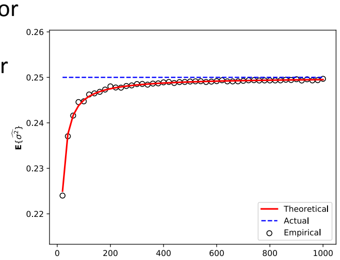
$\mathbf{E}_{p(\mathbf{t})}\left\{\mathbf{t}^\top\mathbf{A}\mathbf{t}\right\} = \mathrm{Tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top\mathbf{A}\boldsymbol{\mu}$ $\quad \mathrm{Tr}(\mathbf{A}) = \sum_{d=1}^{D} A_{dd}.$

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

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

When D < N (that is, the number of attributes we measure for each data point is *smaller* than the number of data points), then our estimates of the variance will, on average, be lower than the true variance.

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



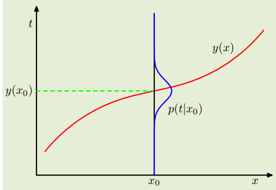
Unlike the estimate for $\hat{\mathbf{w}}$, the MLE for $\hat{\sigma}^2$ is ***biased***.

$$D = 2 \text{ and } N = 20$$

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} = \sigma^2\left(1 - \frac{D}{N}\right) = 0.25\left(1 - \frac{2}{20}\right) = 0.2250$$

SISTA 22

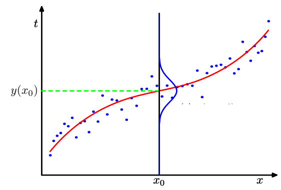
Review



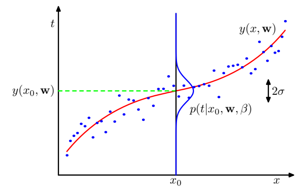
The generating process…

$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n \;\; ; \;\; \epsilon \sim \mathcal{N}(0, \sigma^2)$

$$L = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^\top \mathbf{x}_n, \sigma^2)$$



… generates data …



… that we fit a model to

$$p(\hat{\mathbf{t}}|\mathbf{X}, \hat{\mathbf{w}}, \hat{\sigma^2}) = \prod_{n=1}^{N} p(\hat{t}_n|\mathbf{x}_n, \hat{\mathbf{w}}, \hat{\sigma^2}) = \prod_{n=1}^{N} \mathcal{N}(\hat{\mathbf{w}}^\top \mathbf{x}_n, \hat{\sigma^2})$$

prediction

estimated parameters

Maximum Likelihood Estimates of Params

The MLE is unique

Estimating Uncertainty in Param Estimates via Expected Value

The Fisher Information $\mathcal{I} = \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}$

$\hat{\mathbf{w}}$: $\widehat{\mathbf{w}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{t}$ $\quad \frac{\partial^2 \log L}{\partial \mathbf{w} \partial \mathbf{w}^\top} = -\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}$ $\quad \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\, \mathbf{w}, \sigma^2)}\left\{\widehat{\mathbf{w}}\right\} = \mathbf{w}$ $\quad \mathrm{cov}\{\widehat{\mathbf{w}}\} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} = -\left(\frac{\partial^2 \log L}{\partial \mathbf{w} \partial \mathbf{w}^\top}\right)^{-1}$

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

Predictions: $t_{new} = \hat{\mathbf{w}}^\top \mathbf{x}_{new}$ $\quad \sigma^2_{\text{new}} = \sigma^2 \mathbf{x}_{\text{new}}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_{\text{new}}$ $\quad \sigma^2_{\text{new}} = \mathbf{x}_{\text{new}}^\top \mathrm{cov}\{\widehat{\mathbf{w}}\}\mathbf{x}_{\text{new}}$

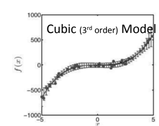
SISTA

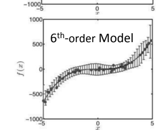
MLE and Model Selection

$$\log L = -\frac{1}{N}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\top\mathbf{x}_n)^2$$

$$\hat{\mathbf{w}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{t}$$

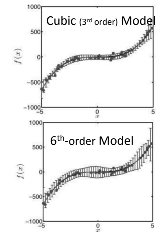
$$\widehat{\sigma^2} = \frac{1}{N}\sum_{n=1}^{N}(t_n - \mathbf{x}^\top\hat{\mathbf{w}})^2 = \frac{1}{N}(\mathbf{t}^\top\mathbf{t} - \mathbf{t}^\top\mathbf{X}\hat{\mathbf{w}})$$





SISTA 24

MLE and Model Selection



$$\log L = -\frac{1}{N}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\top \mathbf{x}_n)^2$$

$$\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t}$$

$$\widehat{\sigma^2} = \frac{1}{N}\sum_{n=1}^{N}(t_n - \mathbf{x}^\top\hat{\mathbf{w}})^2 = \frac{1}{N}(\mathbf{t}^\top\mathbf{t} - \mathbf{t}^\top\mathbf{X}\hat{\mathbf{w}})$$

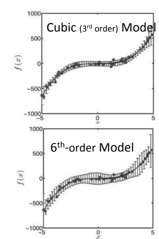
Plug in $\widehat{\sigma^2}$ to the log likelihood:

$$\begin{aligned}\log L &= -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\widehat{\sigma^2} - \frac{1}{2\widehat{\sigma^2}}N\widehat{\sigma^2} \\ &= -\frac{N}{2}(1+\log 2\pi) - \frac{N}{2}\log\widehat{\sigma^2}.\end{aligned}$$

Making $\hat{\sigma}^2$ smaller makes log L larger. Making model more flexible decreases $\hat{\sigma}^2$.

SISTA 25

MLE Prefers Complex Models



$$\log L = -\frac{1}{N}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\top \mathbf{x}_n)^2$$

$$\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t}$$

$$\widehat{\sigma^2} = \frac{1}{N}\sum_{n=1}^{N}(t_n - \mathbf{x}^\top\hat{\mathbf{w}})^2 = \frac{1}{N}(\mathbf{t}^\top\mathbf{t} - \mathbf{t}^\top\mathbf{X}\hat{\mathbf{w}})$$

Plug in $\widehat{\sigma^2}$ to the log likelihood:

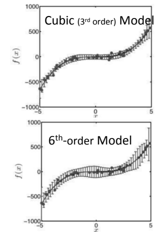
$$\begin{aligned}\log L &= -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\widehat{\sigma^2} - \frac{1}{2\widehat{\sigma^2}}N\widehat{\sigma^2} \\ &= -\frac{N}{2}(1+\log 2\pi) - \frac{N}{2}\log\widehat{\sigma^2}.\end{aligned}$$

Making $\hat{\sigma}^2$ smaller makes log L larger. Making model more flexible decreases $\hat{\sigma}^2$.

Bad news: Increasing the model complexity will *decrease* the variance!

SISTA 26

MLE Prefers Complex Models



$$\log L = -\frac{1}{N}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\top \mathbf{x}_n)^2$$

$$\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t}$$

$$\widehat{\sigma^2} = \frac{1}{N}\sum_{n=1}^{N}(t_n - \mathbf{x}^\top\hat{\mathbf{w}})^2 = \frac{1}{N}(\mathbf{t}^\top\mathbf{t} - \mathbf{t}^\top\mathbf{X}\hat{\mathbf{w}})$$

Plug in $\widehat{\sigma^2}$ to the log likelihood:

$$\begin{aligned}\log L &= -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\widehat{\sigma^2} - \frac{1}{2\widehat{\sigma^2}}N\widehat{\sigma^2} \\ &= -\frac{N}{2}(1+\log 2\pi) - \frac{N}{2}\log\widehat{\sigma^2}.\end{aligned}$$

Making $\hat{\sigma}^2$ smaller makes log L larger. Making model more flexible decreases $\hat{\sigma}^2$.

Bad news: Increasing the model complexity will *decrease* the variance!

Bottom line: Unfortunately, we can't use MLE to do *model selection.* But, with a *particular* model, MLE will choose the parameters that make the data have the highest overall likelihood under the model.

SISTA 27

The Bias Variance Tradeoff

$$Y = f(X) + \epsilon$$

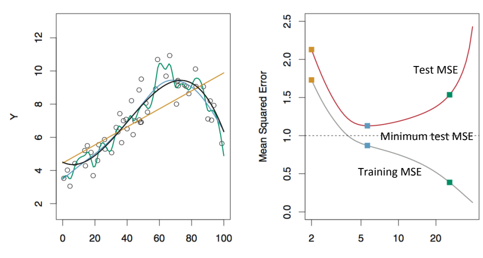


Fig 1: contrasting training with testing error

SISTA 28

The Bias Variance Tradeoff

$$Y = f(X) + \epsilon$$

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}} \end{aligned}$$

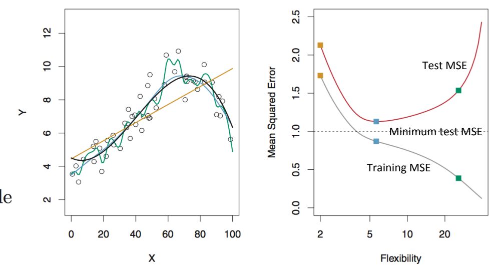


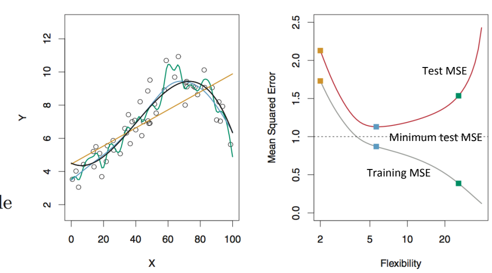
Fig 1: contrasting training with testing error

29

The Bias Variance Tradeoff

$$Y = f(X) + \epsilon$$

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}} \end{aligned}$$



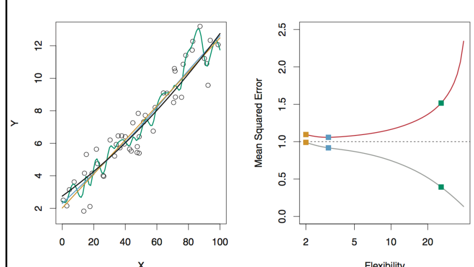


Fig 2: But if data is close to linear, linear fit may do very close to perfect

30

The Bias Variance Tradeoff

$$Y = f(X) + \epsilon$$

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}} \end{aligned}$$

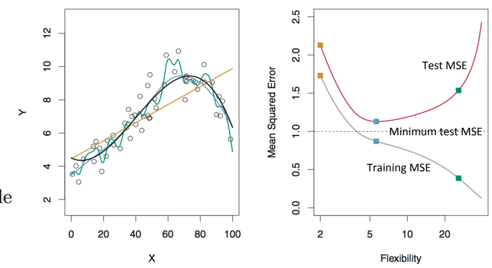
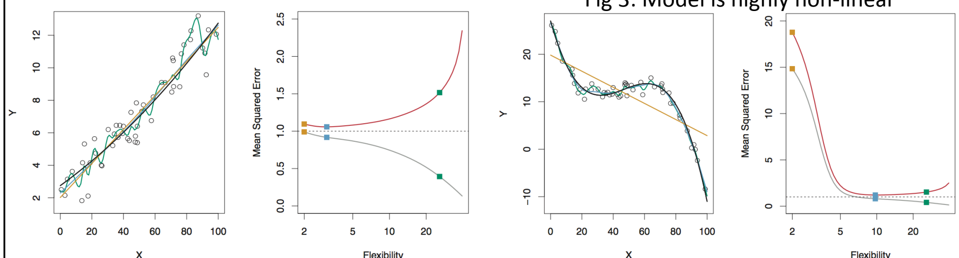


Fig 3: Model is highly non-linear

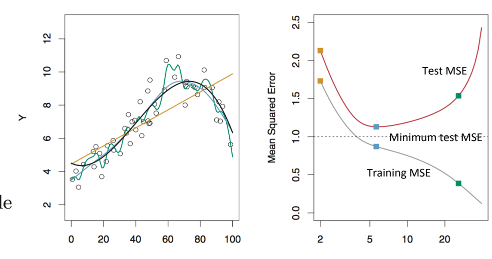


31

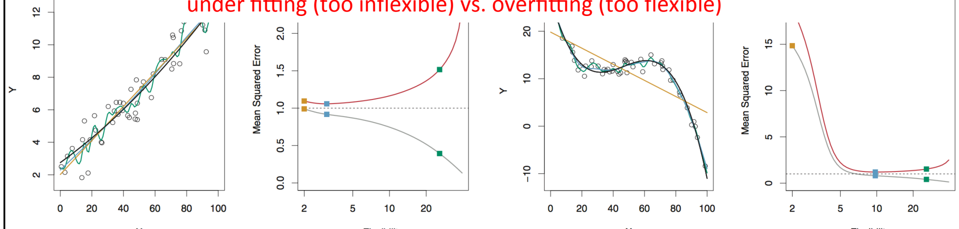
The Bias Variance Tradeoff

$$Y = f(X) + \epsilon$$

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}} \end{aligned}$$



In general: always tend to have a u-shaped curve representing under fitting (too inflexible) vs. overfitting (too flexible)

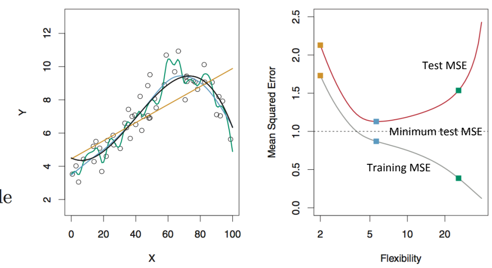


32

The Bias Variance Tradeoff

$$Y = f(X) + \epsilon$$

$$\begin{aligned} E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\ &= \underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}} \end{aligned}$$



$$E\left(y_0-\hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon)$$

Expected Test MSE
(average test MSE if we repeatedly estimated *f* using a large number of training sets)

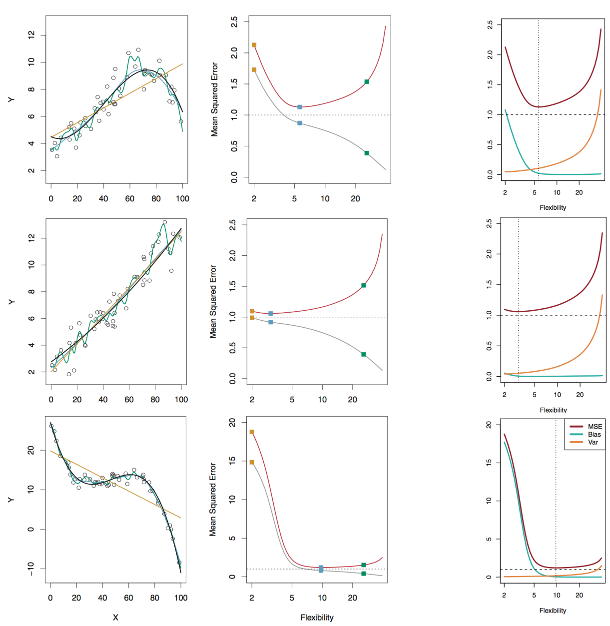
Model variance
(how *f* changes with different sampled data)

Model Bias
(error from inflexibility of the model relative to the true *f*)

Irreducible variance

Bias: the systematic mismatch between our model and the process that generated the data.
Too simple a model == too high a bias (underfitting)
Too complex a model (too many degrees of freedom) == too low a bias (overfitting)
Variance: Squared error between model and data

$$E\left(y_0-\hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon)$$



Blue: Model bias^2
Orange: Model variance
Horizontal dashed: Var($\boldsymbol{\epsilon}$)
Red: test MSE

SISTA 34

STOP