ECS171: Machine Learning
Lecture 12: Overfitting, Regularization

Cho-Jui Hsieh
UC Davis

Feb 26, 2018
What is overfitting?
Illustration of overfitting

- Example: neural network
- Overfitting: $E_{\text{in}}$ decreases while $E_{\text{out}}$ increases
Illustration of overfitting

- Example: neural network
- Overfitting: $E_{\text{in}}$ decreases while $E_{\text{out}}$ increases

Overfitting: “fitting the data more than is warranted”
Fitting the noise, not only useless but also harmful
Two fits for each target

**Noisy low-order target**

<table>
<thead>
<tr>
<th></th>
<th>2nd Order</th>
<th>10th Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{in}$</td>
<td>0.050</td>
<td>0.034</td>
</tr>
<tr>
<td>$E_{out}$</td>
<td>0.127</td>
<td>9.00</td>
</tr>
</tbody>
</table>

**Noiseless high-order target**

<table>
<thead>
<tr>
<th></th>
<th>2nd Order</th>
<th>10th Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{in}$</td>
<td>0.029</td>
<td>$10^{-5}$</td>
</tr>
<tr>
<td>$E_{out}$</td>
<td>0.120</td>
<td>7680</td>
</tr>
</tbody>
</table>
Overfitting
A Detailed Experiment

- Generate data by \( y = f(x) + \epsilon = \sum_{q=0}^{Q_f} \alpha_q x^q + \epsilon \)
  - noise level: \( \sigma^2 \)
  - target complexity: \( Q_f \)
  - data set size: \( N \)

- Fit the data set \((x_1, y_1), \ldots, (x_N, y_N)\) using our two models
  - \( H_2 \): 2nd-order polynomials
  - \( H_{10} \): 10th-order polynomials

- Compare \( E_{\text{out}}(g_{10}) - E_{\text{out}}(g_2) \)
Impact of “noise”

Stochastic noise

Deterministic noise

number of data points \[\uparrow\] Overfitting \[\downarrow\]

stochastic noise \[\uparrow\] Overfitting \[\uparrow\]
deterministic noise \[\uparrow\] Overfitting \[\uparrow\]
The part of $f$ that $\mathcal{H}$ cannot capture: $f(x) - h^*(x)$

Difference to stochastic noise:
- depends on $\mathcal{H}$
- fixed for a given $x$
Two cures

- Regularization: Putting the brakes
- Validation (next class)
Regularization
The polynomial model

- $\mathcal{H}_Q$: polynomials of order $Q$

\[ \mathcal{H}_Q = \left\{ \sum_{q=0}^{Q} w_q L_q(x) \right\} \]

- Linear regression in the $\mathcal{Z}$ space with

\[ z = [1, L_1(x), \ldots, L_Q(x)] \]

Legendre polynomials:

- $L_1 = x$
- $L_2 = \frac{1}{2}(3x^2 - 1)$
- $L_3 = \frac{1}{2}(5x^3 - 3x)$
- $L_4 = \frac{1}{8}(35x^4 - 30x^2 + 3)$
- $L_5 = \frac{1}{8}(63x^5 \ldots)$
Unconstrained solution

- Input \((x_1, y_1), \ldots, (x_N, y_N) \rightarrow (z_1, y_1), \ldots, (z_N, y_N)\)

- Linear regression:

  \[
  \text{Minimize } E_{\text{in}}(w) = \frac{1}{N} \sum_{n=1}^{N} (w^T z_n - y_n)^2
  \]

  \[
  \text{Minimize } \frac{1}{N} (Zw - y)^T (Zw - y)
  \]

- Solution \(w_{\text{lin}} = (Z^T Z)^{-1} Z^T y\)
Hard constraint: $\mathcal{H}_2$ is constrained version of $\mathcal{H}_{10}$
(with $w_q = 0$ for $q > 2$)
Constraining the weights

- Hard constraint: $\mathcal{H}_2$ is constrained version of $\mathcal{H}_{10}$
  (with $w_q = 0$ for $q > 2$)
- Soft-order constraint: $\sum_{q=0}^{Q} w_q^2 \leq C$
Constraining the weights

- Hard constraint: $\mathcal{H}_2$ is constrained version of $\mathcal{H}_{10}$
  
  (with $w_q = 0$ for $q > 2$)

- Soft-order constraint: $\sum_{q=0}^{Q} w_q^2 \leq C$

- The problem given soft-order constraint:

  $$\text{Minimize} \frac{1}{N} (Z\mathbf{w} - \mathbf{y})^T(Z\mathbf{w} - \mathbf{y}) \quad \text{s.t.} \quad \mathbf{w}^T\mathbf{w} \leq C$$

  smaller hypothesis space

- Solution $\mathbf{w}_{\text{reg}}$ instead of $\mathbf{w}_{\text{lin}}$
Equivalent to the unconstrained version

- Constrained version:

\[
\min_w E_{\text{in}}(w) = \frac{1}{N} (Zw - y)^T (Ze - y) \quad \text{s.t. } w^T w \leq C
\]

- Optimal when

\[
\nabla E_{\text{in}}(w_{\text{reg}}) \propto -w_{\text{reg}}
\]

Why? If \(-\nabla E_{\text{in}}(w)\) and \(w\) are not parallel, can decrease \(E_{\text{in}}(w)\) without violating the constraint.
Equivalent to the unconstrained version

- Constrained version:

\[
\min_w E_{\text{in}}(w) = \frac{1}{N} (Zw - y)^T (Ze - y) \quad \text{s.t.} \quad w^T w \leq C
\]

- Optimal when

\[
\nabla E_{\text{in}}(w_{\text{reg}}) \propto -w_{\text{reg}}
\]

- Assume \( \nabla E_{\text{in}}(w_{\text{reg}}) = -2\frac{\lambda}{N} w_{\text{reg}} \)

\[\Rightarrow \nabla E_{\text{in}}(w_{\text{reg}}) + 2\frac{\lambda}{N} w_{\text{reg}} = 0\]
Equivalent to the unconstrained version

- **Constrained version:**
  \[
  \min_w E_{\text{in}}(w) = \frac{1}{N}(Zw - y)^T(Ze - y) \quad \text{s.t.} \quad w^T w \leq C
  \]

- **Optimal when**
  \[
  \nabla E_{\text{in}}(w_{\text{reg}}) \propto -w_{\text{reg}}
  \]

- **Assume**
  \[
  \nabla E_{\text{in}}(w_{\text{reg}}) = -2\frac{\lambda}{N}w_{\text{reg}}
  \]
  \[
  \Rightarrow \nabla E_{\text{in}}(w_{\text{reg}}) + 2\frac{\lambda}{N}w_{\text{reg}} = 0
  \]

- **\(w_{\text{reg}}\)** is also the solution of **unconstrained problem**

  \[
  \min_w E_{\text{in}}(w) + \frac{\lambda}{N}w^T w
  \]

  (Ridge regression!)
Equivalent to the unconstrained version

- Constrained version:
  \[
  \min_w E_{\text{in}}(w) = \frac{1}{N} (Zw - y)^T (Ze - y) \quad \text{s.t.} \quad w^T w \leq C
  \]

- Optimal when
  \[
  \nabla E_{\text{in}}(w_{\text{reg}}) \propto -w_{\text{reg}}
  \]

- Assume \( \nabla E_{\text{in}}(w_{\text{reg}}) = -2 \frac{\lambda}{N} w_{\text{reg}} \)
  \[
  \Rightarrow \nabla E_{\text{in}}(w_{\text{reg}}) + 2 \frac{\lambda}{N} w_{\text{reg}} = 0
  \]

- \( w_{\text{reg}} \) is also the solution of unconstrained problem

  \[
  \min_w E_{\text{in}}(w) + \frac{\lambda}{N} w^T w
  \]

  (Ridge regression!)

  \( C \uparrow \lambda \downarrow \)
Ridge regression solution

\[
\min_w E_{\text{aug}}(w) = \frac{1}{N} \left( (Zw - y)^T(Zw - y) + \lambda w^T w \right)
\]

\[\nabla E_{\text{aug}}(w) = 0 \implies Z^T Z (w - y) + \lambda w = 0
\]
Ridge regression solution

$$\min_w E_{\text{aug}}(w) = \frac{1}{N} \left( (Zw - y)^T (Zw - y) + \lambda w^T w \right)$$

$$\nabla E_{\text{aug}}(w) = 0 \ \Rightarrow \ \ Z^T Z(w - y) + \lambda w = 0$$

So, $$w_{\text{reg}} = (Z^T Z + \lambda I)^{-1} Z^T y \quad \text{(with regularization)}$$

as opposed to $$w_{\text{lin}} = (Z^T Z)^{-1} Z^T y \quad \text{(without regularization)}$$
The result

$$\min_w E_{\text{in}}(w) + \frac{\lambda}{N} w^T w$$
Equivalent to “weight decay”

Consider the general case

$$\min_w E_{\text{in}}(w) + \frac{\lambda}{N} w^T w$$
Consider the general case

\[ \min_w E_{\text{in}}(w) + \frac{\lambda}{N} w^T w \]

Gradient descent:

\[ w_{t+1} = w_t - \eta \left( \nabla E_{\text{in}}(w_t) + 2\frac{\lambda}{N} w_t \right) \]

\[ = w_t \left( 1 - 2\eta \frac{\lambda}{N} \right) - \eta \nabla E_{\text{in}}(w_t) \]

weight decay
Variations of weight decay

- Emphasis of certain weights:

\[
\sum_{q=0}^{Q} \gamma_q w_q^2
\]

- Example 1: \(\gamma_q = 2^q\) \(\Rightarrow\) low-order fit
- Example 2: \(\gamma_q = 2^{-q}\) \(\Rightarrow\) high-order fit
Variations of weight decay

- Emphasis of certain weights:

\[ \sum_{q=0}^{Q} \gamma_q w_q^2 \]

- Example 1: \( \gamma_q = 2^q \Rightarrow \) low-order fit
- Example 2: \( \gamma_q = 2^{-q} \Rightarrow \) high-order fit

- General Tikhonov regularizer:

\[ w^T H w \]

with a positive semi-definite \( H \)
Calling the regularizer $\Omega = \Omega(h)$, we minimize

$$E_{\text{aug}}(h) = E_{\text{in}}(h) + \frac{\lambda}{N} \Omega(h)$$
General form of regularizer

- Calling the regularizer $\Omega = \Omega(h)$, we minimize

$$E_{\text{aug}}(h) = E_{\text{in}}(h) + \frac{\lambda}{N}\Omega(h)$$

- Recall our VC bound:

$$E_{\text{out}}(h) \leq E_{\text{in}}(h) + \Omega(\mathcal{H})$$
General form of regularizer

- Calling the regularizer $\Omega = \Omega(h)$, we minimize

$$E_{\text{aug}}(h) = E_{\text{in}}(h) + \frac{\lambda}{N}\Omega(h)$$

- Recall our VC bound:

$$E_{\text{out}}(h) \leq E_{\text{in}}(h) + \Omega(\mathcal{H})$$

- $E_{\text{aug}}$ is better than $E_{\text{in}}$ as a proxy for $E_{\text{out}}$
General regularizers $\Omega(h)$

- stochastic noise: high-frequency, non-smooth
- deterministic noise: non-smooth (higher-order terms)
General regularizers $\Omega(h)$

- stochastic noise: high-frequency, non-smooth
- deterministic noise: non-smooth (higher-order terms)

Regularization:
Constraint in the “direction” of the target function
General regularizers $\Omega(h)$

- stochastic noise: high-frequency, non-smooth
- deterministic noise: non-smooth (higher-order terms)

Regularization:
- Constraint in the “direction” of the target function

General principal:
- Plausible: direction towards smoother or simpler models
- Friendly: easy to optimize
L2 vs L1 regularizer

- **L1-regularizer:** $\Omega(w) = \|w\|_1 = \sum_q |w_q|$
- usually leads to a **sparse solution**
  (only few $w_q$ will be nonzero)
Conclusions

- Next class: LFD 4.3

Questions?