CSE 417T: Introduction to Machine Learning

Lecture 10: Overfitting

Henry Chai
09/27/18
Recall

- Decide on a transformation $\Phi: \mathcal{X} \rightarrow \mathcal{Z}$

- Convert $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ to $\tilde{\mathcal{D}} = \{(\Phi(x_1) = \frac{x_1}{z_1}, y_1), \ldots, (\Phi(x_n) = \frac{x_n}{z_n}, y_n)\}$

- Fit a linear model using $\tilde{\mathcal{D}}, \tilde{g}(\tilde{z})$

- Return the corresponding predictor in the original space: $g(\tilde{x}) = \tilde{g}(\Phi(\tilde{x}))$
Nonlinear Transforms for Approving Credit

- Input: \( x_1 = \text{age}, x_2 = \text{income}, x_3 = \text{credit score} \)

- How close is this person to some optimal age, \( a^* \)?

- What is this person’s income scaled by age?

- Having a very low credit score is more significant than having a very high credit score.

- Transformation: \( \Phi(\vec{x}) = [x_1, x_2, x_3, |a^* - x_1|, \frac{x_2}{x_1}, \sqrt{x_3}] \)
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
\begin{itemize}
  \item $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \mathbb{R}$ and $n = 20$
  \item $f$ is a $10^{\text{th}}$-order polynomial in $x$ with additive Gaussian noise
    \begin{equation*}
      y = \sum_{d=0}^{10} a_d x^d + \epsilon \quad \text{where} \quad \epsilon \sim N(0, \sigma^2)
    \end{equation*}
  \item $\mathcal{H}_2 = 2^{\text{nd}}$-order polynomials
    \begin{itemize}
      \item $\tilde{z} = \Phi_2(x) = [x, x^2]$
    \end{itemize}
  \item $\mathcal{H}_{10} = 10^{\text{th}}$-order polynomials
    \begin{itemize}
      \item $\tilde{z} = \Phi_{10}(x) = [x, x^2, x^3, x^4, x^5, x^6, x^7, x^8, x^9, x^{10}]$
    \end{itemize}
\end{itemize}
Noisy Targets

- 10-d target function with additive Gaussian noise

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

- \( \mathcal{H}_2 = 2^{nd}\)-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>0.016</td>
<td>0.011</td>
</tr>
<tr>
<td>$E_{out}$</td>
<td>0.009</td>
<td>3797</td>
</tr>
</tbody>
</table>

Target Function

- 2\textsuperscript{nd}-Order Hypothesis
- 10\textsuperscript{th}-Order Hypothesis
- Noisy Samples
Simple model

Complex model

Number of training points, $n$

$E_{out}$

$E_{in}$

Expected error

Expected error
Bias-Variance Tradeoff (Example)

Bias of $\tilde{g}(\tilde{x}) \approx 0.50$
Variance of $g_D(\tilde{x}) \approx 0.25$
$\mathbb{E}_D[E_{out}(g_D)] \approx 0.75$

Bias of $\tilde{g}(\tilde{x}) \approx 0.21$
Variance of $g_D(\tilde{x}) \approx 1.74$
$\mathbb{E}_D[E_{out}(g_D)] \approx 1.95$
Experimental Setup

- $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \mathbb{R}$ and $n = 100$

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

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

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

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

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

<table>
<thead>
<tr>
<th>$H_2$</th>
<th>$H_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{\text{in}}$</td>
<td>0.003</td>
</tr>
<tr>
<td>$E_{\text{out}}$</td>
<td>0.004</td>
</tr>
</tbody>
</table>

![Graph showing the target function and hypothesis orders.](image-url)
Two Types of Noise

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

- Deterministic noise
  - Limitations 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>Number of points</td>
<td>↑</td>
<td>↓</td>
</tr>
<tr>
<td></td>
<td>↓</td>
<td>↑</td>
</tr>
<tr>
<td>Stochastic noise</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td></td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>Deterministic noise</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td></td>
<td>↓</td>
<td>↓</td>
</tr>
</tbody>
</table>
\[ \mathbb{E}_D [E_{out}(g_D)] = \mathbb{E}_D [\mathbb{E}_{\tilde{x}} [(g_D(\tilde{x}) - y)^2]] \text{ where } y = f(\tilde{x}) + \epsilon \]

\[ = \mathbb{E}_{\tilde{x}} [\text{Variance of } g_D(\tilde{x})] + \mathbb{E}_{\tilde{x}} [\text{Bias of } g(\tilde{x})] + \text{Stochastic noise} \]
\[ \mathbb{E}_D[E_{out}(g_D)] = \mathbb{E}_D[\mathbb{E}_{\tilde{x}}[ (g_D(\tilde{x}) - y)^2] \] where \( y = f(\tilde{x}) + \epsilon \)

\[
\begin{align*}
&= \mathbb{E}_{\tilde{x}}[\text{Variance of } g_D(\tilde{x})] \\
&+ \mathbb{E}_{\tilde{x}}[\text{Deterministic noise of } g(\tilde{x})] \\
&+ \text{Stochastic noise}
\end{align*}
\]