Linear Models
Linear Models
Linear Models?
Linear Models?
Nonlinear Models
Feature Transformation: Model

\[ h(\tilde{x}) = f(\overrightarrow{w}^T \tilde{x}) \]

\[ \tilde{h}(\tilde{z}) = f(\overrightarrow{w}^T \tilde{z}) \text{ where } \tilde{z} = \Phi(\tilde{x}) \]
Feature Transformation: Example

\[ h(\tilde{x}) = f(\tilde{w}^T \tilde{x}) \]

\[ \tilde{h}(\tilde{z}) = f(\tilde{w}^T \tilde{z}) \text{ where } \tilde{z} = [(x_1 - 0.5)^2, (x_2 - 0.5)^2] \]
Nonlinear Models

\[ x^2 \]

\[ x_1 \]
Nonlinear Models

\[ z_1 = (x_1 - 0.5)^2 \]

\[ z_2 = (x_2 - 0.5)^2 \]
Nonlinear Models

$$z_2 = (x_2 - 0.5)^2$$

$$z_1 = (x_1 - 0.5)^2$$
Nonlinear Models
Nonlinear Models

- Decide on a transformation $\Phi: \mathcal{X} \rightarrow \mathcal{Z}$
- Convert $\mathcal{D} = \{(\overrightarrow{x_1}, y_1), \ldots, (\overrightarrow{x_n}, y_n)\}$ to $\widetilde{\mathcal{D}} = \{(\Phi(\overrightarrow{x_1}) = \overrightarrow{z_1}, y_1), \ldots, (\Phi(\overrightarrow{x_n}) = \overrightarrow{z_n}, y_n)\}$
- Fit a linear model using $\widetilde{\mathcal{D}}, \tilde{g}(\overrightarrow{z})$
- Return the corresponding predictor in the original space: $g(\overrightarrow{x}) = \tilde{g}(\Phi(\overrightarrow{x}))$
... But at what cost?

- VC dimension of linear separators: $d_{VC} = d + 1$ where $d = \text{the dimensionality of the input space } \mathcal{X}$

- Let $\tilde{d} = \text{the dimensionality of the transformed space } \mathcal{Z}$: if $\tilde{d} \gg d$, then the learned hypothesis will not generalize well:

$$E_{out}(g) \leq E_{in}(g) + O\left(\sqrt{(d + 1) \frac{\log(n)}{n}}\right)$$

vs.

$$E_{out}(g) \leq E_{in}(g) + O\left(\sqrt{(\tilde{d} + 1) \frac{\log(n)}{n}}\right)$$
• But what if $\tilde{d} = d$?

• From the example: $\tilde{x} = [x_1, x_2]$ so $d = 2$ and $\tilde{z} = [(x_1 - 0.5)^2, (x_2 - 0.5)^2]$ so $\tilde{d} = 2$.

• Data snooping: looking at the training data to decide what transformation to use.

• Pick a transformation before looking at training data.
Nonlinear Models

- Decide on a transformation $\Phi : \mathcal{X} \rightarrow \mathcal{Z}$
- Convert $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ to
  $\tilde{\mathcal{D}} = \{(\Phi(x_1) = z_1, y_1), \ldots, (\Phi(x_n) = z_n, y_n)\}$
- Fit a linear model using $\tilde{\mathcal{D}}, \tilde{g}(\tilde{z})$
- Return the corresponding predictor in the original space: $g(\tilde{x}) = \tilde{g}(\Phi(\tilde{x}))$
General \( k^{th} \)-order Transforms

- \( \Phi_2 ([x_1, x_2]) = [x_1, x_2, x_1^2, x_1 x_2, x_2^2] \)
- \( \Phi_3 ([x_1, x_2]) = [x_1, x_2, x_1^2, x_1 x_2, x_2^2, x_1^3, x_1^2 x_2, x_1 x_2^2, x_2^3] \)
- \( \Phi_4 ([x_1, x_2]) = [x_1, x_2, x_1^2, x_1 x_2, x_2^2, x_1^3, x_1^2 x_2, x_1 x_2^2, x_2^3, x_1^4, x_1^3 x_2, x_1^2 x_2^2, x_1 x_2^3, x_2^4] \)

- \( \Phi_Q \) maps a 2-d vector to a \( \frac{Q(Q+3)}{2} \)-d output
Linear Models
Nonlinear Models?
## Tradeoffs

<table>
<thead>
<tr>
<th></th>
<th>Low-Dimensional Input Space</th>
<th>High-Dimensional Input Space</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{in}$</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Generalization</td>
<td>Good</td>
<td>Bad</td>
</tr>
</tbody>
</table>

Overfitting
Overfitting

- Overfitting is fitting the training data “more than is warranted”
- Fitting noise rather than signal
- Fitting “outliers”
Experimental Setup

- $\mathcal{X} = \mathbb{R}, \mathcal{Y} = \mathbb{R}$ and $n = 20$

- $f$ is a $10^{\text{th}}$-order polynomial in $x$ with additive Gaussian noise

\[
y = \sum_{d=0}^{10} a_d x^d + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2)
\]

- $\mathcal{H}_2 = 2^{\text{nd}}$-order polynomials
  - $\vec{z} = \Phi_2(x) = [x, x^2]$

- $\mathcal{H}_{10} = 10^{\text{th}}$-order polynomials
  - $\vec{z} = \Phi_{10}(x) = [x, x^2, x^3, x^4, x^5, x^6, x^7, x^8, x^9, x^{10}]$
Break

- 10-d target function with additive Gaussian noise
- \( y = f(x) + \epsilon \) where \( \epsilon \sim N(0, \sigma^2) \)
- \( \mathcal{H}_2 = 2^{\text{nd}} \)-order polynomial
- \( \mathcal{H}_{10} = 10^{\text{th}} \)-order polynomial

Which hypothesis set do you expect to have a lower out-of-sample error?

- 2nd order polynomials
- 10th order polynomials
Noisy Targets

- 10-d target function with additive Gaussian noise
- \( y = f(x) + \epsilon \) where \( \epsilon \sim N(0, \sigma^2) \)
- \( \mathcal{H}_2 = 2^{\text{nd}}\)-order polynomial
- \( \mathcal{H}_{10} = 10^{\text{th}}\)-order polynomial
Noisy Targets

- 10-d target function with additive Gaussian noise
- $y = f(x) + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$
- $\mathcal{H}_2 = 2^{\text{nd}}$-order polynomial
- $\mathcal{H}_{10} = 10^{\text{th}}$-order polynomial
Noisy Targets

- 10-d target function with additive Gaussian noise
- \( y = f(x) + \epsilon \) where \( \epsilon \sim N(0, \sigma^2) \)
- \( \mathcal{H}_2 = 2^{nd}\)-order polynomial
- \( \mathcal{H}_{10} = 10^{th}\)-order polynomial
Noisy Targets

- 10-d target function with additive Gaussian noise
- \( y = f(x) + \epsilon \) where \( \epsilon \sim N(0, \sigma^2) \)
- \( H_2 = 2^{nd} \)-order polynomial
- \( H_{10} = 10^{th} \)-order polynomial
Noisy Targets

<table>
<thead>
<tr>
<th>$\mathcal{H}_2$</th>
<th>$\mathcal{H}_10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{in}$</td>
<td>0.016</td>
</tr>
<tr>
<td>$E_{out}$</td>
<td>0.009</td>
</tr>
</tbody>
</table>

$\mathcal{H}_2$ and $\mathcal{H}_{10}$ represent different hypotheses, and $E_{in}$ and $E_{out}$ are the in-sample and out-of-sample errors, respectively.
Simple model

Complex model

\[ E_{\text{out}} \]

\[ E_{\text{in}} \]

Number of training points, \( n \)
Bias-Variance Tradeoff

\( n = 2 \)

Bias of \( \bar{g}(\tilde{x}) \) \( \approx 0.50 \)

Variance of \( g_D(\tilde{x}) \) \( \approx 0.25 \)

\( \mathbb{E}_D[E_{out}(g_D)] \approx 0.75 \)

Bias of \( \bar{g}(\tilde{x}) \) \( \approx 0.21 \)

Variance of \( g_D(\tilde{x}) \) \( \approx 1.74 \)

\( \mathbb{E}_D[E_{out}(g_D)] \approx 1.95 \)
Bias-Variance Tradeoff

\( n = 5 \)

Bias of \( \bar{g}(\bar{x}) \) \( \approx 0.50 \)

Variance of \( g_\mathcal{D}(\bar{x}) \) \( \approx 0.10 \)

\( \mathbb{E}_\mathcal{D}[E_{out}(g_\mathcal{D})] \approx 0.60 \)

Bias of \( g_\mathcal{D}(\bar{x}) \) \( \approx 0.21 \)

Variance of \( g_\mathcal{D}(\bar{x}) \) \( \approx 0.21 \)

\( \mathbb{E}_\mathcal{D}[E_{out}(g_\mathcal{D})] \approx 0.42 \)
Experimental Setup

• \( \mathcal{X} = \mathbb{R}, \mathcal{Y} = \mathbb{R} \) and \( n = 100 \)

• \( f \) is a 10\textsuperscript{th}-order polynomial in \( x \) with additive Gaussian noise

\[
y = \sum_{d=0}^{10} a_d x^d + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2)
\]

• \( \mathcal{H}_2 = 2\textsuperscript{nd}-order polynomials
  • \( \vec{z} = \Phi_2(x) = [x, x^2] \)

• \( \mathcal{H}_{10} = 10\textsuperscript{th}-order polynomials
  • \( \vec{z} = \Phi_{10}(x) = [x, x^2, x^3, x^4, x^5, x^6, x^7, x^8, x^9, x^{10}] \)
Noisy Targets

- 10-d target function with additive Gaussian noise
  - $y = f(x) + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$
- $\mathcal{H}_2 = 2^{\text{nd}}$-order polynomial
- $\mathcal{H}_{10} = 10^{\text{th}}$-order polynomial
Noisy Targets

- 10-d target function with additive Gaussian noise
- \( y = f(x) + \epsilon \) where \( \epsilon \sim N(0, \sigma^2) \)
- \( \mathcal{H}_2 = 2^{nd}\)-order polynomial
- \( \mathcal{H}_{10} = 10^{th}\)-order polynomial
Noisy Targets

<table>
<thead>
<tr>
<th></th>
<th>$\mathcal{H}_2$</th>
<th>$\mathcal{H}_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{in}$</td>
<td>0.018</td>
<td>0.010</td>
</tr>
<tr>
<td>$E_{out}$</td>
<td>0.009</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Target Function

2nd-Order Hypothesis

10th-Order Hypothesis

Noisy Samples
Noiseless Targets

- 50-d target function with no noise
- \( y = \sum_{d=0}^{50} a_d x^d \)
- \( \mathcal{H}_2 = 2^{\text{nd}}\)-order polynomial
- \( \mathcal{H}_{10} = 10^{\text{th}}\)-order polynomial
Noiseless Targets

- 50-d target function with no noise
  \[ y = \sum_{d=0}^{50} a_d x^d \]
- \( \mathcal{H}_2 = 2^{nd}\)-order polynomial
- \( \mathcal{H}_{10} = 10^{th}\)-order polynomial
Noiseless Targets

<table>
<thead>
<tr>
<th>$\mathcal{H}_2$</th>
<th>$\mathcal{H}_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{in}$</td>
<td>0.003</td>
</tr>
<tr>
<td>$E_{out}$</td>
<td>0.004</td>
</tr>
</tbody>
</table>

- $E_{in}$: 0.003
- $E_{out}$: 0.004

Target Function

- $2^{nd}$-Order Hypothesis
- $10^{th}$-Order Hypothesis

- Noiseless Samples

Graph showing the comparison of $2^{nd}$-Order and $10^{th}$-Order hypotheses with noiseless samples.
Two Types of Noise

- Stochastic noise
  - Measurement error
  - Random
  - Not affected by choice of $\mathcal{H}$

- Deterministic noise
  - Limitation of $\mathcal{H}$
  - Not random
  - Dependent on $\mathcal{H}$ and $f$

- Given a single dataset $\mathcal{D}$ and a fixed $\mathcal{H}$, the two types of noise are indistinguishable
<table>
<thead>
<tr>
<th></th>
<th>Direction</th>
<th>Overfitting</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of points</strong></td>
<td>↑</td>
<td>↓</td>
</tr>
<tr>
<td></td>
<td>↓</td>
<td>↑</td>
</tr>
<tr>
<td><strong>Stochastic noise</strong></td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td></td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td><strong>Deterministic noise</strong></td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td></td>
<td>↓</td>
<td>↓</td>
</tr>
</tbody>
</table>