Lecture 2: Noise, Overfitting, and Bias vs Variance

Iain Styles

11 October 2019

Noise, Overfitting, and Bias vs Variance

By the end of this lecture you should be able to:

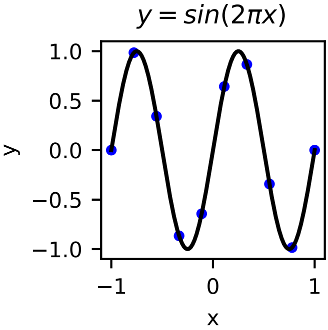
1. Understand the effect of noise on machine learning problems
2. Understand and explain the concepts of over and underfitting
3. Be able to explain these concepts using the idea of bias-variance decomposition

Choosing a model $f(\mathbf{w}, x)$

- If we know something about our data we may be able to deduce what f should be
  - $F = ma$, $s - ut + \frac{1}{2}at^2$, etc
- More often than not, we will not be able to do this and we will have to choose a representation (basis)
- This has to be done very carefully
- We will explore the implications of different choices in this lecture

A model problem

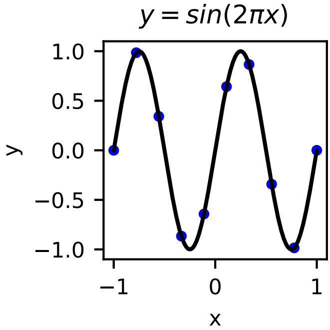
- We will continue to work with linear models but need example that
  - allows us to explore the power of linear models
  - study the effect of model choice in depth
- Our (my?) choice: $y(x) = \sin(2\pi x)$
- $f(\mathbf{w}, x) = w_0 \sin(2\pi x)$ is trivial, but we will assume no knowledge
- $f(\mathbf{w}, x) = \sum_{i=0}^{M-1} w_i x^i$ is a common and powerful choice



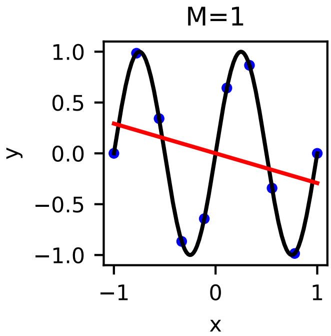
A reasonable expectation?

- $y(x) = \sin(2\pi x) = \sum_{i=0}^{M-1} w_i x^i$?
- Maclaurin series: $\sin(ax) = ax - \frac{a^3x^3}{3!} + \frac{a^5x^5}{5!} - \frac{a^7x^7}{7!} + \cdots$
- So we can evaluate the quality of the estimation of the underlying function
- True weights for $a = 2\pi$ are $\mathbf{w} \approx (0, 6.28, -41.34, 0, 81.61, 0, -76.7, 0, 42.1, \dots)$
- Start by generating "pure" data with no added noise
- Fit polynomial expansion up to order $M = 9$

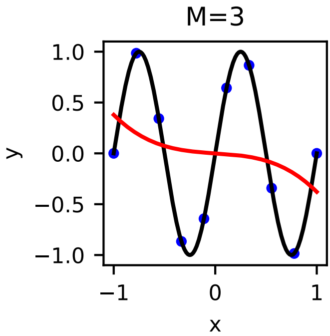
Polynomial fit of $y = sin(2\pi x)$



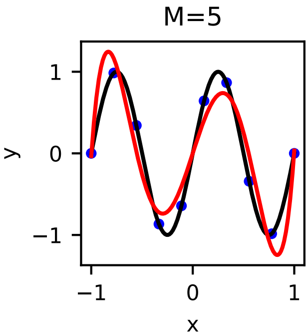
Polynomial fit of $y = sin(2\pi x)$



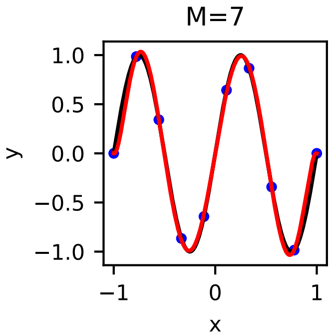
Polynomial fit of $y = sin(2\pi x)$



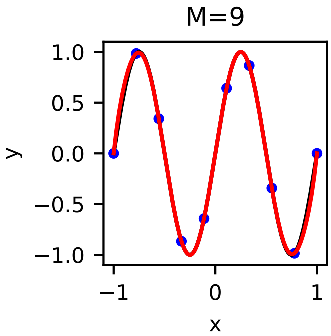
Polynomial fit of $y = sin(2\pi x)$



Polynomial fit of $y = sin(2\pi x)$

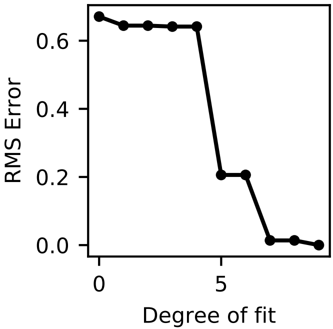


Polynomial fit of $y = sin(2\pi x)$



Evaluation

- How do we evaluate the quality of these results?
- Root-mean-square (RMS) error, $R = \sqrt{\frac{1}{N}\sum_i r_i^2} = \sqrt{\mathcal{L}_{\mathrm{LSE}}/N}$
- Normalises for number of data points
- Converges rapidly towards zero, with little change after $M = 7$
- Zero coefficients show as plateaux



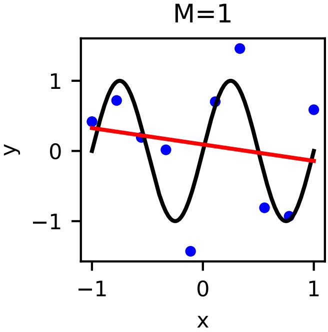
What are the coefficients of the fitted polynomial?

M	w_0	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9
0	0.00									
1	0.00	-0.29								
2	0.00	-0.29	-0.00							
3	0.00	-0.07	-0.00	-0.31						
4	0.00	-0.07	0.00	-0.31	-0.00					
5	0.00	3.85	-0.00	-16.51	0.00	12.69				
6	0.00	3.85	-0.00	-16.51	0.00	12.69	-0.00			
7	-0.00	6.00	0.00	-35.84	-0.00	54.04	0.00	-24.20		
8	-0.00	6.00	0.00	-35.84	-0.00	54.04	0.00	-24.20	-0.00	
9	-0.00	6.28	0.00	-41.12	-0.00	78.61	0.00	-63.77	-0.00	20.00
True	0	6.28	0	-41.34	0	81.61	0	-76.7	0	42.1

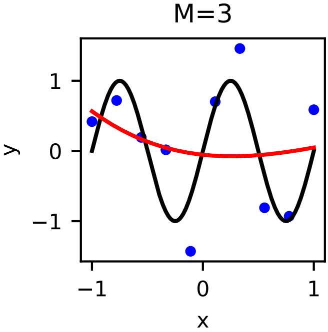
Analysis

- Coefficients are not quite correct
- Effect of limited sample domain (Maclaurin series is over $x \in [-\infty, \infty]$)
- Low order terms match well
- But note $M = 9$ has zero error
- Exactly fits all data point
- A strong hint as to what can go wrong
- Repeat, with added noise: $y = \sin(2\pi x) + \epsilon$

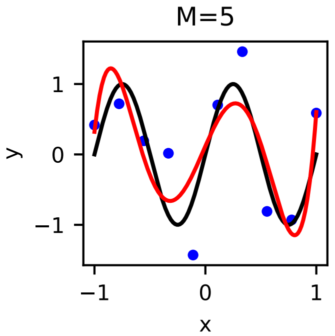
Polynomial fit of $y = sin(2\pi x) + \epsilon$



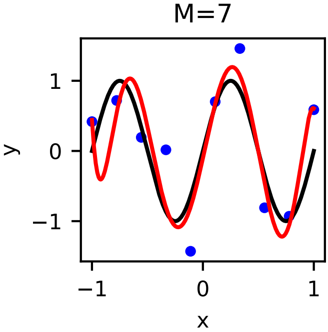
Polynomial fit of $y = sin(2\pi x) + \epsilon$



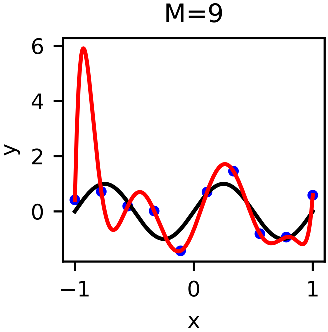
Polynomial fit of $y = sin(2\pi x) + \epsilon$



Polynomial fit of $y = sin(2\pi x) + \epsilon$



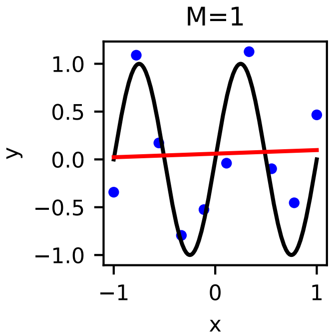
Polynomial fit of $y = sin(2\pi x) + \epsilon$



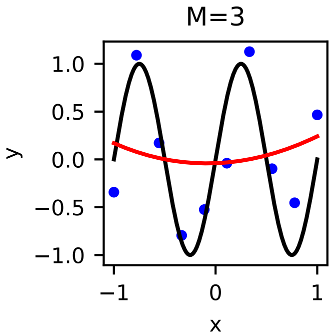
Noise can dramatically changes the result

- Low order fits are similar in noise-free and noisy cases
- High order fits differ quite dramatically
- Match data points exactly but do not model the underlying function, even though they are clearly able towards
- High order models also expressive enough to fit model + noise
- What happens if we have a different realisation of noise?

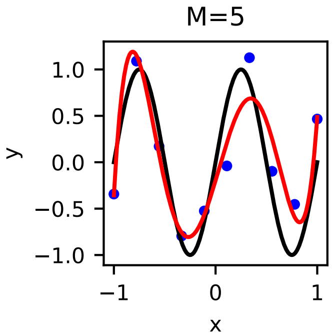
Polynomial fit of $y = sin(2\pi x) + \epsilon$



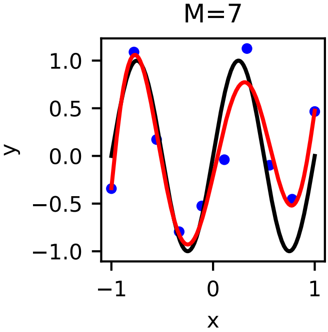
Polynomial fit of $y = sin(2\pi x) + \epsilon$



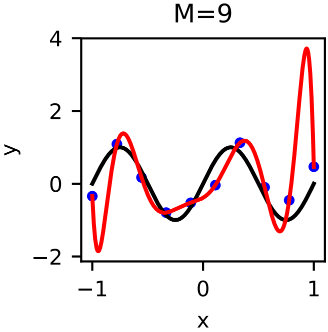
Polynomial fit of $y = sin(2\pi x) + \epsilon$



Polynomial fit of $y = sin(2\pi x) + \epsilon$



Polynomial fit of $y = sin(2\pi x) + \epsilon$



Noise corrupts

- Low-order fits similar
- High-order fits very different
- Noise in the data leads to noise in the estimated model
- Robust models cannot model the data very well
- How can we understand this?

Bias-Variance Decomposition

- Underlying data generating function $h(x)$
- Data $y = h(x) + \epsilon$
- Estimated model $f(x)$

Bias-Variance Decomposition

- Underlying data generating function $h(x)$
- Data $y = h(x) + \epsilon$
- Estimated model $f(x)$

What is the expected value of the least-squares loss?

$$\mathbb{E}[\mathcal{L}] = \mathbb{E}[(y - f)^2] \tag{1}$$

Bias-Variance Decomposition

We first expand the square

$$\mathbb{E}[\mathcal{L}] = \mathbb{E}[(y - f)^2] \tag{2}$$
$$= \mathbb{E}[y^2] + \mathbb{E}[f^2] - 2\mathbb{E}[yf] \tag{3}$$

Bias-Variance Decomposition

We first expand the square

$$\begin{aligned} \mathbb{E}[\mathcal{L}] &= \mathbb{E}[(y-f)^2] \quad (2)\\ &= \mathbb{E}[y^2] + \mathbb{E}[f^2] - 2\mathbb{E}[yf] \quad (3) \end{aligned}$$

The variance of a random variable is:

$$\text{var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2, \quad (4)$$

and for independent variables X and Y

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \quad (5)$$

Bias-Variance Decomposition

We first expand the square

$$\begin{aligned} \mathbb{E}[\mathcal{L}] &= \mathbb{E}[(y-f)^2] && (2) \\ &= \mathbb{E}[y^2] + \mathbb{E}[f^2] - 2\mathbb{E}[yf] && (3) \end{aligned}$$

The variance of a random variable is:

$$\mathrm{var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2, \tag{4}$$

and for independent variables X and Y

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \tag{5}$$

This allows us to rewrite the loss as

$$\mathbb{E}[\mathcal{L}] = \mathrm{var}[y] + (\mathbb{E}[y])^2 + \mathrm{var}[f] + (\mathbb{E}[f])^2 - 2\mathbb{E}[y]\mathbb{E}[f] \tag{6}$$

Bias-Variance Decomposition

- Recall $y = h(x) + \epsilon$
- Noise distribution: $\mathbb{E}[\epsilon] = 0$ and $\text{var}[\epsilon] = \sigma^2$
- So $\mathbb{E}[y] = h$ and $\text{var}[y] = \sigma^2$.

Bias-Variance Decomposition

- Recall $y = h(x) + \epsilon$
- Noise distribution: $\mathbb{E}[\epsilon] = 0$ and $\mathrm{var}[\epsilon] = \sigma^2$
- So $\mathbb{E}[y] = h$ and $\mathrm{var}[y] = \sigma^2$.

The expected loss becomes

$$
\begin{aligned}
\mathbb{E}[\mathcal{L}] &= \sigma^2 + h^2 + \mathrm{var}[f] + (\mathbb{E}[f])^2 - 2h\mathbb{E}[f] && (7)\\
&= \sigma^2 + \mathrm{var}[f] + h^2 + (\mathbb{E}[f])^2 - 2h\mathbb{E}[f] && (8)\\
&= \sigma^2 + \underbrace{\mathrm{var}[f]}_{\text{variance}} + \underbrace{(h - \mathbb{E}[f])^2}_{\text{bias}} && (9)
\end{aligned}
$$

Interpretation

- How can we interpret this result?
- Only contribution from data y is its variance σ^2.
- All dependency on the *specific sample*, y, of the data has been absorbed into the other terms.
- The variance of f is a consequence of the variance in the data
  - No noise $\rightarrow$ always learn the same model
  - Noisy samples $\rightarrow$ different models.
  - $\operatorname{var} f$ is sensitivity of learned model to the choice of data.

Interpretation

- $h(x) - \mathbb{E}[f(x)]$ is the ability of the estimated model to accurately represent the true model
- It is the *bias* of the estimate.
- Fitting $f(x) = mx * c$ to $h(x) = \sin(2\pi x)$ has a high bias: cannot represent the data
- But it has a low variance: insensitive to particular data choice
- Loss minimisation requires simultaneous minimisation of both bias and variance
  - Models that both fit *and* generalise well
  - Nearly always in conflict
- A fundamental limitation of machine learning.