

# Machine Learning I

## Supervised Learning: Bias and variance decomposition

Souhaib Ben Taieb

University of Mons



# Table of contents

A note on the data distribution

The bias and variance decomposition

The bias and variance tradeoff

# Table of contents

**A note on the data distribution**

The bias and variance decomposition

The bias and variance tradeoff

# Data distribution in regression

The data distribution  $p_{x,y}$  is often **implicitly specified**, i.e.  $p_{x,y}$  is not given explicitly. In regression, the following (additive error) data generating process is often considered:

$$y = f(x) + \varepsilon, \tag{1}$$

where

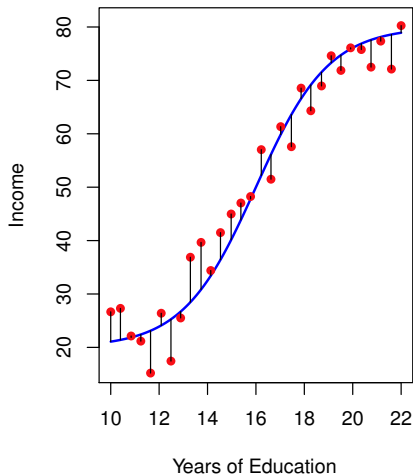
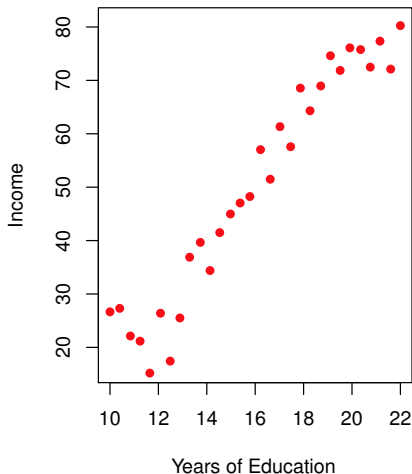
- ▶  $x \sim p_x$  (e.g.  $p_x(x) = \frac{1}{2}$  for  $x \in [-1, 1]$ )
- ▶  $f$  is a fixed unknown function (e.g.  $f(x) = x^2$ )
- ▶  $\varepsilon$  is random noise, where
  - ▶  $\mathbb{E}[\varepsilon|x] = 0$
  - ▶  $\text{Var}(\varepsilon|x) = \sigma^2$ , with  $\sigma \in [0, \infty)$ .

Note that we have

- ▶  $\mathbb{E}[y|x] = f(x)$  and  $\text{Var}[y|x] = \sigma^2$

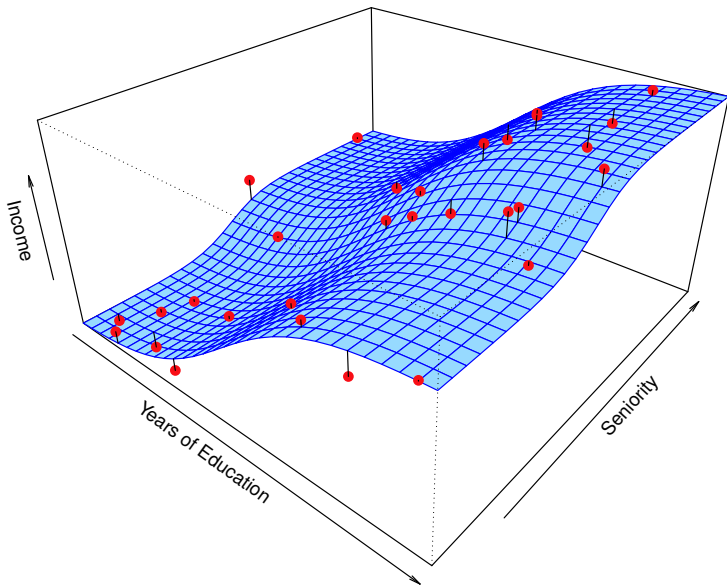
i.e.  $p_{y|x}$  depends on  $x$  only through the conditional expectation.

# Data distribution in regression



→ Try to visualize  $p_{x,y}$

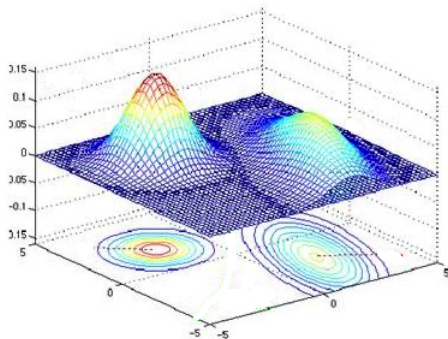
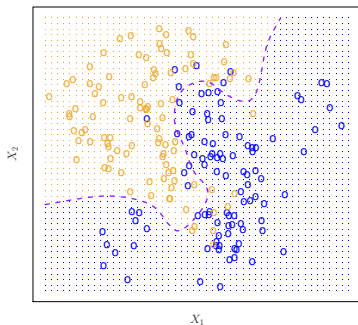
# Data distribution in regression



# Data distribution in classification

Using Bayes' rule, we can write

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \propto p(x|y)p(y) \stackrel{y \text{ uniform}}{\propto} p(x|y)$$



# Table of contents

A note on the data distribution

**The bias and variance decomposition**

The bias and variance tradeoff



# The bias-variance decomposition

- ▶ Previously, we considered the **unrealistic scenario** where we know  $p_{x,y}$ . As a result, we were able to compute the optimal hypothesis/predictions for different loss functions.
- ▶ In practice, we only observe a **dataset**  $\mathcal{D}$  where each data point is assumed to be an i.i.d. realization from  $p_{x,y}$ .
- ▶ Overly simple models underfit and complex models overfit. There is an **approximation-generalization** tradeoff:

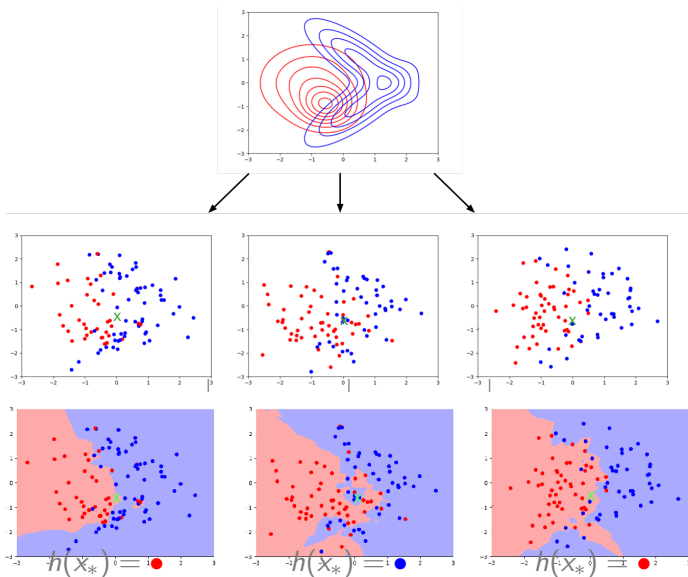
$$E_{\text{out}}(g) - E_{\text{out}}(f) = \underbrace{[E_{\text{out}}(g^*) - E_{\text{out}}(f)]}_{\text{Approximation error}} + \underbrace{[E_{\text{out}}(g) - E_{\text{out}}(g^*)]}_{\text{Estimation error}}$$

- ▶ The **bias-variance** decomposition allows to quantify this tradeoff for the **squared error** loss function.

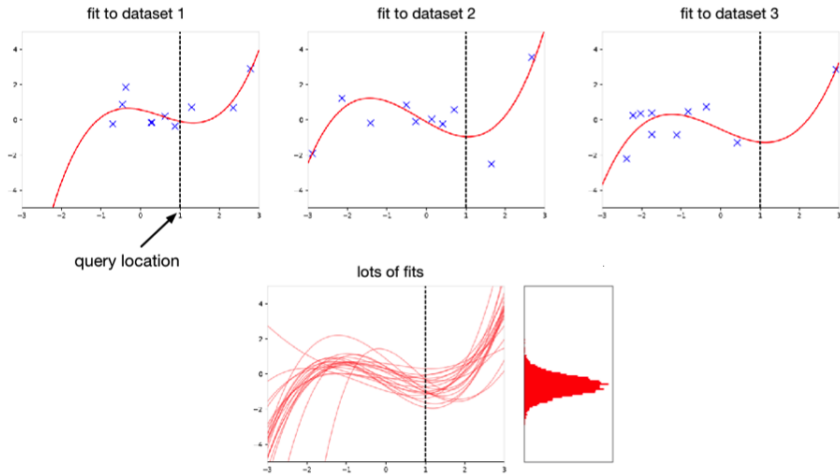
# An experiment

- ▶ Consider an experiment where we sample **lots of training sets** independently from  $p_{\mathbf{x},\mathbf{y}}$ .
- ▶ Pick a fixed **query point**  $x_*$ .
- ▶ Let's run our learning algorithm on each training set, and compute its prediction  $g(x_*)$  at the query point  $x_*$ .
- ▶ We can view  $g(x_*)(=g_{\mathcal{D}}(x_*))$  as a **random variable**, where the randomness comes from the training set  $\mathcal{D}$ .

# Classification example



# Regression example



# An experiment (continued)

- ▶ Fix a query point  $x_*$ .
- ▶ Repeat:
  - ▶ Sample a dataset  $\mathcal{D}$  i.i.d. from  $p_{x,y}$
  - ▶ Run the learning algorithm on  $\mathcal{D}$  to obtain  $g$
  - ▶ Compute the prediction for  $x_*$ , i.e.  $g(x_*)$
  - ▶ Sample the (true) output  $y_*$  from  $p_{y|x}(\cdot|x = x_*)$
  - ▶ Compute the loss  $L(y_*, g(x_*))$

$L(y_*, g(x_*))$  contains two **sources of randomness**:  $\mathcal{D}$  and  $y_*$ . This gives a distribution over the loss at  $x_*$ .

Let us expand

$$\mathbb{E}_{\mathcal{D}} [\mathbb{E}_{y|x} [L(y, g(x))|x]]$$

for the squared error loss  $L(y, \hat{y}) = (y - \hat{y})^2$ .

# An experiment (continued)

- ▶ Fix a query point  $x_*$ .
- ▶ Repeat:
  - ▶ Sample a dataset  $\mathcal{D}$  i.i.d. from  $p_{x,y}$
  - ▶ Run the learning algorithm on  $\mathcal{D}$  to obtain  $g$
  - ▶ Compute the prediction for  $x_*$ , i.e.  $g(x_*)$
  - ▶ Sample the (true) output  $y_*$  from  $p_{y|x}(\cdot|x = x_*)$
  - ▶ Compute the loss  $L(y_*, g(x_*))$

$L(y_*, g(x_*))$  contains two **sources of randomness**:  $\mathcal{D}$  and  $y_*$ . This gives a distribution over the loss at  $x_*$ .

Let us expand

$$\mathbb{E}_{\mathcal{D}} [\mathbb{E}_{y|x} [L(y, g(x))|x]]$$

for the squared error loss  $L(y, \hat{y}) = (y - \hat{y})^2$ .

# The bias-variance decomposition

Recall that

$$\mathbb{E}_{y|x}[(y - g(x))^2|x] = \text{Var}(y|x) + (f(x) - g(x))^2 \text{ where } f(x) = \mathbb{E}[y|x].$$

We can write

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}} [\mathbb{E}_{y|x} [(y - g(x))^2|x]] \\ &= \text{Var}(y|x) + \mathbb{E}_{\mathcal{D}} [(f(x) - g(x))^2] \\ &= \text{Var}(y|x) + f(x)^2 - 2f(x)\mathbb{E}_{\mathcal{D}}[g(x)] + \mathbb{E}_{\mathcal{D}}[g(x)^2] \\ &= \text{Var}(y|x) + f(x)^2 - 2f(x)\mathbb{E}_{\mathcal{D}}[g(x)] + \text{Var}(g(x)) + \mathbb{E}_{\mathcal{D}}[g(x)]^2 \\ &= \underbrace{\text{Var}(y|x)}_{\text{Bayes error at } x} + \underbrace{(f(x) - \mathbb{E}_{\mathcal{D}}[g(x)])^2}_{\text{Bias at } x} + \underbrace{\text{Var}(g(x))}_{\text{Variance at } x} \end{aligned}$$

# The bias-variance decomposition

Recall that

$$\mathbb{E}_{y|x}[(y - g(x))^2|x] = \text{Var}(y|x) + (f(x) - g(x))^2 \text{ where } f(x) = \mathbb{E}[y|x].$$

We can write

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}} [\mathbb{E}_{y|x} [(y - g(x))^2|x]] \\ &= \text{Var}(y|x) + \mathbb{E}_{\mathcal{D}}[(f(x) - g(x))^2] \\ &= \text{Var}(y|x) + f(x)^2 - 2f(x)\mathbb{E}_{\mathcal{D}}[g(x)] + \mathbb{E}_{\mathcal{D}}[g(x)^2] \\ &= \text{Var}(y|x) + f(x)^2 - 2f(x)\mathbb{E}_{\mathcal{D}}[g(x)] + \text{Var}(g(x)) + \mathbb{E}_{\mathcal{D}}[g(x)]^2 \\ &= \underbrace{\text{Var}(y|x)}_{\text{Bayes error at } x} + \underbrace{(f(x) - \mathbb{E}_{\mathcal{D}}[g(x)])^2}_{\text{Bias at } x} + \underbrace{\text{Var}(g(x))}_{\text{Variance at } x} \end{aligned}$$



# The bias-variance decomposition

Recall that

$$\mathbb{E}_{y|x}[(y - g(x))^2|x] = \text{Var}(y|x) + (f(x) - g(x))^2 \text{ where } f(x) = \mathbb{E}[y|x].$$

We can write

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}} [\mathbb{E}_{y|x} [(y - g(x))^2|x]] \\ &= \text{Var}(y|x) + \mathbb{E}_{\mathcal{D}} [(f(x) - g(x))^2] \\ &= \text{Var}(y|x) + f(x)^2 - 2f(x)\mathbb{E}_{\mathcal{D}}[g(x)] + \mathbb{E}_{\mathcal{D}}[g(x)^2] \\ &= \text{Var}(y|x) + f(x)^2 - 2f(x)\mathbb{E}_{\mathcal{D}}[g(x)] + \text{Var}(g(x)) + \mathbb{E}_{\mathcal{D}}[g(x)]^2 \\ &= \underbrace{\text{Var}(y|x)}_{\text{Bayes error at } x} + \underbrace{(f(x) - \mathbb{E}_{\mathcal{D}}[g(x)])^2}_{\text{Bias at } x} + \underbrace{\text{Var}(g(x))}_{\text{Variance at } x} \end{aligned}$$

# The bias-variance decomposition

Recall that

$$\mathbb{E}_{y|x}[(y - g(x))^2|x] = \text{Var}(y|x) + (f(x) - g(x))^2 \text{ where } f(x) = \mathbb{E}[y|x].$$

We can write

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}} [\mathbb{E}_{y|x} [(y - g(x))^2|x]] \\ &= \text{Var}(y|x) + \mathbb{E}_{\mathcal{D}}[(f(x) - g(x))^2] \\ &= \text{Var}(y|x) + f(x)^2 - 2f(x)\mathbb{E}_{\mathcal{D}}[g(x)] + \mathbb{E}_{\mathcal{D}}[g(x)^2] \\ &= \text{Var}(y|x) + f(x)^2 - 2f(x)\mathbb{E}_{\mathcal{D}}[g(x)] + \text{Var}(g(x)) + \mathbb{E}_{\mathcal{D}}[g(x)]^2 \\ &= \underbrace{\text{Var}(y|x)}_{\text{Bayes error at } x} + \underbrace{(f(x) - \mathbb{E}_{\mathcal{D}}[g(x)])^2}_{\text{Bias at } x} + \underbrace{\text{Var}(g(x))}_{\text{Variance at } x} \end{aligned}$$

# The bias-variance decomposition

Recall that

$$\mathbb{E}_{y|x}[(y - g(x))^2|x] = \text{Var}(y|x) + (f(x) - g(x))^2 \text{ where } f(x) = \mathbb{E}[y|x].$$

We can write

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}} [\mathbb{E}_{y|x} [(y - g(x))^2|x]] \\ &= \text{Var}(y|x) + \mathbb{E}_{\mathcal{D}} [(f(x) - g(x))^2] \\ &= \text{Var}(y|x) + f(x)^2 - 2f(x)\mathbb{E}_{\mathcal{D}}[g(x)] + \mathbb{E}_{\mathcal{D}}[g(x)^2] \\ &= \text{Var}(y|x) + f(x)^2 - 2f(x)\mathbb{E}_{\mathcal{D}}[g(x)] + \text{Var}(g(x)) + \mathbb{E}_{\mathcal{D}}[g(x)]^2 \\ &= \underbrace{\text{Var}(y|x)}_{\text{Bayes error at } x} + \underbrace{(f(x) - \mathbb{E}_{\mathcal{D}}[g(x)])^2}_{\text{Bias at } x} + \underbrace{\text{Var}(g(x))}_{\text{Variance at } x} \end{aligned}$$

# The bias-variance decomposition

$$\mathbb{E}_{\mathcal{D}, y|x}[(y - g(x))^2|x] = \underbrace{\text{Var}(y|x)}_{\text{Bayes error at } x} + \underbrace{(f(x) - \mathbb{E}_{\mathcal{D}}[g(x)])^2}_{\text{Bias at } x} + \underbrace{\text{Var}(g(x))}_{\text{Variance at } x}$$

We split the expected error at  $x$  into three terms:

- ▶ Bayes error: the inherent unpredictability of the output
- ▶ **bias**: how wrong the expected prediction is (underfitting)
- ▶ **variance**: the variability of the predictions (overfitting)

# The bias-variance decomposition

If we take the expectation with respect to  $x$ , we obtain

$$\begin{aligned} \mathbb{E}_{\mathcal{D}, y, x}[(y - g(x))^2] \\ = \underbrace{\text{Var}(y)}_{\text{Bayes error}} + \underbrace{\mathbb{E}_x[(f(x) - \mathbb{E}_{\mathcal{D}}[g(x)])^2]}_{\text{Bias}} + \underbrace{\mathbb{E}_x[\text{Var}(g(x))]}_{\text{Variance}} \end{aligned}$$

While the analysis only applies to squared error, we often use “bias” / “variance” as synonyms for “underfitting” / “overfitting”.

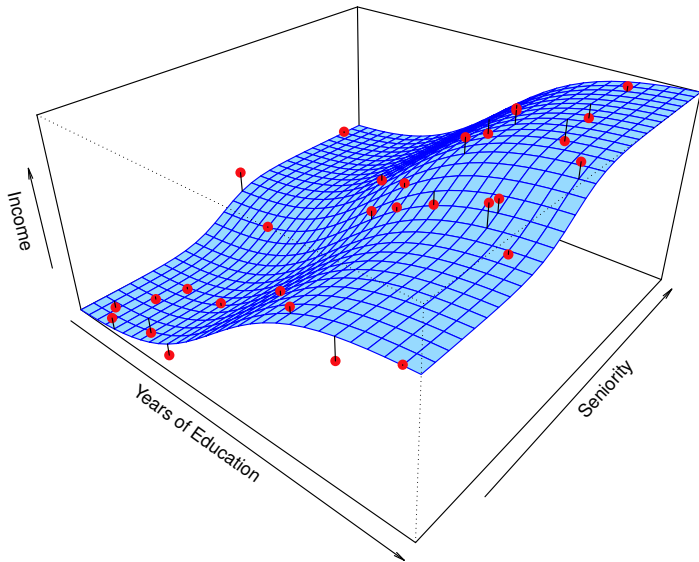
# Table of contents

A note on the data distribution

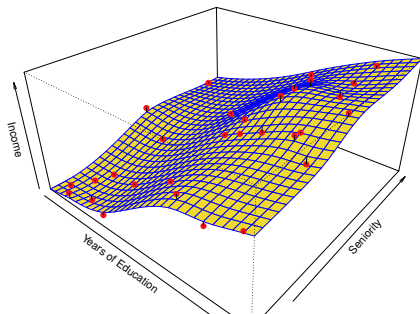
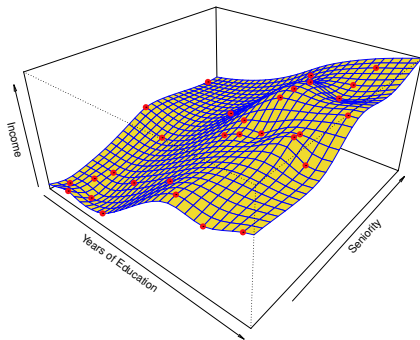
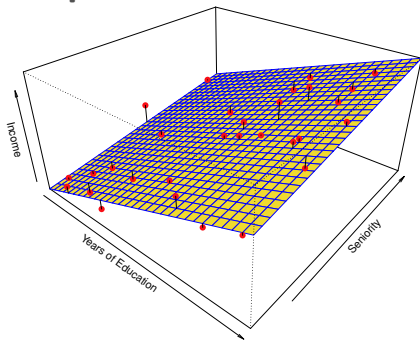
The bias and variance decomposition

**The bias and variance tradeoff**

# Example

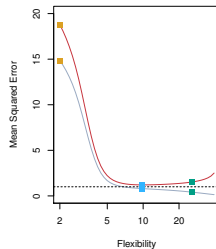
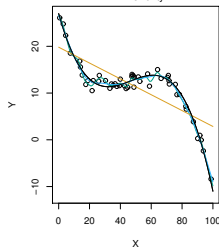
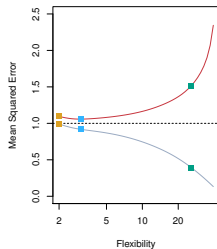
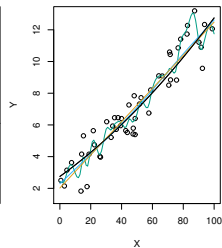
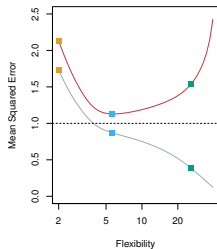
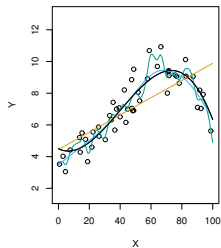


# Example

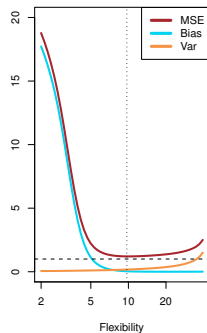
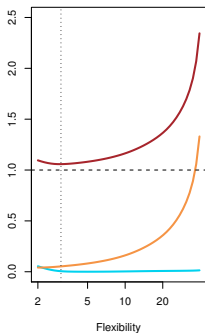
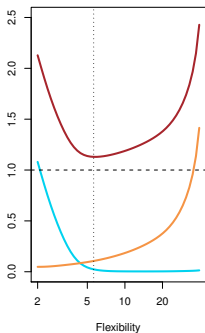
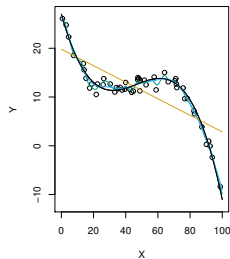
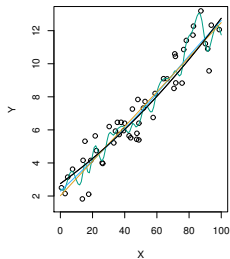
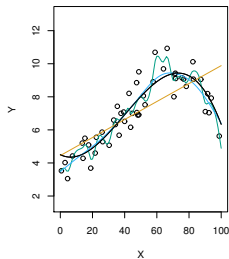




# Training and test errors (recap)



# The bias-variance tradeoff



# The bias-variance tradeoff

Throwing darts = predictions for each draw of a dataset.

