Learning From Data
Lecture 11
Overfitting

What is Overfitting
When does Overfitting Occur
Stochastic and Deterministic Noise

Superstitions – Myth or Reality?

- Paraskevedekatriaphobia – fear of Friday the 13th.
  - Are future Friday the 13ths really more dangerous?

- OCD
  the subjects performs an action which leads to a good outcome and thereby generalizes it as cause and effect: the action will always give good results. Having overfit the data, the subject compulsively engages in that activity.

Humans are overfitting machines, very good at “finding coincidences”.

An Illustration of Overfitting on a Simple Example

Quadratic $f$
5 data points
A little noise (measurement error)
5 data points $\rightarrow$ 4th order polynomial fit

Superstitions – Myth or Reality?
• Paraskevedekatriaphobia – fear of Friday the 13th.
  - Are future Friday the 13ths really more dangerous?

• OCD
  the subjects performs an action which leads to a good outcome and thereby generalizes it as cause and effect: the action will always give good results. Having overfit the data, the subject compulsively engages in that activity.

Humans are overfitting machines, very good at “finding coincidences”.

An Illustration of Overfitting on a Simple Example

Quadratic $f$
5 data points
A little noise (measurement error)
5 data points $\rightarrow$ 4th order polynomial fit
An Illustration of Overfitting on a Simple Example

Quadratic $f$
5 data points
A little noise (measurement error)
5 data points $\rightarrow$ 4th order polynomial fit

Classic overfitting: simple target with excessively complex $\mathcal{H}$
$E_{in} \approx 0$, $E_{out} \gg 0$

The noise did us in. (why?)

Overfitting is Not Just Bad Generalization

VC Analysis:
Covers bad generalization but with lots of slack – the VC bound is loose

Overfitting:
Going for lower and lower $E_{in}$ results in higher and higher $E_{out}$
Case Study: 2nd vs 10th Order Polynomial Fit

10th order $f$ with noise.

50th order $f$ with no noise.

$H_2$: 2nd order polynomial fit

$H_{10}$: 10th order polynomial fit

←− special case of linear models with feature transform $x \mapsto (1, x, x^2, \cdots)$.

Which model do you pick for which problem and why?

Overfitting: 9/24

H_2 wins for both cases

Case Study: 2nd vs 10th Order Polynomial Fit

Is there Really “No Noise” with the Complex $f$?

Simple $f$ with noise.

Complex $f$ with no noise.

Go figure:

Simpler $H$ is better even for the more complex target with no noise.

Overfitting: 10/24

Look only at the data

Is there really “no noise” →

Overfitting: 11/24

Is there really “no noise” →

Overfitting: 12/24

Is there really “no noise” →
Is there Really “No Noise” with the Complex $f$?

Simple $f$ with noise.

Complex $f$ with no noise.

$H$ should match quantity and quality of data, not $f$.

Overfit Measure: $E_{out}(H_{10}) - E_{out}(H_{2})$

When is $H_{2}$ Better than $H_{10}$?

Overfitting: $E_{out}(H_{10}) > E_{out}(H_{2})$
Noise

That part of \( y \) we \textit{cannot} model

\[ \text{it has two sources} \ldots \]

**Deterministic Noise — Model Error**

We would like to learn from \( \mathcal{O} \):

\[ y_n = h^*(x_n) \]

Unfortunately, we only observe \( \mathcal{O} \):

\[ y_n = f(x_n) = h^*(x_n) + \text{‘deterministic noise’} \]

\[ \text{We cannot model this} \]

\[ \text{Deterministic Noise: the part of } f \text{ we cannot model.} \]

**Stochastic Noise — Data Error**

We would like to learn from \( \mathcal{O} \):

\[ y_n = f(x_n) \]

Unfortunately, we only observe \( \mathcal{O} \):

\[ y_n = f(x_n) + \text{‘stochastic noise’} \]

\[ \text{no one can model this} \]

**Stochastic Noise**: fluctuations/measurement errors we cannot model.

**Stochastic & Deterministic Noise Hurt Learning**

\[ y = f(x) \]

\[ \text{source: random measurement errors} \]

\[ \text{re-measure } y_n \text{, stochastic noise changes.} \]

\[ \text{change } \mathcal{H} \text{, stochastic noise the same.} \]

\[ \text{source: learner’s } \mathcal{H} \text{ cannot model } f \]

\[ \text{re-measure } y_n \text{, deterministic noise the same.} \]

\[ \text{change } \mathcal{H} \text{, deterministic noise changes.} \]

\[ \text{We have single } \mathcal{D} \text{ and fixed } \mathcal{H} \text{ so we cannot distinguish} \]
Noise and the Bias-Variance Decomposition

\[ y = f(x) + \epsilon \]

\[ \text{measurement error} \]

\[ \mathbb{E}[E_{\text{out}}(x)] = \mathbb{E}_{D,\epsilon}[(g(x) - f(x) - \epsilon)^2] \]

\[ = \mathbb{E}_{D,\epsilon}[(g(x) - f(x))^2 + 2(g(x) - f(x))\epsilon + \epsilon^2] \]

\[ \downarrow \]

bias + var

\[ 0 \]

\[ \sigma^2 \]

Noise is the Culprit

Overfitting is the disease

Noise is the cause

Learning is led astray by fitting the noise more than the signal

Cures

Regularization: Putting on the brakes.

Validation: A reality check from peeking at \( E_{\text{out}} \) (the bottom line).

Regularization

no regularization

regularization!

- Data
- Target
- Fit