ECE595 / STAT598: Machine Learning I
Lecture 31.2: Regularization - Two Regularization Techniques

Spring 2020
Stanley Chan

School of Electrical and Computer Engineering
Purdue University
Outline

- Lecture 30 Overfit
- Lecture 31 Regularization
- Lecture 32 Validation

Today's Lecture:

- Motivation for Regularization
  - VC Analysis
  - Example
- Two Regularization Techniques
  - Weight Decay
  - Augmented Error
- Choosing a Regularization
  - Pill or Poisson?
  - Role of $\lambda$
Soft Order Constraint

Consider the following example

- $\mathcal{H} = \text{set of polynomials in one variable } x \in [-1, 1].$
- E.g., $h(x) = 2x^2 + 3x + 7.$
- Want to express $h(x)$ using basis function.
- Basis functions for polynomials are Legendre polynomials $L_q(x)$, $q = 1, 2, \ldots$ 
- So, any $h(x)$ can be expressed as

$$h(x) = \sum_{q=1}^{Q} w_q L_q(x)$$  \hspace{1cm} (2)
Soft Order Constraint

This model is indeed linear! (Why?)

- You define a nonlinear transform $\Phi$,

$$
\begin{bmatrix}
1 \\
L_1(x) \\
\vdots \\
L_Q(x)
\end{bmatrix}
$$

- The hypothesis set is

$$
\mathcal{H}_Q = \left\{ h \mid h(x) = w^T z = \sum_{q=0}^{Q} w_q L_q(x) \right\}
$$

- So now you can define training error (for linear regression) as

$$
E_{\text{in}}(w) = \frac{1}{N} \sum_{n=1}^{N} (w^T z_n - y_n)^2
$$
Soft Order Constraint

There are multiple ways of constraining the weights.

- **Hard** constraint:
  - Force coefficients to be zero.
  - For example,
    \[ \mathcal{H}_2 = \{ \mathbf{w} \mid \mathbf{w} \in \mathcal{H}_{10}; w_q = 0, \text{for} q \geq 3 \}. \]

- **Soft** constraint:
  - Force coefficients to be small.
  - For example,
    \[ \sum_{q=0}^{Q} w_q^2 \leq C \]

  - It encourages weights to be small without changing the order of the polynomial by explicitly forcing some weights to zero.
**VC Perspective of Soft Order Constraint**

- The optimization is

  \[
  \text{minimize } E_{\text{in}}(\mathbf{w}) \quad \text{subject to } \mathbf{w}^T \mathbf{w} \leq C
  \]

- We know \( E_{\text{in}}(\mathbf{w}) = \frac{1}{M} \| \mathbf{Z} \mathbf{w} - \mathbf{y} \|^2 \)

- The hypothesis set is

  \[
  \mathcal{H}(C) = \{ h \mid h(x) = \mathbf{w}^T \mathbf{z}, \mathbf{w}^T \mathbf{w} \leq C \}
  \]

- So the optimization is equivalent to minimize \( E_{\text{in}} \) over \( \mathcal{H}(C) \)

- If \( C_1 < C_2 \), then \( \mathcal{H}(C_1) \subset \mathcal{H}(C_2) \) and \( d_{\text{vc}}(\mathcal{H}(C_1)) \leq d_{\text{vc}}(\mathcal{H}(C_2)) \)

- So we should expect better generalization with \( \mathcal{H}(C_1) \)
Solving the Soft Order Constraint Problem

The optimization problem is

\[
\minimize_w \frac{1}{N} \| Z w - y \|^2 \quad \text{subject to} \quad w^T w \leq C
\]  

(4)

- Using Lagrangian techniques we can show that the minimization is equivalent to

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

for some choices of \( \lambda_C \).

- You can further change the constraint to

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

- \( \gamma_q = q \) or \( \gamma_q = e^q \) encourages a low-order fit

- \( \gamma_q = (1 + q)^{-1} \) or \( \gamma_q = e^{-q} \) encourages a high-order fit
Another type of regularization is **augmented error**

\[
E_{\text{aug}}(w) = E_{\text{in}}(w) + \lambda w^T w
\]  

- Unconstrained minimization is often easier than constrained minimization
- But you are paying the price of interpretability
- For a given \( C \), soft order constraint corresponds to selecting a hypothesis from a smaller set \( \mathcal{H}(C) \)
- VC analysis says we will get a better generalization when \( C \) decreases (but not too much)
- The optimal \( C \) is sum square magnitude we allow.
- For augmented error, you need to find the optimal parameter \( \lambda^* \)
- This is not very interpretable.
The augmented error for a hypothesis $h \in \mathcal{H}$ is

$$E_{\text{aug}}(h, \lambda, \Omega) = E_{\text{in}}(h) + \frac{\lambda}{N} \Omega(h)$$  \hspace{1cm} (6)$$

- Here, $\Omega(h) = w^T w$
- There are two components of the penalty:
  - The regularizer $\Omega(h)$ which penalizes a particular property of $h$
  - The regularization parameter $\lambda$ which controls the amount of regularization
- As $N$ increases, the need for regularization goes down
- This equation resembles VC bound
Choice of $\lambda$

Minimizing $E_{\text{in}}(w) + \frac{\lambda}{N} w^T w$ for different $\lambda$'s:

$\lambda = 0$

$\lambda = 0.0001$

$\lambda = 0.01$

$\lambda = 1$

overfitting $\rightarrow$ $\rightarrow$ $\rightarrow$ underfitting