Bias-Variance Tradeoff

Mathematics for Machine Learning

Lecturer: Matthew Wicker

Logistics

- Thank you to all who filled out the survey!
- Our final lecture will be an overview lecture where we connect all of the topics and processes together
- After the final lecture, GTA's will be assigned problem topics to discuss in small groups
- You will get the following materials before next Friday's lecture: Practice Exam, formula sheet, slides containing list of examinable material

Material Covered

Models: Linear models, basis expansion, logistic regression, neural networks, Prob. densities, Bayesian density estimation

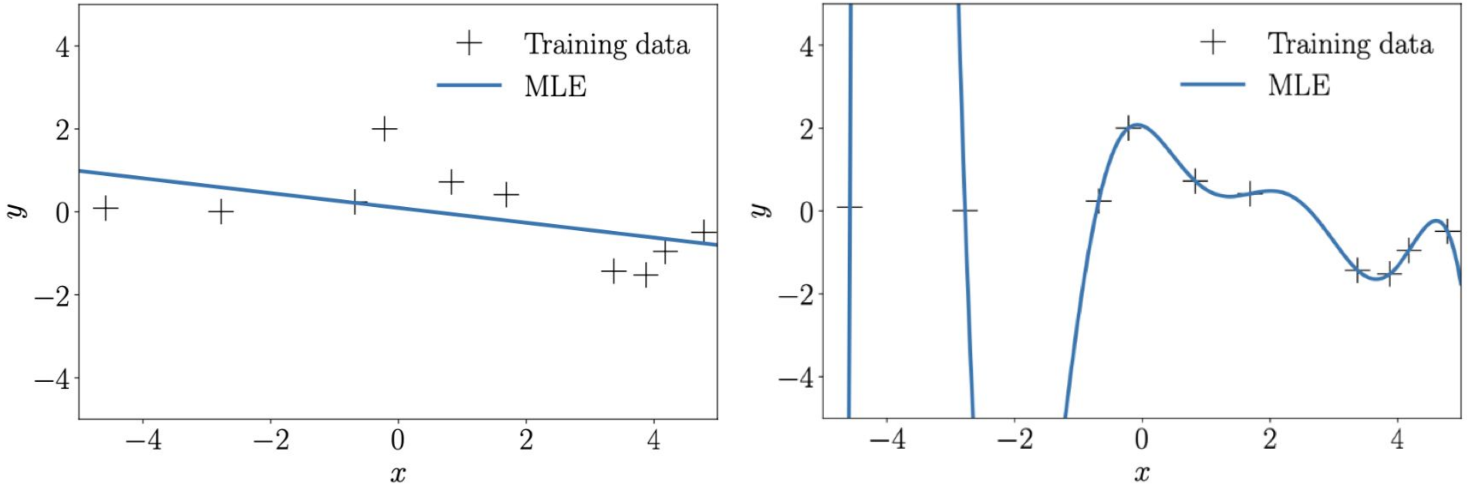
Techniques: Least squares estimation, forward AD, reverse AD, computational graphs, gradient descent, convergence, convexity, Lipschitz continuity, Maximum likelihood, maximum a posteriori, Bayesian inference, LOTUS, change of variables, expectation identities, equating coefficients, joint Gaussian, epistemic/aleatoric uncertainty, concentration inequalities, cross-validation, regularization, generalization error bounds

Settings: Regression, Classification, Density Estimation

This lecture: Estimators, bias of an estimator, variance of an estimator, bias variance trade off

Recall from last lectures

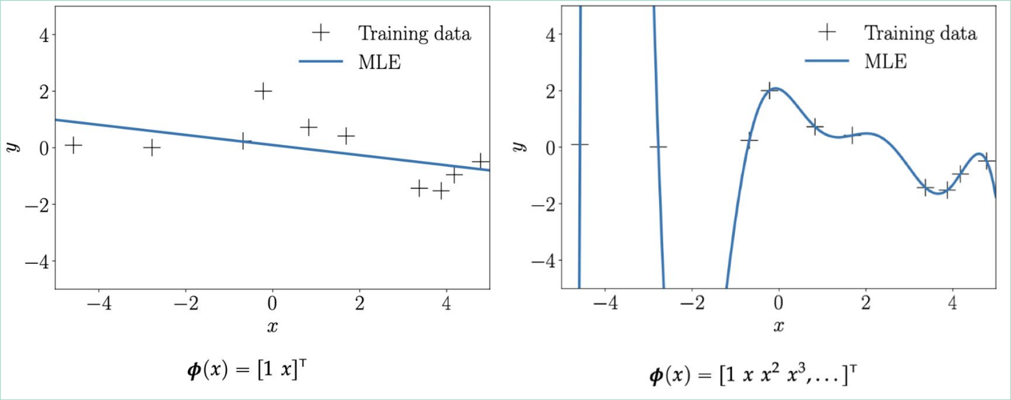
fits better but worse generalisation



$\boldsymbol{\phi}(x) = [1 \ x]^\top$

$\boldsymbol{\phi}(x) = [1 \ x \ x^2 \ x^3, \ldots]^\top$

Recall from last lectures



Model validation ideas:

- Test-set -> generalization error w/concentration
- Generalization error bounds for class of models
- Regularization (implicit/explicit)
- *Cross-validation*

We talk of estimators, let's define them

Definition 2.1. Statistic A statistic S is a random variable that is a function of some data $\mathcal{D}$, $S = g(\mathcal{D})$ where the data $\mathcal{D}$ is a collection of random variables.

An estimator is a statistic that aims at approximating a parameter/property of a distribution.

RV

$\downarrow$

$S = g(D)$

estimator -> S that approxs dist. param.

We talk of estimators, let's define them

Definition 2.1. Statistic A statistic S is a random variable that is a function of some data $\mathcal{D}$, $S = g(\mathcal{D})$ where the data $\mathcal{D}$ is a collection of random variables.

An estimator is a statistic that aims at approximating a parameter/property of a distribution.

sample mean estimator

$$\hat{Z} = \frac{Z_1 + Z_2 + \ldots + Z_N}{N}$$

f^n applied to data

We talk of estimators, let's define them

An estimator is a statistic that aims at approximating a parameter/property of a distribution.

estimator of Z

$$\text{Bias}(\hat{Z}_n) = \mathbb{E}[\hat{Z}_n - Z]$$

Expected diff between estimator and actual statistic

We talk of estimators, let's define them

An estimator is a statistic that aims at approximating a parameter/property of a distribution.

$$\mathrm{Bias}(\hat{Z}_n) = \mathbb{E}[\hat{Z}_n - Z] = 0$$

Unbiased estimator

We proved unbiased estimator for the sample mean

From lecture 10:

$$\hat{Z} = \frac{Z_1 + Z_2 + \ldots + Z_N}{N}$$

$$\mathbb{E}[\hat{Z}] = \frac{\mathbb{E}[Z_1] + \mathbb{E}[Z_2] + \ldots + \mathbb{E}[Z_N]}{N}$$

i.i.d

$$= \frac{\mu_1 + \mu_2 + \ldots \mu_N}{N} = \mu$$

$= \mu + \mu + \mu \ldots$

Sample variance estimators

$$\hat{\sigma}^2_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

$$\hat{\sigma}^{2'}_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2$$

Here we have two sample variance estimators, but which one is the biased estimator and which is unbiased?

Letting: $\mu = \hat{\mu}_n$

Sample variance estimators

Let's check the simpler looking equation for bias

$$\hat{\sigma^2}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

$$\text{bias}(\hat{\sigma^2}_n) = \mathbb{E}[\hat{\sigma^2}_n - \sigma^2]$$

Bias $= E[\hat{S} - S]$

$= E[\hat{S}] - E[S] = E[\hat{S}] - S$

Var : $E[(\hat{S} - E[\hat{S}])^2]$

Err : $E[(\hat{S} - S)^2]$

$= \text{Var} + \text{Bias}^2$

$= E[(\hat{S} - E[\hat{S}])^2] + (E[\hat{S}] - S)^2$

Sample variance estimators

Let's check the simpler looking equation for bias

$$\hat{\sigma^2}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

$$\begin{aligned}\mathrm{bias}(\hat{\sigma^2}_n) &= \mathbb{E}[\hat{\sigma^2}_n - \sigma^2] \\ &= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \sigma^2]\end{aligned}$$

Sample variance estimators

Let's check the simpler looking equation for bias

$$\text{bias}(\hat{\sigma^2}_n) = \mathbb{E}[\hat{\sigma^2}_n - \sigma^2]$$

$$= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \sigma^2]$$

$$= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2] - \sigma^2$$

$E(\sigma^2) = \sigma^2$

Sample variance estimators

Simplifying the expression in the expectation:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2=\frac{1}{n}\sum_{i=1}^{n}(X_i^2-2X_i\mu+\mu^2)$$

Sample variance estimators

Simplifying the expression in the expectation:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2X_i\mu + \mu^2)$$

$$\propto \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} 2X_i\mu + \sum_{i=1}^{n}\mu^2$$

Remember we have to mult by in 1/n at the end

Sample variance estimators

Simplifying the expression in the expectation:

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i^2-2X_i\mu+\mu^2) \\
&\propto \sum_{i=1}^{n}X_i^2-\sum_{i=1}^{n}2X_i\mu+\sum_{i=1}^{n}\mu^2 \\
&= \left(\sum_{i=1}^{n}X_i^2\right)-2\mu\left(\sum_{i=1}^{n}X_i\right)+n\mu^2
\end{aligned}
$$

Remember we have to mult by in 1/n at the end

Sample variance estimators

Simplifying the expression in the expectation:

$$= \left(\sum_{i=1}^{n} X_i^2\right) - 2\mu\left(\sum_{i=1}^{n} X_i\right) + n\mu^2$$

$$= \left(\sum_{i=1}^{n} X_i^2\right) - 2n\mu^2 + n\mu^2$$

$\mu = \frac{\sum x_i}{n}$

$\hookrightarrow \sum x_i = n\mu$

Remember we have to mult by in 1/n at the end

Sample variance estimators

Simplifying the expression in the expectation:

$$= \left(\sum_{i=1}^{n} X_i^2\right) - 2\mu\left(\sum_{i=1}^{n} X_i\right) + n\mu^2$$

$$= \left(\sum_{i=1}^{n} X_i^2\right) - 2n\mu^2 + n\mu^2$$

$$= \left(\sum_{i=1}^{n} X_i^2\right) - n\mu^2$$

Remember we have to mult by in 1/n at the end

Sample variance estimators

Simplifying the expression in the expectation:

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} X_i^2\right) - n\mu^2\right] = \left(\sum_{i=1}^{n} \mathbb{E}[X_i^2]\right) - \mathbb{E}[n\mu^2]$$

Expectation over data (X_i)

Remember we have to mult by in 1/n at the end

Sample variance estimators

Simplifying the expression in the expectation:

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} X_i^2\right) - n\mu^2\right] = \left(\sum_{i=1}^{n} \mathbb{E}[X_i^2]\right) - \mathbb{E}[n\mu^2]$$

$$\sigma^2 = E[x^2] - \underbrace{E[x]^2}_{\mu^2}$$

Aside, identities we will prove later:

$$\mathbb{E}[X_i^2] = \sigma^2 + \mu^2$$

Hint: this is just rearranging the variance definition

$$\mathbb{E}[\mu^2] = \frac{\sigma^2}{n} + \mu^2$$

Sample variance estimators

Simplifying the expression in the expectation:

$$= \left(\sum_{i=1}^{n} \mathbb{E}[X_i^2]\right) - \mathbb{E}[n\mu^2]$$

$$= \sum_{i=1}^{n} (\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)$$

$$= n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2$$

$$= (n-1)\sigma^2$$

Remember we have to mult by in 1/n at the end

Sample variance estimators

Simplifying the expression in the expectation:

$$= \left(\sum_{i=1}^{n} \mathbb{E}[X_i^2]\right) - \mathbb{E}[n\mu^2]$$

$$= n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2$$

$$= (n-1)\sigma^2$$

Remember we have to mult by in 1/n at the end

$$\text{bias}(\hat{\sigma}^2{}_n) = \frac{n-1}{n}\sigma^2 - \sigma^2$$

Sample variance estimators

$$\hat{\sigma^2}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

By plugging in our estimator into the definition of bias, making the i.i.d. assumption and being comfortable with manipulating expectation identities, we showed the bias is non-zero:

$$\mathrm{bias}(\hat{\sigma^2}_n) = \frac{n-1}{n}\sigma^2 - \sigma^2$$

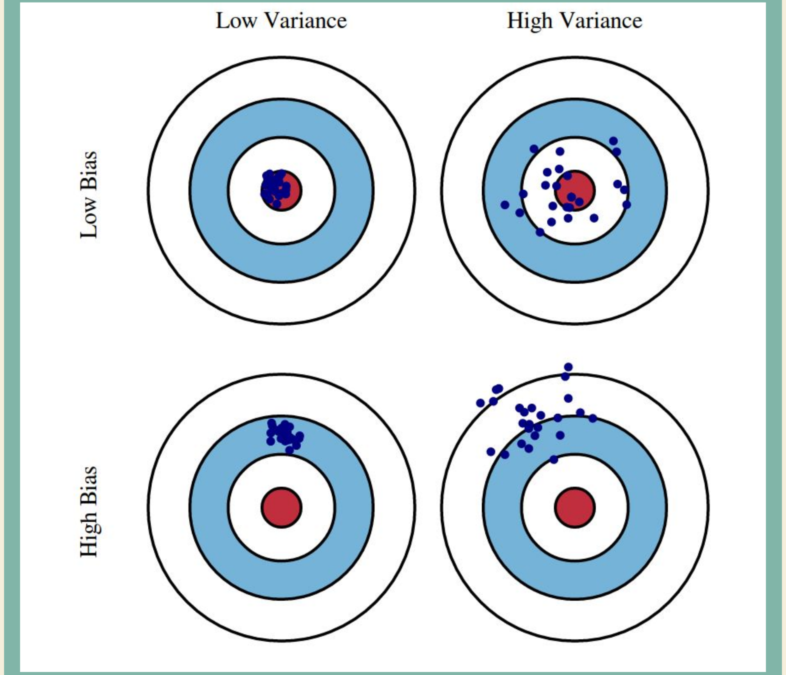
Variance of an estimator

Definition 2.1. Statistic A statistic S is a random variable that is a function of some data $\mathcal{D}$, $S = g(\mathcal{D})$ where the data $\mathcal{D}$ is a collection of random variables.

An estimator is a statistic that aims at approximating a parameter/property of a distribution.

$$\mathrm{Var}(\hat{S}) = \mathbb{E}\left[(\hat{S} - \mathbb{E}(\hat{S}))^2\right]$$

Visualization of bias-variance



Bias-variance decomposition

$$\mathrm{Err}(\hat{S}) = \mathbb{E}[(\hat{S} - S)^2] = \mathrm{Bias}^2 + \mathrm{Var}$$

$E[\hat{S}-S] + E[(\hat{S}-E[\hat{S}])^2]$

error of $\hat{S}$ is Bias^2 + variance

The bias-variance decomposition simply tells us that the (mean squared error) of an estimator is completely accounted for by the sum of the squared bias and variance of the estimator.

Bias-variance decomposition

$$\mathrm{Err}(\hat{S}) = \mathbb{E}[(\hat{S} - S)^2] = \mathrm{Bias}^2 + \mathrm{Var}$$

$$= \mathbb{E}[(g(\mathcal{D}) - S)^2]$$

only works for MSE of estimator

The bias-variance decomposition simply tells us that the *(mean squared error)* of an estimator is completely accounted for by the sum of the squared bias and variance of the estimator.

bias: $E[\hat{S}-S]$ var: $E[(E[\hat{S}]-\hat{S})^2]$

Bias-variance proof

$$\mathbb{E}[(\hat{S} - S)^2] = \mathbb{E}\left[\left(\hat{S} - \mu + \mu - S\right)^2\right]$$

$\pm \mu$

$$= \mathbb{E}\left[\left((\hat{S} - \mu) + (\mu - S)\right)^2\right]$$

$= E[(\hat{S} - E[\hat{S}])^2] + E[(E[\hat{S}] - S)^2]$

Letting: $\mu = \mathbb{E}[\hat{S}]$

Bias-variance proof

$$\begin{aligned}\mathbb{E}[(\hat{S}-S)^2] &= \mathbb{E}\left[\left(\hat{S}-\mu+\mu-S\right)^2\right] \\ &= \mathbb{E}\left[\left((\hat{S}-\mu)+(\mu-S)\right)^2\right] \\ &= \mathbb{E}\left[(\hat{S}-\mu)^2+2(\hat{S}-\mu)(\mu-S)+(\mu-S)^2\right]\end{aligned}$$

Bias-variance proof

$$
\begin{aligned}
\mathbb{E}[(\hat{S}-S)^2] &= \mathbb{E}\left[\left(\hat{S}-\mu+\mu-S\right)^2\right] \\
&= \mathbb{E}\left[\left((\hat{S}-\mu)+(\mu-S)\right)^2\right] \\
&= \mathbb{E}\left[(\hat{S}-\mu)^2+2(\hat{S}-\mu)(\mu-S)+(\mu-S)^2\right] \\
&= \mathbb{E}\left[(\hat{S}-\mu)^2+(\mu-S)^2\right] \qquad (\star)
\end{aligned}
$$

$\mu = E[\hat{S}]$

$E[E[\hat{S}]] = \hat{S}$

$E[(\hat{S}-\mu)(\mu-S)] = E[\hat{S}\mu - \hat{S}S - \mu\mu + \mu S] = E[\hat{S}E[S]] - E[\hat{S}S] - E[\mu^2] + E[\mu S]$

$= E[S]E[\hat{S}] - E[S]E[\hat{S}] - \mu^2 + \mu \underbrace{E[S]}_{=\mu}$

$= 0$

Bias-variance proof

bias: $E[\hat{S}-S]$

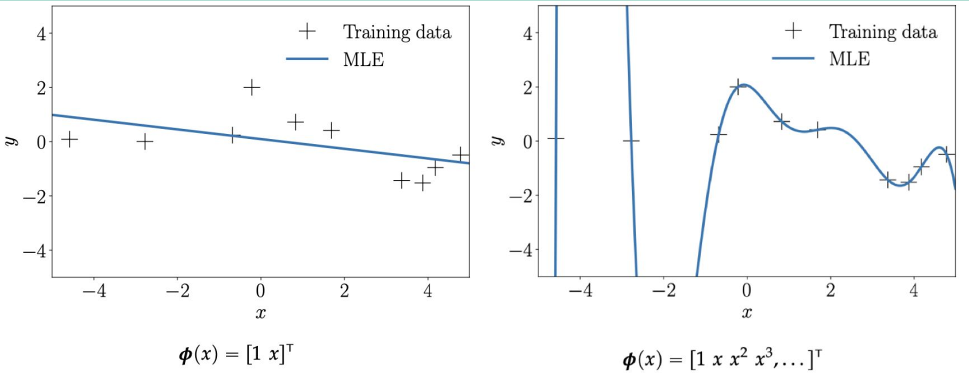
$= E[\hat{S}] - E[S]$

$= \mu - S$

$$\begin{aligned}
\mathbb{E}[(\hat{S}-S)^2] &= \mathbb{E}\left[\left(\hat{S}-\mu+\mu-S\right)^2\right] \\
&= \mathbb{E}\left[\left((\hat{S}-\mu)+(\mu-S)\right)^2\right] \\
&= \mathbb{E}\left[(\hat{S}-\mu)^2 + 2(\hat{S}-\mu)(\mu-S) + (\mu-S)^2\right] \\
&= \mathbb{E}\left[(\hat{S}-\mu)^2 + (\mu-S)^2\right] \qquad (\star) \\
&= \mathbb{E}\left[(\hat{S}-\overset{E[\hat{S}]}{\mu})^2\right] + \mathbb{E}\left[(\overset{E[\hat{S}]}{\mu}-S)^2\right] \\
&= \mathbb{E}\left[(\hat{S}-\mu)^2\right] + (\mu-S)^2 = \text{var.} + \text{bias}^2
\end{aligned}$$

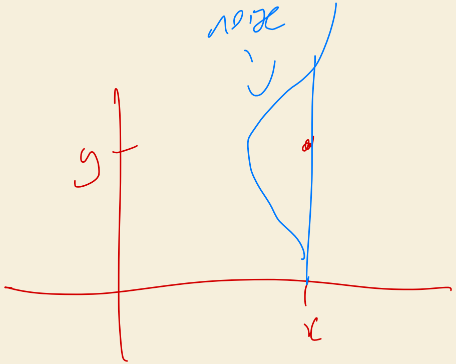
Bias-Variance in ML

Our course notes discuss the particulars of bias-variance decomposition in linear models, which you should study (in order to answer exercises), but in lecture we will be more general/conceptual.



https://www.cs.cornell.edu/courses/cs4780/2022sp/notes/LectureNotes17.html

Bias-Variance in ML



$$\hat{y} = \int_y P(y|\mathbf{x})dy$$

Consider the true label of our machine learning models to be the integral over the likelihood

Bias-Variance in ML

$$\hat{y} = \int_y y_x P(y|\mathbf{x})dy$$

We can write the error of our model as the integral over both the output and input space of our mean squared error:

$$\mathbb{E}_{P(\mathbf{x},y)}[\underbrace{(f^\theta(\mathbf{x}) - y)^2}_{\text{MSE error}}] = \int_x \int_y (\underset{\overset{\uparrow}{\text{model}}}{f^\theta}(\mathbf{x}) - y)^2 P(\mathbf{x}, y)dyd\mathbf{x}$$

Bias-Variance in ML

$$\mathbb{E}_{P(\mathbf{x},y)}[(f^\theta(\mathbf{x}) - y)^2] = \int_x \int_y (\overset{\hat{S}}{f^\theta(\mathbf{x})} \overset{-S}{-\, y})^2 P(\mathbf{x}, y)\,dy\,d\mathbf{x}$$

need $E[\hat{S}]$

When we compute the bias and variance of an estimator (in this case the estimator of our risk of a model), we need to do so with respect to the "average model"

$$\hat{f}^\theta = \underbrace{\mathbb{E}_{\mathcal{D}\sim P(\mathbf{x},y)}}_{\text{avg. model given datasets}} \underbrace{\int_{\mathcal{D}} f^\theta_{\mathcal{D}} P(\mathcal{D})\,d\mathcal{D}}_{\text{integrate over all potential datasets}}$$

Bias-Variance in ML

Now, we have defined all of the components we need to quantify the bias and variance of our model error estimator:

$$\mathbb{E}_{\mathbf{x},y,\mathcal{D}}\left[\left(f^{\theta}_{\mathcal{D}}(\mathbf{x}) - y\right)^2\right] \quad \text{MSE}$$

$$E[\hat{s} - s] \rightarrow \underbrace{E[(\hat{s} - E(\hat{s}))^2]}_{\text{var}} + \underbrace{E[(E(\hat{s}) - s)^2]}_{\text{bias}^2}$$

$$\rightarrow E[(f^{\theta}_{D}(x) - \hat{f}^{\theta}(x))^2] + E[(\hat{f}^{\theta}(x) - y)^2]$$

Bias-Variance in ML

$$\mathbb{E}_{\mathbf{x},y,\mathcal{D}}\left[\left(f^{\theta}_{\mathcal{D}}(\mathbf{x}) - y\right)^2\right]$$

$\text{Err} = E[(\hat{s} - s)^2] = \text{var} + \text{bias}^2 = E[(\hat{s} - E[\hat{s}])^2] + E[(E[\hat{s}] - s)^2]$

$= (E[\hat{s}] - s)^2$

As before, we consider subtracting and adding the mean from the estimated quantity:

$$= \mathbb{E}_{\mathbf{x},y,\mathcal{D}}\left[\left(\left(f^{\theta}_{\mathcal{D}}(\mathbf{x}) - \hat{f}^{\theta}(\mathbf{x})\right) + \left(\hat{f}^{\theta}(\mathbf{x}) - y\right)\right)^2\right]$$

add / subtract avg model

$\hat{s} - E[s]$

Bias-Variance in ML

By the same cancellation logic, we get that the middle terms of completing the square drop off, so:

$$= \mathbb{E}_{\mathbf{x},y,\mathcal{D}}\left[\left((f^{\theta}_{\mathcal{D}}(\mathbf{x}) - \hat{f}^{\theta}(\mathbf{x}))^2 + (\hat{f}^{\theta}(\mathbf{x}) - y)^2\right)\right]$$

And then by linearity of expectation we have the definition of bias-variance again:

$$\mathbb{E}_{\mathbf{x},\mathcal{D}}\left[(f^{\theta}_{\mathcal{D}}(\mathbf{x}) - \hat{f}^{\theta}(\mathbf{x}))^2\right] + \mathbb{E}_{\mathbf{x},y}[(\hat{f}^{\theta}(\mathbf{x}) - y)^2]$$

how much model deviates around avg model w/ diff datasets (variance of model)

risk of avg classifier

Bias-Variance in ML

$$\mathbb{E}_{\mathbf{x},\mathcal{D}}\left[\left(f_{\mathcal{D}}^{\theta}(\mathbf{x}) - \hat{f}^{\theta}(\mathbf{x})\right)^2\right] + \underbrace{\mathbb{E}_{\mathbf{x},y}[(\hat{f}^{\theta}(\mathbf{x}) - y)^2]}$$

Notice! Another MSE term here, this time the risk of our average model. We can repeat our bias-variance steps to decompose this further

Bias-Variance in ML

Again, we can add and subtract the estimator for y to further decompose this term:

$$\mathbb{E}_{\mathbf{x},\mathcal{D}}\left[\left(f^{\theta}_{\mathcal{D}}(\mathbf{x}) - \hat{f}^{\theta}(\mathbf{x})\right)^2\right] + \underbrace{\mathbb{E}_{\mathbf{x},y}[(\hat{f}^{\theta}(\mathbf{x}) - y)^2]}$$

$$\mathbb{E}_{\mathbf{x},y}[(\hat{f}^{\theta}(\mathbf{x}) - y)^2] = \mathbb{E}_{\mathbf{x},y}\left[(\hat{f}^{\theta}(\mathbf{x}) - \hat{y} + \hat{y} - y)^2\right]$$

Bias-Variance in ML

Again, we can add and subtract the estimator for y to further decompose this term:

$$\mathbb{E}_{\mathbf{x},\mathcal{D}}\left[(f_{\mathcal{D}}^{\theta}(\mathbf{x}) - \hat{f}^{\theta}(\mathbf{x}))^2\right] + \underbrace{\mathbb{E}_{\mathbf{x},y}[(\hat{f}^{\theta}(\mathbf{x}) - y)^2]}$$

$$\begin{aligned}
\mathbb{E}_{\mathbf{x},y}[(\hat{f}^{\theta}(\mathbf{x}) - y)^2] &= \mathbb{E}_{\mathbf{x},y}\left[(\hat{f}^{\theta}(\mathbf{x}) - \hat{y} + \hat{y} - y)^2\right] \\
&= \underbrace{\mathbb{E}_{\mathbf{x},y}[(\hat{y} - y)^2]}_{\text{noise}} + \underbrace{\mathbb{E}_{\mathbf{x}}[(\hat{f}^{\theta}(\mathbf{x}) - \hat{y})^2]}_{\text{bias}^2}
\end{aligned}$$

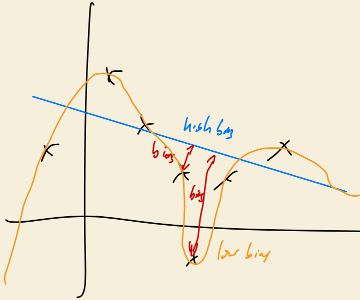
Final bias-variance decomposition

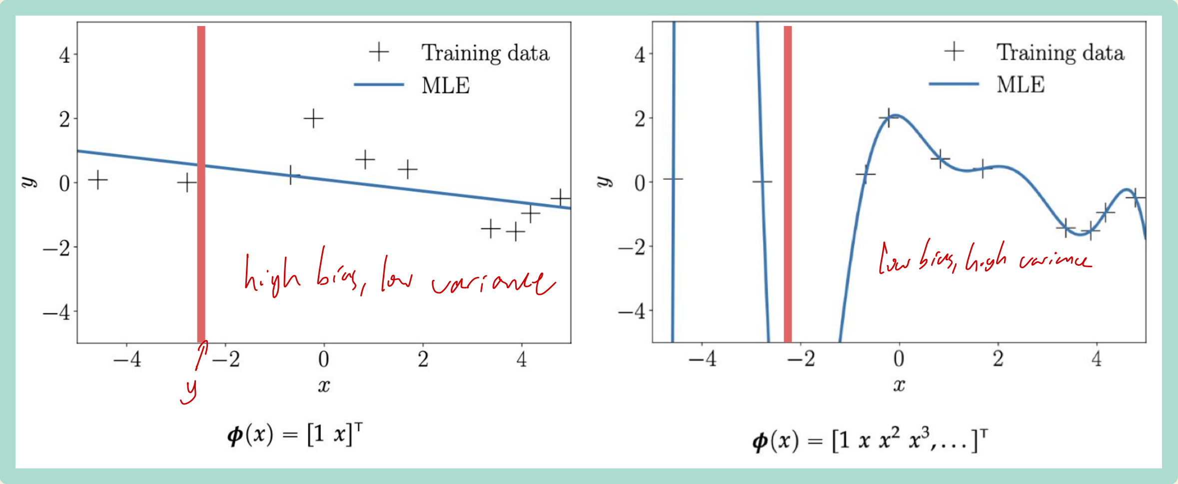
$$\mathbb{E}_{\mathbf{x},y,\mathcal{D}}\left[\left(f_{\mathcal{D}}^{\theta}(\mathbf{x}) - y\right)^2\right]$$

We decomposed the error into three terms:

$$\underbrace{\mathbb{E}_{\mathbf{x},y}[(\hat{y} - y)^2]}_{\text{Noise}} + \underbrace{\mathbb{E}_{\mathbf{x}}\ [(\hat{f}^{\theta}(\mathbf{x}) - \hat{y})^2]}_{\text{Squared bias}} + \underbrace{\mathbb{E}_{\mathbf{x},\mathcal{D}}[(\hat{f}_{\mathcal{D}}^{\theta}(\mathbf{x}) - \hat{f}^{\theta}(\mathbf{x}))^2]}_{\text{Variance}}$$

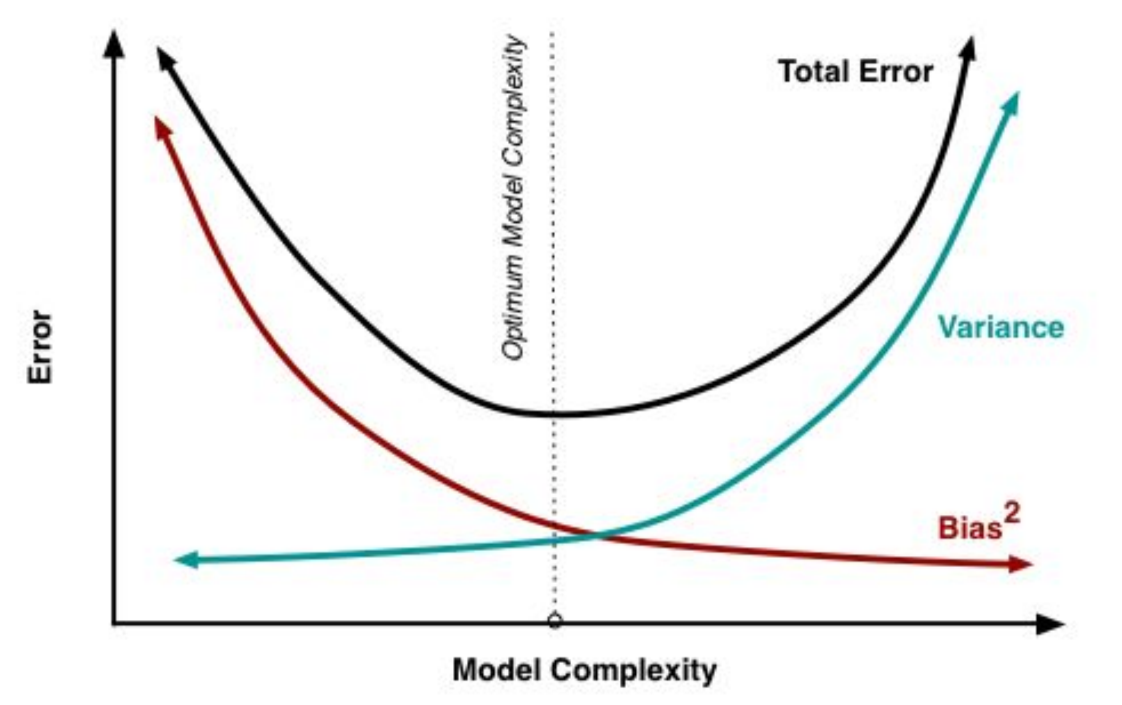
Visualizing this decomposition





$$\underbrace{\mathbb{E}_{\mathbf{x},y}[(\hat{y}-y)^2]}_{\text{Noise}} + \underbrace{\mathbb{E}_{\mathbf{x}}[(\hat{f}^{\theta}(\mathbf{x})-\hat{y})^2]}_{\text{Squared bias}} + \underbrace{\mathbb{E}_{\mathbf{x},\mathcal{D}}[(\hat{f}^{\theta}_{\mathcal{D}}(\mathbf{x})-\hat{f}^{\theta}(\mathbf{x}))^2]}_{\text{Variance}}$$

Model selection



Next lecture:
Principal Component-Analysis