Lecture 9: Model Selection

Machine Learning 1

TU Berlin – WiSe 2016/2017

Outline

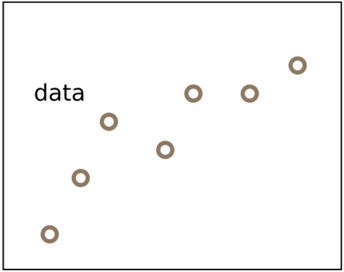
This week:

- Basics
  - Occam's razor
  - Model complexity
- Statistical Learning Theory (1)
  - Bias-variance decomposition
- Validation Techniques
  - Hold-out, k-fold validation

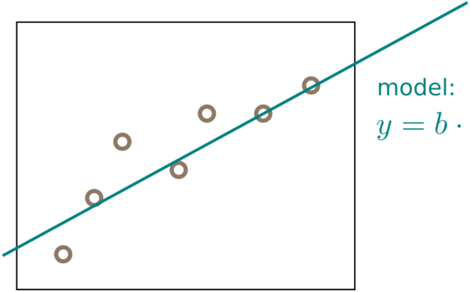
Next week:

- Statistical Learning Theory (2)
  - Bounds on generalization error, VC dimension
- Kernels
  - Kernel trick, induced feature spaces

Example: Predicting the Future



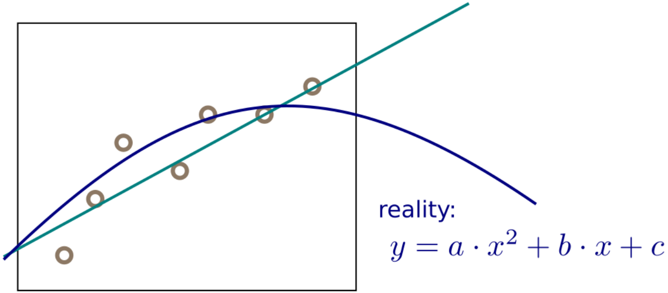
Example: Predicting the Future



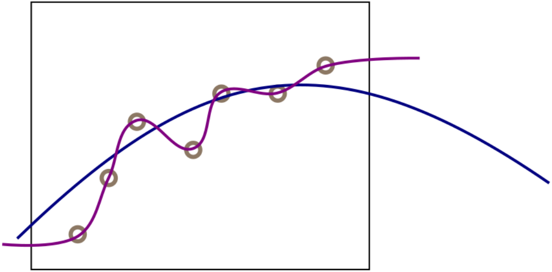
model:

$y = b \cdot x + c$

Example: Predicting the Future



Example: Predicting the Future



more complex model:

$$y = a_7x^7 + a_6x^6 + a_5x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x^1 + a_0$$

Occam's Razor

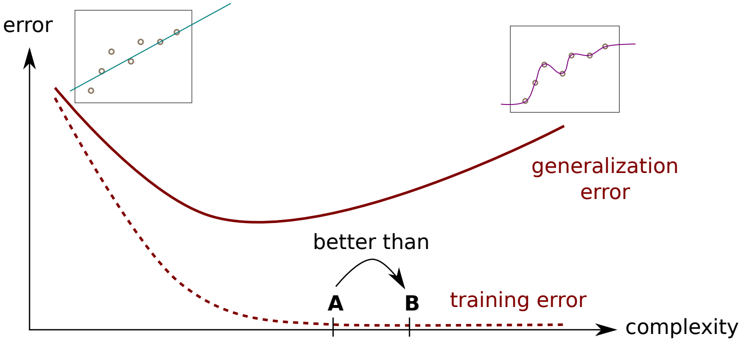
"Among competing hypotheses, the one with the fewest assumptions should be selected."

Domingos (1998): "Occam's two Razors: The Sharp and the Blunt" lists **two common interpretations** of it in a ML setting:

- **1st Razor:** *"Given two models with the same generalization error, the simpler one should be preferred because simplicity is desirable in itself."*
- **2nd Razor:** *"Given two model with the same training-set error the simpler one should be preferred because it is likely to have lower generalization error."*

(and warns against a too literal use of the second interpretation).

Occam's (2nd) Razor in ML



- A too simple model causes "underfitting".
- A too complex model causes "overfitting".
- **Question:** How to define model complexity

Some possible measures of model complexity

Number of parameters of the model

- $f(\boldsymbol{x}) = c$ has 1 parameters.
- $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + c$ has $(d+1)$ parameters.
- $f(\boldsymbol{x}) = \boldsymbol{x}^\top A \boldsymbol{x} + \boldsymbol{w}^\top \boldsymbol{x} + c$ has (d^2+d+1) parameters.

Some possible measures of model complexity

Number of parameters of the model

- $f(\boldsymbol{x}) = c$ has 1 parameters.
- $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + c$ has $(d+1)$ parameters.
- $f(\boldsymbol{x}) = \boldsymbol{x}^\top A \boldsymbol{x} + \boldsymbol{w}^\top \boldsymbol{x} + c$ has (d^2+d+1) parameters.

Choice of variables the model receives as input

- Feature selection.
- PCA dimensionality reduction.

Some possible measures of model complexity

Number of parameters of the model

- $f(\boldsymbol{x}) = c$ has 1 parameters.
- $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + c$ has $(d+1)$ parameters.
- $f(\boldsymbol{x}) = \boldsymbol{x}^\top A \boldsymbol{x} + \boldsymbol{w}^\top \boldsymbol{x} + c$ has $(d^2 + d + 1)$ parameters.

Choice of variables the model receives as input

- Feature selection.
- PCA dimensionality reduction.

Properties of the function

- e.g. continuity, slope.

Some possible measures of model complexity

Number of parameters of the model

- $f(\boldsymbol{x}) = c$ has 1 parameters.
- $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + c$ has $(d+1)$ parameters.
- $f(\boldsymbol{x}) = \boldsymbol{x}^\top A \boldsymbol{x} + \boldsymbol{w}^\top \boldsymbol{x} + c$ has $(d^2 + d + 1)$ parameters.

Choice of variables the model receives as input

- Feature selection.
- PCA dimensionality reduction.

Properties of the function

- e.g. continuity, slope.

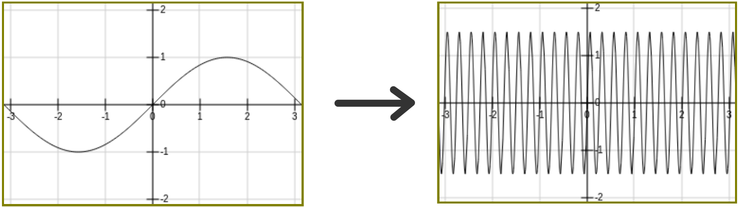
VC-dimension

- Provide a bound on generalization error!

A Second Look at Occam's Razor

> *"Given two model with the same training-set error the simpler one should be preferred because it is likely to have lower generalization error."*

Counter-example for "simplicity = few parameters": The two-parameters model $f(x) = a\sin(\omega x)$ can fit almost *any* dataset in $\mathbb{R}$.



I.e. $(a = 1, \omega = 21833.5)$ is "simple" (only 2 numbers), but does not lead to low generalization error.

From Occam's Razor to Prediction

"Given two model with the same training-set error the simpler one should be preferred because it is likely to have lower generalization error."

Problem: what does "simple" or "intuitive" mean?

From Occam's Razor to Prediction

> *"Given two model with the same training-set error the simpler one should be preferred because it is likely to have lower generalization error."*

Problem: what does "simple" or "intuitive" mean?

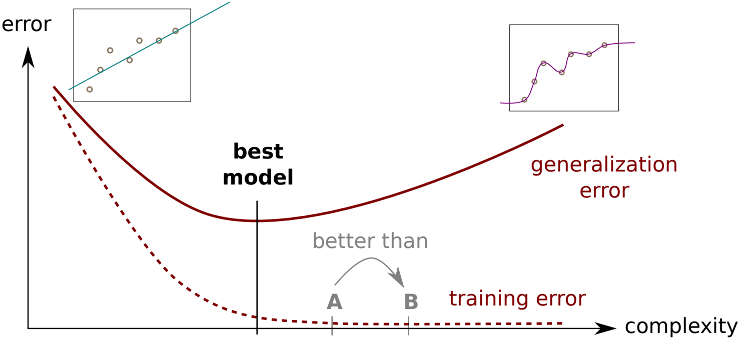
Falsifiability/prediction strength (S. Hawking, after K. Popper)

> *"[a good model] must accurately describe a large class of observations on the basis of a model that contains only a few arbitrary elements, and it must make definite predictions about the results of future observations."*

means: The model with lowest generalization error is preferable.

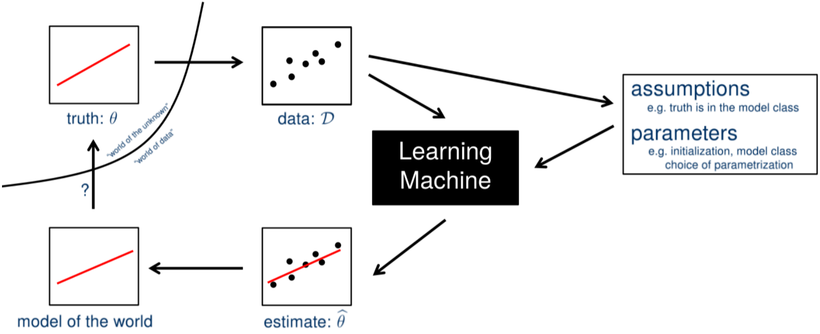
From Occam's Razor to Prediction

> *"[a good model] must accurately describe a large class of observations on the basis of a model that contains only a few arbitrary elements, and it must make definite predictions about the results of future observations."*



Statistical Learning Theory

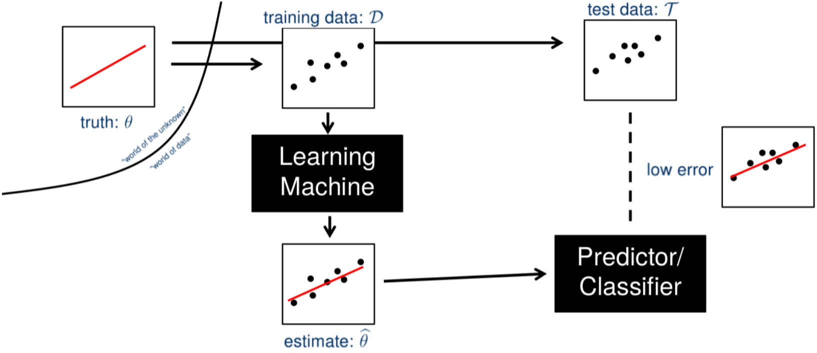
What is a Learning Machine?



Learning Machine $=$ function $\mathcal{D} \mapsto \widehat{\theta}$

Statistical Learning Theory

What is a good Learning Machine?



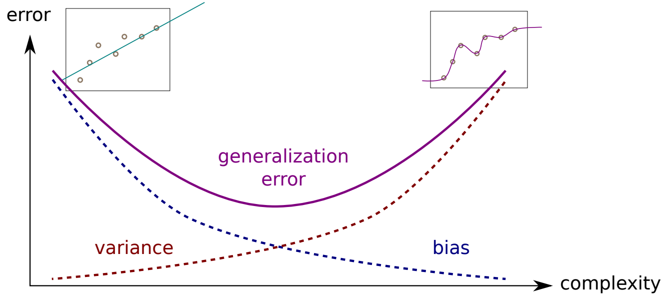
$$\mathrm{MSE}(\widehat{\theta}) = \mathbb{E}\left[(\text{predicted data - test data})^2\right]$$

Statistical Learning Theory

Let $\mathcal{D}$ and $\hat{\theta}$ be *random variables*. For several error measures (e.g. mean square error, KL divergence), generalization error can be decomposed as follows:

$$\boxed{\textbf{Error}(f_{\hat{\theta}}) = \textbf{Bias}(f_{\hat{\theta}}) + \textbf{Variance}(f_{\hat{\theta}})}$$

Bias and variance contribute in different proportions to the error depending on the complexity:



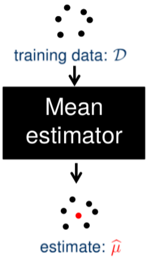
Example: Parameters of a Gaussian

parametric estimation:

θ is a value in $\mathbb{C}^n$ (e.g. $\theta = (\mu, \Sigma)$ for Gaussians)

$\widehat{\theta}$ is function in the data $\mathcal{D} = \{X_1, \ldots, X_N\}$

(X_i are random variables giving back data points)



e.g. mean estimator $$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

covariance estimator $$\widehat{\Sigma} = \frac{1}{N-1}(X_i - \widehat{\mu})(X_i - \widehat{\mu})^\top$$

Here, the learning machine is a simple estimator (can be computed analytically).

Bias, Variance, and MSE of an Estimator

parametric estimation:

θ is a value in $\mathbb{C}^n$ (e.g. $\theta = (\mu, \Sigma)$ for Gaussians)

$\widehat{\theta}$ is function in the data $\mathcal{D} = \{X_1, \ldots, X_N\}$

(X_i are random variables giving back data points)

bias of $\widehat{\theta}$: $\text{Bias}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta} - \theta]$

measures expected deviation of the mean

variance of $\widehat{\theta}$: $\text{Var}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^2\right]$

measures scatter around estimator mean

MSE of $\widehat{\theta}$: $\text{MSE}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \theta)^2\right]$

measures prediction error

Note: for $\theta \in \mathbb{C}^n$, we use the notation $\theta^2 = \theta^\top \theta$.

Bias-Variance Analysis for the Mean Estimator

$$\mathsf{Bias}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta} - \theta] \qquad \mathsf{Var}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^2\right] \qquad \mathsf{MSE}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \theta)^2\right]$$

$$\mathsf{MSE}(\widehat{\theta}) = \mathsf{Bias}(\widehat{\theta})^2 + \mathsf{Var}(\widehat{\theta})$$

Example: estimation of mean

$X_1, \ldots, X_N$ i.i.d Gaussian $\sim \mathcal{N}(\mu, \sigma)$

$$\widehat{\mu} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

"natural" estimator

$$\mathsf{Bias}(\widehat{\mu}) = 0 \qquad \mathsf{MSE}(\widehat{\mu}) = \mathsf{Var}(\widehat{\mu}) = \frac{\sigma^2}{N}$$

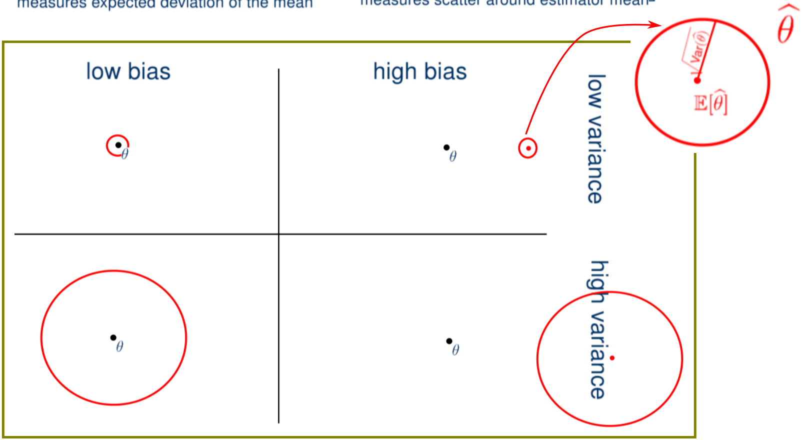
Visualizing Bias and Variance

$$\text{Bias}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta} - \theta]$$

measures expected deviation of the mean

$$\text{Var}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^2\right]$$

measures scatter around estimator mean



Bias-Variance Decomposition of the MSE

$$\text{Bias}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta} - \theta]$$

measures expected deviation of the mean

$$\text{Var}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^2\right]$$

measures scatter around estimator mean

$$\text{MSE}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \theta)^2\right]$$

measures prediction error

Bias-Variance Decomposition of the MSE

$$\text{Bias}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta} - \theta]$$
measures expected deviation of the mean

$$\text{Var}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^2\right]$$
measures scatter around estimator mean

$$\text{MSE}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \theta)^2\right]$$
measures prediction error

Proposition: $\text{MSE}(\widehat{\theta}) = \text{Bias}(\widehat{\theta})^2 + \text{Var}(\widehat{\theta})$

proof:
$$\begin{aligned}
\text{MSE}(\widehat{\theta}) &= \mathbb{E}\left[(\widehat{\theta} - \theta)^2\right] = \mathbb{E}[\widehat{\theta}^2] - 2\theta\mathbb{E}[\widehat{\theta}] + \theta^2 \\
&= \mathbb{E}[\widehat{\theta}^2] - 2\left(\mathbb{E}[\widehat{\theta}]\right)^2 + \left(\mathbb{E}[\widehat{\theta}]\right)^2 + \left(\mathbb{E}[\widehat{\theta}]\right)^2 - 2\theta\mathbb{E}[\widehat{\theta}] + \theta^2 \\
&= \mathbb{E}[\widehat{\theta}^2] - 2\mathbb{E}\left[\widehat{\theta}\left(\mathbb{E}[\widehat{\theta}]\right)\right] + \mathbb{E}\left(\mathbb{E}[\widehat{\theta}]\right)^2 + \left(\mathbb{E}[\widehat{\theta}] - \theta\right)^2 \\
&= \mathbb{E}\left[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^2\right] + \left(\mathbb{E}[\widehat{\theta}] - \theta\right)^2 \\
&= \text{Var}(\widehat{\theta}) + \text{Bias}(\widehat{\theta})^2
\end{aligned}$$

Bias-Variance Analysis for the James-Stein Estimator

$\text{Bias}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta} - \theta]$ $\quad$ $\text{Var}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^2\right]$ $\quad$ $\text{MSE}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \theta)^2\right]$

$$\text{MSE}(\widehat{\theta}) = \text{Bias}(\widehat{\theta})^2 + \text{Var}(\widehat{\theta})$$

Example: estimation of mean

$X_1, \ldots, X_N \in \mathbb{R}^n, n > 2$ i.i.d Gaussian $\sim \mathcal{N}(\mu, \sigma \cdot I)$

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

"natural" estimator

$$\text{Bias}(\widehat{\mu}) = 0$$

$$\text{MSE}(\widehat{\mu}) = \text{Var}(\widehat{\mu}) = \frac{\sigma^2}{N}$$

$$\widehat{\mu}_{\text{JS}} = \widehat{\mu} - \frac{(n-2)\sigma^2}{\widehat{\mu}^2}\widehat{\mu}$$

"shrinkage"

James-Stein estimator

$$\text{Bias}(\widehat{\mu}_{\text{JS}}) > 0$$

$$\text{MSE}(\widehat{\mu}_{\text{JS}}) < \text{MSE}(\widehat{\mu})$$

Comparing the Mean and James-Stein Estimator

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

"natural" estimator

$\text{Bias}(\widehat{\mu}) = 0$

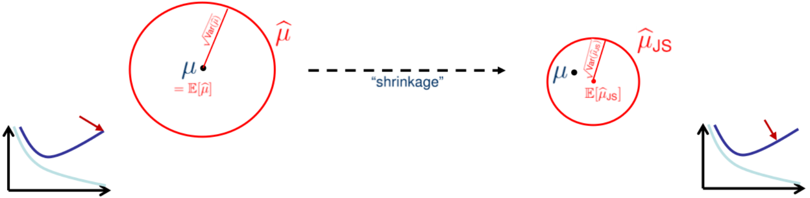
$\text{MSE}(\widehat{\mu}) = \text{Var}(\widehat{\mu}) = \frac{\sigma^2}{N}$

$$\widehat{\mu}_{\text{JS}} = \widehat{\mu} - \frac{(n-2)\sigma^2}{\widehat{\mu}^2}\widehat{\mu}$$

James-Stein estimator

$\text{Bias}(\widehat{\mu}_{\text{JS}}) > 0$

$\text{MSE}(\widehat{\mu}_{\text{JS}}) < \text{MSE}(\widehat{\mu})$



Estimator of Functions

supervised learning:

training data $\mathcal{D}$ is $X_1, \ldots, X_N$ with labels $Y_1, \ldots, Y_N$

(e.g. in regression, $X_i \in \mathbb{R}^n, Y_i \in \mathbb{R}$)

parameter θ "is" a generative function $f = f_\theta$:

$$Y_i = f(X_i) + \varepsilon_i$$

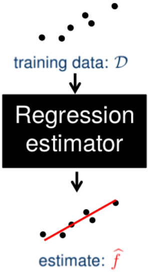
ε_i is error with $\mathbb{E}[\varepsilon_i] = 0$

Learning Machine learns approximation $\widehat{f} = f_{\widehat{\theta}}$

such that $Y_i \approx \widehat{f}(X_i)$

Example (Linear Regression):

$$f(x) = \beta^\top x + \alpha, \quad \theta = (\alpha, \beta)$$



Bias-Variance Analysis of the Function Estimator (locally)

supervised learning:

training data $\mathcal{D}$ is $X_1, \dots, X_N$ with labels $Y_1, \dots, Y_N$
(e.g. in regression, $X_i \in \mathbb{R}^n, Y_i \in \mathbb{R}$)

parameter θ "is" a generative function $f = f_\theta$:
$Y_i = f(X_i) + \varepsilon_i$

bias of $\widehat{f}$ at X_i: $\text{Bias}(\widehat{f}|X_i) = \mathbb{E}_Y[\widehat{f}(X_i) - f(X_i)]$

variance of $\widehat{f}$ at X_i: $\text{Var}(\widehat{f}|X_i) = \mathbb{E}_Y\left[(\widehat{f}(X_i) - \mathbb{E}_Y[\widehat{f}(X_i)])^2\right]$

MSE of $\widehat{f}$ at X_i: $\text{MSE}(\widehat{f}|X_i) = \mathbb{E}_Y\left[(\widehat{f}(X_i) - Y_i)^2\right]$

Proposition: $\text{MSE}(\widehat{f}|X_i) = \text{Var}(\varepsilon_i) + \text{Bias}(\widehat{f}|X_i)^2 + \text{Var}(\widehat{f}|X_i)$

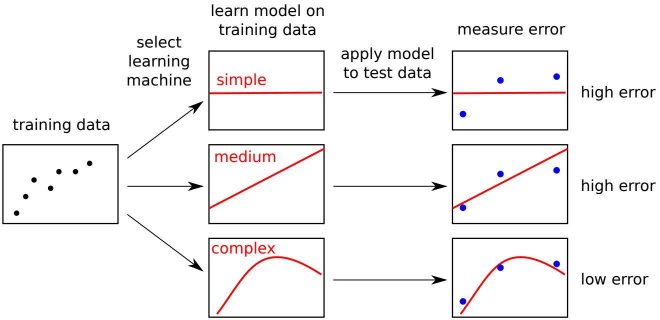
Predicting the Generalization Error

To establish the superiority of a learning machine over the other (e.g. JS vs. mean) we must make data-generating assumptions (e.g. $X \sim \mathcal{N}(\mu, \sigma I)$) and know the optimal parameter under these data-generating assumptions (here μ).

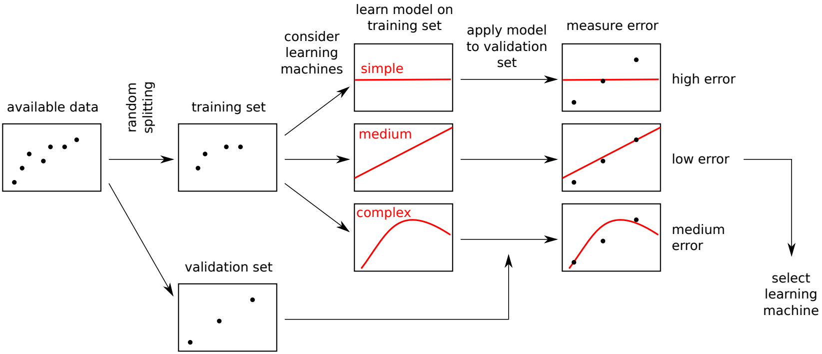
Predicting the Generalization Error

To establish the superiority of a learning machine over the other (e.g. JS vs. mean) we must make data-generating assumptions (e.g. $X \sim \mathcal{N}(\mu, \sigma I)$) and know the optimal parameter under these data-generating assumptions (here μ).

But in the *general case*, **how do we know in advance which learning machine to select?**



The Holdout Selection Procedure

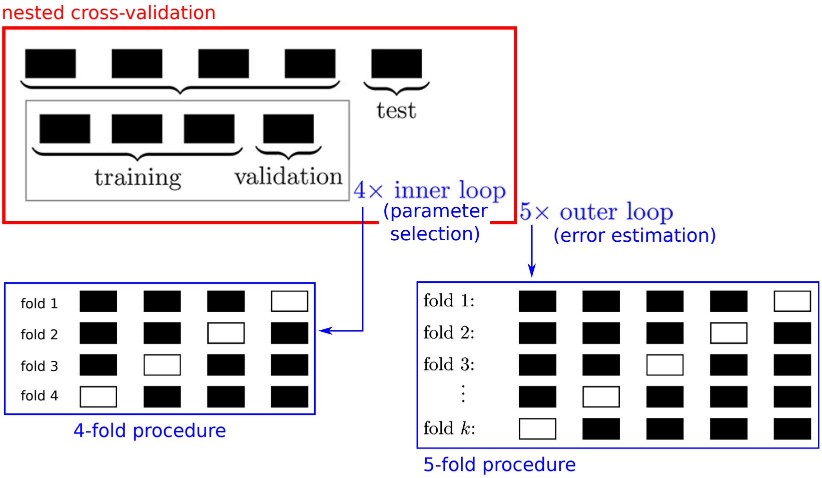


Remarks:

- For limited data, splitting data in two sets can reduce the model quality.
- There is a tradeoff between the amount of data used to train the model and the amount of data used to measure the error.
- To improve the error estimate, the process can be repeated for several splits, and the measured errors averaged (k-fold validation).

Model Selection and Validation in Practice

Nested k-fold validation:



Wrap-up

- **(1st) Occam's Razor:** Given two models with the same generalization error, the simpler one should be preferred, because simplicity is desirable in itself.
- **Model Selection/Validation:** How to make sure that a model predicts well? By testing it on out-of-sample data. Holdout+k-fold validation can be used to produce such performance estimate.
- **Bias-Variance Decomposition:** The performance of a predictive model can be decomposed into bias and variance.

Never use the test set to train the model or select the parameters.