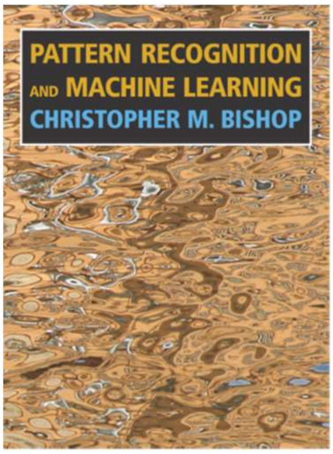
Bias-Variance Decomposition

Song Liu (song.liu@bristol.ac.uk)

Reference



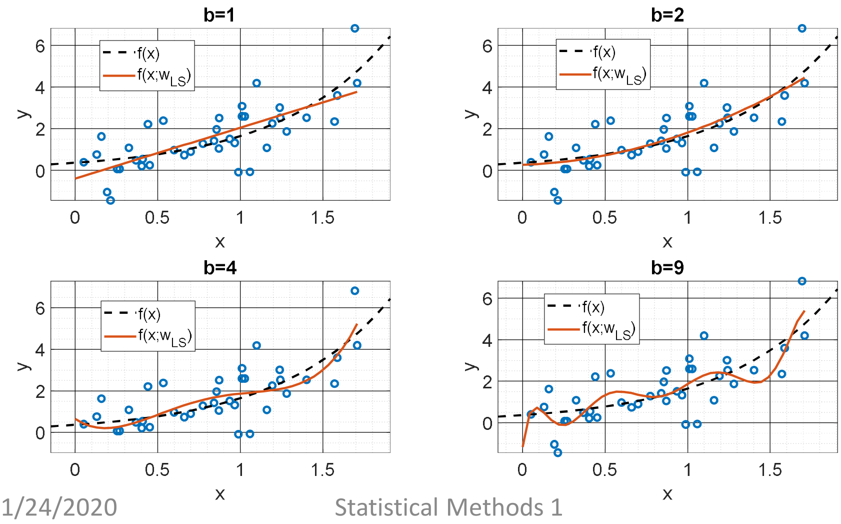
Today's class *roughly* follows Chapter 3.2.

Pattern Recognition and Machine Learning

Christopher Bishop, 2006

Poly. Feature with various b

- $y = g(x) + \epsilon, g(x) = \exp(1.5x - 1), \epsilon \sim N(0, .64)$



Q, write down conditional mean.

What Really Happened?

- We mentioned that $f(\boldsymbol{x}; \boldsymbol{w}_{\mathrm{LS}})$ is too flexible to generalize well on unobserved dataset, but why?
- What is the mathematical explanation of OF?
- Why testing error is a good measurement of the generalization of a prediction $f(\boldsymbol{x}; \boldsymbol{w}_{\mathrm{LS}})$?

- We are introducing a frequentist analysis of explaining this phenomenon, called **Variance and Bias decomposition**.

20

This explanation only valid under the assumption introduced later.

This analysis requires a squared loss function. You can expand this idea to other loss functions, but the analysis procedure is usually less obvious than the one for squared loss function.

From Training Error to Expected Loss

- $E(D, \boldsymbol{w}_{\text{LS}})$ is the **training error** of $\boldsymbol{w}_{\text{LS}}$ on a training set D**.**
- We do not care $E(D, \boldsymbol{w}_{\text{LS}})$ on a specific training dataset, let us take expectation with respect to D:

$$\mathbb{E}_D[E(D, w_{\text{LS}})] = \mathbb{E}_D\left[\sum_{i \in D}[y_i - f(\boldsymbol{x}_i; \boldsymbol{w}_{\text{LS}})]^2\right]$$

$$= \sum_{i=1..n} \underbrace{\mathbb{E}_D\left[[y - f(\boldsymbol{x}_i; \boldsymbol{w}_{\text{LS}})]^2 | \boldsymbol{x}_i\right]}_{\textbf{Expected Loss!}}$$

To investigate the expected loss further, we need to make some assumptions on the randomness of D.

21

Additive Noise Assumption

- First, assume an outcome y_i is generated by
- $y_i = g(\boldsymbol{x}_i) + \epsilon_i$.
  - $g(\boldsymbol{x})$: $R^d \rightarrow R$ is some deterministic function.
  - $\forall_i, \epsilon_i$ is independent of $\boldsymbol{x}_i$ and $\mathbb{E}[\epsilon_i] = 0$
  - We call ϵ_i **additive noise**.

- For example, if we assume ϵ_i comes from normal dist. with mean 0 and variance σ^2, y_i follows a normal distribution with mean $g(\boldsymbol{x}_i)$ and variance σ^2.

This analysis does NOT require a distributional assumption on epsilon.

Decomposition of Expected Loss

- $\mathbb{E}_D\left[[y - f_{\mathrm{LS}}(\boldsymbol{x}_i)]^2 | \boldsymbol{x}_i\right] = \mathbb{E}_\epsilon\left[[y - f_{\mathrm{LS}}(\boldsymbol{x}_i)]^2 | \boldsymbol{x}_i\right]$

$$= \underbrace{\mathrm{var}_\epsilon[\epsilon]}_{\text{Irreducible error}} + \underbrace{\left[g(\boldsymbol{x}_i) - \mathbb{E}_\epsilon[f_{\mathrm{LS}}(\boldsymbol{x}_i)|\boldsymbol{x}_i]\right]^2}_{\text{bias}} + \underbrace{\mathrm{var}_\epsilon[f_{\mathrm{LS}}(\boldsymbol{x}_i)|\boldsymbol{x}_i}_{\text{variance}}$$

- “Variance and Bias decomposition”
- Prove it, hint, by our data generating assumption:
- $\mathbb{E}_\epsilon\left[[y - f_{\mathrm{LS}}(\boldsymbol{x}_i)]^2 | \boldsymbol{x}_i\right] = \mathbb{E}_\epsilon\left[[g(\boldsymbol{x}_i) + \epsilon - f_{\mathrm{LS}}(\boldsymbol{x}_i)]^2 | \boldsymbol{x}_i\right]$

23

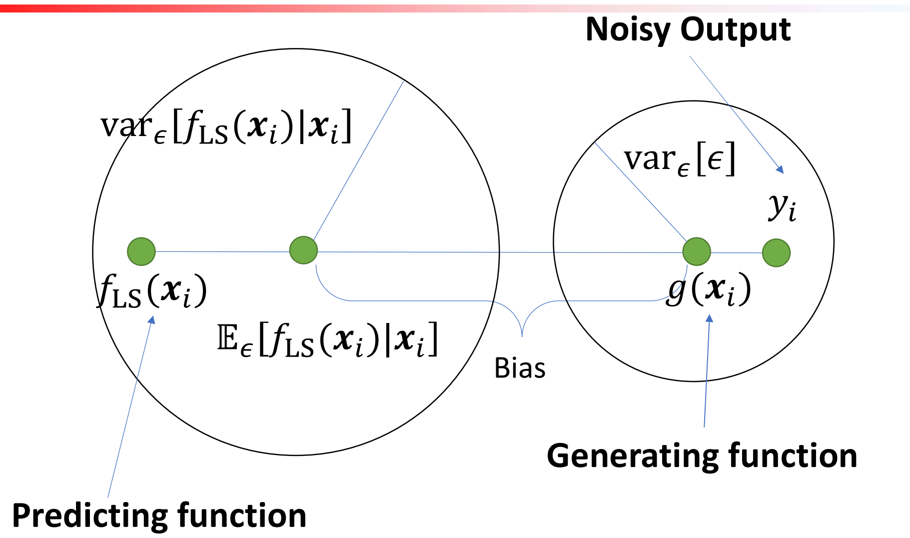
This decomposition does **not** require the explicit expression of our prediction function f_{LS}.

"Variance and Bias decomposition"

- $\mathrm{var}[\epsilon] + \left[g(\boldsymbol{x}_i) - \mathbb{E}[f_{\mathrm{LS}}(\boldsymbol{x}_i)|\boldsymbol{x}_i]\right]^2 + \mathrm{var}[f_{\mathrm{LS}}(\boldsymbol{x}_i)|\boldsymbol{x}_i]$
  - 1st term measures the randomness of our data generating process, which is beyond our control.
  - 2nd term shows the accuracy of our expected prediction.
  - 3rd term shows how easily our fitted prediction function is affected by the randomness of the dataset.

The reason that our prediction function wrapped inside of an expectation is because that the prediction is also influenced by randomness of our dataset.

A Visualization of V-B Decomposition

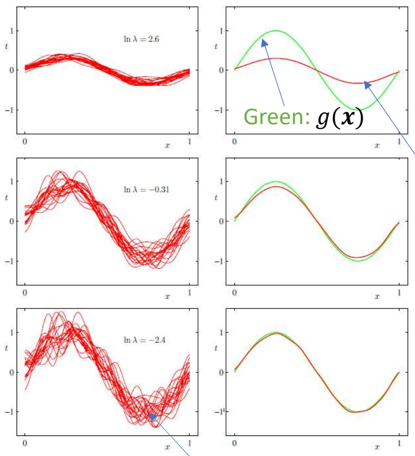


Variance and Bias Tradeoff

- $\mathrm{var}[\epsilon] + \left[g(\boldsymbol{x}_i) - \mathbb{E}[f_{\mathrm{LS}}(\boldsymbol{x}_i)|\boldsymbol{x}_i]\right]^2 + \mathrm{var}[f_{\mathrm{LS}}(\boldsymbol{x}_i)|\boldsymbol{x}_i]$
  - As we increase b, f_{LS} becomes more **complex** and can adapt to more complex underlying function, thus 2nd term keeps reducing.
  - As we increase b, f_{LS} becomes more **sensitive** to the noise in our dataset, thus 3rd term keeps increasing.
  - A **balance** between 2nd and 3rd term gives the minimum expected error.

26

Variance and Bias Tradeoff



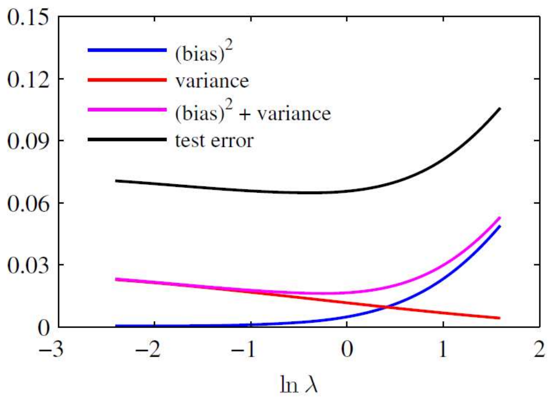
- As flexibility increases (λ decreases), the bias decreases, and the variance increases.

Red: **Expected** f_{LS}

PRML Figure 3.5

Red: f_{LS} over different datasets, see the variances

Variance and Bias Tradeoff



PRML Figure 3.6

- As the flexibility decreases (λ increase), bias increases and the variance decreases.

28

Notice the behavior testing error, almost the same as the bias+variance, only up to a constant.

We will investigate this later.

In-Sample Error

- $\mathbb{E}_\epsilon[(y - f_{\text{LS}}(\boldsymbol{x}_i))^2 | \boldsymbol{x}_i]$ is conditional on $\boldsymbol{x}_i$.
- To calculate the collective error, we need to average over all $\boldsymbol{x}_i$.
  - $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\epsilon[(y - f_{\text{LS}}(\boldsymbol{x}_i))^2 | \boldsymbol{x}_i]$
  - is called **in sample error**

- Can we use in sample error to measure the performance of our f_{LS}?

29

Can we use this expected loss evaluating the performance of our prediction?

Out-Sample Error

- In sample error is not useful in practice.
  - We cannot calculate $\mathbb{E}_\epsilon[(y - f_{\mathrm{LS}}(\boldsymbol{x}_i))^2 | \boldsymbol{x}_i]$
  - We do not know $g(\boldsymbol{x})$ and the distribution of ϵ.
- Instead, we use **out-sample error**:
  - Error over the entire distribution of $\boldsymbol{x}$:
  - $\boxed{\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_\epsilon[(y - f_{\mathrm{LS}}(\boldsymbol{x}))^2 | \boldsymbol{x}]}$
  - $\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_\epsilon[(y - f_{\mathrm{LS}}(\boldsymbol{x}))^2 | \boldsymbol{x}] = \mathbb{E}_{\boldsymbol{x}}\mathbb{E}_y[(y - f_{\mathrm{LS}}(\boldsymbol{x}))^2 | \boldsymbol{x}]$
    $= \mathbb{E}_{p(y,\boldsymbol{x})}[(y - f_{\mathrm{LS}}(\boldsymbol{x}))^2]$
- Can we approximate out-sample error?

30

The first equality is due to the law of the unconscious statistician (LOTUS):

https://en.wikipedia.org/wiki/Law_of_the_unconscious_statistician

It is said that many statistician uses this law without noticing it, hence the name.

Approx. Out-Sample Error

- Train least-squares on dataset D_0, getting f_0 ,
- Obtain a fresh batch datapoints $D_1 := \{(y_i', \boldsymbol{x}_i')\}_{i=1}^{n'}$,
- D_1 and D_0 are independently and identically distributed:
- $\frac{1}{n'}\sum_{(y',\boldsymbol{x}')\in D_1}\left(y' - f_0(\boldsymbol{x}')\right)^2 \approx \mathbb{E}_{p(y,\boldsymbol{x})}[(y - f_0(\boldsymbol{x}))^2]$
  - due to law of large numbers.
- $\mathbb{E}_{p(y,\boldsymbol{x})}[(y - f_0(\boldsymbol{x}))^2] \approx \mathbb{E}_{p(y,\boldsymbol{x})}[(y - f_{\mathrm{LS}}(\boldsymbol{x}))^2]$

- $\frac{1}{n'}\sum_{(y',\boldsymbol{x}')\in D_1}(y' - f_0(\boldsymbol{x}'))^2$ is $E(D_1, f_0)$!
- This justifies the usage of $E(D_1, f_0)$ for evaluating the overfitting of our prediction f_0.

31

This is NOT a mathematical proof!!

As a matter of fact, the second approximation, using $\mathbb{E}_{p(y,\boldsymbol{x})}[(y - f_0(\boldsymbol{x}))^2]$ to approximate $\mathbb{E}_{p(y,\boldsymbol{x})}[(y - f_{\mathrm{LS}}(\boldsymbol{x}))^2]$ is very rough, as we are replacing one of the random variables $f_{\mathrm{LS}}(\boldsymbol{x})$ with a fixed value $f_0(\boldsymbol{x})$. The approximation accuracy may vary.

Conclusion

- The phenomenon of OF can be explained by decomposition of expected error.
- Two types of expected errors can be used for measuring the performance of f_{LS}:
  - In-sample error, cannot be computed, unless we know g and dist. of ϵ.
  - Out-sample error, can be roughly approximated by $E(D_1, f_0)$, which is the testing error.

Homework

- Prove variance and bias decomposition.
  - Page 23