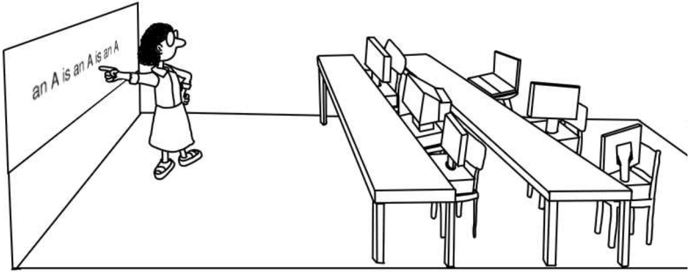
Machine Learning

Krystian Mikolajczyk & Deniz Gunduz

Department of Electrical and Electronic Engineering
Imperial College London



Machine Learning - Part 2.3 Summary

Department of Electrical and Electronic Engineering
Imperial College London

- Bias-variance trade-off

Bias-Variance Trade-Off

VC analysis: test error $\leqslant$ training error $+$ complexity penalty

Another approach: **bias-variance** analysis:

test error $=$ bias $+$ variance

- Bias: how well can $\mathcal{H}$ approximate f? (as before)
- Variance: how well can we select a good $h \in \mathcal{H}$?

Setup:

- e.g. $\mathbf{x}$ - patient record, $\mathcal{D}$ - a hospital, y - cost prediction.
- $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ where $y_i = f(\mathbf{x}_i) \in \mathbb{R}$.
- Test error within $\mathcal{D}$ (squared):
  $R(g^{(\mathcal{D})}) = \mathbb{E}\left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2 \middle| \mathcal{D}\right] = \mathbb{E}_{\mathbf{x}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2\right]$

Bias-Variance Analysis

Test error within $\mathcal{D}$:

$$R(g^{(\mathcal{D})}) = \mathbb{E}_x\left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2\right]$$

Expected test error (over many $\mathcal{D}_1, \ldots, \mathcal{D}_K$) and $(\mathbf{x}_1, \ldots, \mathbf{x}_n)^{(\mathcal{D})}$:

$$\begin{aligned}
\mathbb{E}\left[R(g^{(\mathcal{D})})\right] &= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{x}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2\right]\right] \\
&= \mathbb{E}\left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2\right] \\
&= \mathbb{E}_x\left[\mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2\right]\right]
\end{aligned}$$

The Average Hypothesis

Concentrate on $\mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2\right]$ for a given $\mathbf{x} \in \mathcal{X}$ (patient x!).

Average hypothesis over many datasets $\mathcal{D}_1, \ldots, \mathcal{D}_K$

The best possible : $\bar{g}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\mathbf{x})\right] \approx \frac{1}{K}\sum_1^K g^{(\mathcal{D}_k)}(\mathbf{x})$

Expected error for patient $\mathbf{x}$ with costs predicted by many $g^{(\mathcal{D})}(\mathbf{x})$

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2\right] &= \mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}) + \bar{g}(\mathbf{x}) - f(\mathbf{x}))^2\right] \\
&= \mathbb{E}_{\mathcal{D}}\Big[(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2 + (\bar{g}(\mathbf{x}) - f(\mathbf{x}))^2 \\
&\qquad + 2(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}))(\bar{g}(\mathbf{x}) - f(\mathbf{x}))\Big] \\
&= \mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2\right] + (\bar{g}(\mathbf{x}) - f(\mathbf{x}))^2 \\
&\qquad + 2\underbrace{\mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}))(\bar{g}(\mathbf{x}) - f(\mathbf{x}))\right]}_{2(\bar{g}(\mathbf{x}) - \bar{g}(\mathbf{x}))(const)}
\end{aligned}
$$

Bias and Variance

$$\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right] = \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right]}_{\text{var}(\mathbf{x})} + \underbrace{\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2}_{\text{bias}(\mathbf{x})} \; .$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[R(g^{(\mathcal{D})})\right] &= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{x}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\text{bias}(\mathbf{x}) + \text{var}(\mathbf{x})\right] \\
&= \text{bias} + \text{var} \, .
\end{aligned}$$

$\text{bias} = \mathbb{E}_{\mathbf{x}}\left[\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]$

- how far $\bar{g}(\mathbf{x})$ from $f(\mathbf{x})$
- large if $\mathcal{H}$ is small
- small if $\mathcal{H}$ is large

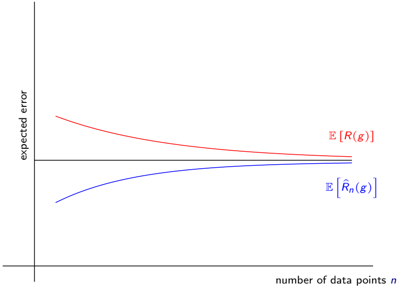
$\text{var} = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})^2\right]\right]$

- how far $g^{(\mathcal{D})}(\mathbf{x})$ from $\bar{g}(\mathbf{x})$
- small if $\mathcal{H}$ is small
- large if $\mathcal{H}$ is large

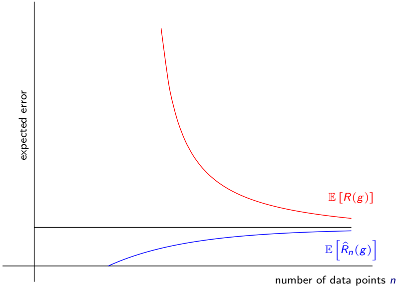
Complexity-Performance Trade-off

Match the model complexity to the data not to the target complexity!

Learning curves:

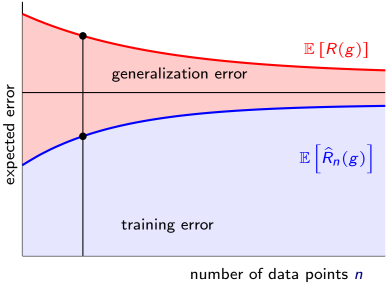


simple model

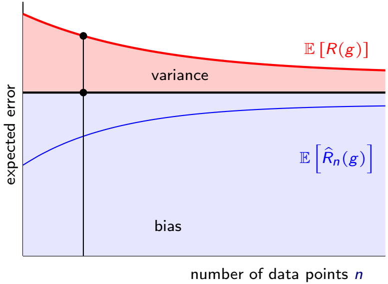


complex model

VC vs Bias-Variance



VC analysis



bias-variance

- best approximation in between $\mathbb{E}[R(g)]$ and $\mathbb{E}\left[\widehat{R}_n(g)\right]$
- in VC the error is on the training sample
- bias based on the best approximation $\bar{g}(x)$ (over all $^{(\mathcal{D})}$)
- bias constant, only depends on $\mathcal{H}$ not on n

Terminology

- Training set: $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.
- Test set: $\mathcal{D}' = \{(x'_1, y'_1), \ldots, (x'_m, y'_m)\}$.
- Loss function: $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$.

	statistics	learning theory	machine learning
$\frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i), y_i)$	in-sample error E_{in}	empirical risk $\widehat{R}_n, L_n, \mathcal{L}_n$	training error
$\mathbb{E}\left[\ell(h(x), y)\right]$ $\mathbb{E}\left[\ell(g^{(\mathcal{D})}(x), y) \middle\vert \mathcal{D}\right]$	out-of-sample error E_{out}	risk, generalization error $R, L, \mathcal{L}$	(true) test error
$\mathbb{E}\left[\ell(g^{(\mathcal{D})}(x), y)\right]$	expected out-of-sample error	expected risk	expected test error
$\frac{1}{n}\sum_{i=1}^{m} \ell(h(x'_i), y'_i)$	test error E_{test}	empirical test error $\widehat{R}'_m$	(empirical) test error

Relations:

- $R(h) = \mathbb{E}\left[\widehat{R}_n(h)\right]$ + high prob. by Hoeffding
- $R(g^{(\mathcal{D})})$ and $\widehat{R}_n(g^{(\mathcal{D})})$: typically $\mathbb{E}\left[\widehat{R}_n(g^{(\mathcal{D})}) \middle\vert \mathcal{D}\right] \neq R(g^{(\mathcal{D})})$ but h.p. by VC inequality
- $R(h) = \mathbb{E}\left[\widehat{R}'_m(h)\right]$ and $R(g^{(\mathcal{D})}) = \mathbb{E}\left[\widehat{R}'_m(g^{(\mathcal{D})}) \middle\vert \mathcal{D}\right]$; h.p. by Hoeffding (in both cases!)

Part 2 Summary

- Feasibility of learning
- Hoeffding's inequality
- Target distribution and error cost
- Multiple hypothesis
- Growth function
- VC inequality
- Bias-variance trade-off