Homework 3: Model Order Selection

Liheng Cao

March 8, 2021

1

(a) (i) No.

(ii) There is no underfitting because the power of the model is higher than the true function.

(iii) $\vec{w} = [1, 2, 0]$ (bold and arrow used interchangeably)

(b) (i) No.

(ii) There is underfitting because the power of the model is lower than the true function.

(c) (i) Yes. The true function has x_1, x_2, but, according to the footnote, our example only has 1 feature.

(ii) There is underfitting because the true function has the term $x_1 x_2$, which the model lacks.

2

(a)

$$x_0 = 1, x_1 = \text{cancer volume}, x_2 = \text{patient's age}, x_3 = \text{type of cancer}$$

x_0 let's the x, w subscripts line up more, which looks better.

(Model 1) $h(\mathbf{x}) = w_0 + w_1 x_1$

(Model 2) $h(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2$

(Model 3) $h(\mathbf{x}) = w_0 + (w_1 + x_3 w_3) x_1 + w_2 x_2$

(b) There are 2 and 3 parameters, respectively. Model 2 is the most complex out of the 2, however Model 3 is the most complex.

(c) (Model 1) $X = \begin{bmatrix} 1 & 0.7 \\ 1 & 1.3 \\ 1 & 1.6 \end{bmatrix}$

(Model 2) $X = \begin{bmatrix} 1 & 0.7 & 55 \\ 1 & 1.3 & 65 \\ 1 & 1.6 & 70 \end{bmatrix}$

(Model 3) $X = \begin{bmatrix} 1 & 0.7 & 55 & 1 \\ 1 & 1.3 & 65 & 0 \\ 1 & 1.6 & 70 & 0 \end{bmatrix}$

(d) Model 2 should be selected. It has similar numbers to Model 3, but Model 2 is a much simpler model. We want to avoid having an overly complicated model that doesn't even give us better numbers.

3

A is probably underfitting; this is because its error is much higher than the other models and because its error drops to around its final value at smaller sample sizes, implying a simpler model.

B is probably neither. If a model is overfitted, then its error would be extremely low for small N, and then much higher for larger N.

C is probably overfitted. It has very large fluctuating error, and we see that it is be most complex model because it has the highest N for which it still has 0 training error.

4

Flexible is good for its higher chance to find the true model, while less flexible is better for runtime. If you don't know (or have a good guess from the graphing the data or otherwise know something about the data's background), it'll probably a good idea to sacrifice some time to get a better model. On the other hand, if you somehow already know the true model, then you could use a less flexible approach since you wouldn't need to "consider" other models and instead just train a single model.

5

- Use the Hoeffding Inequality. $N = 100$

$$1 - 2e^{-2 \cdot 0.1^2 \cdot 100} \approx 0.729$$

- $N = 200$

$$1 - 2e^{-2 \cdot 0.1^2 \cdot 200} \approx 0.995$$

6

$$X = \begin{bmatrix} 1 & 6.6 & 1 & 4.0 \\ 1 & 6.4 & 2 & 5.0 \\ 1 & 7.2 & 2 & 5.0 \\ 1 & 6.4 & 2 & 5.0 \\ 1 & 7.2 & 2 & 5.0 \end{bmatrix}$$

$$y = \begin{bmatrix} 24.0 \\ 21.6 \\ 34.7 \\ 21.6 \\ 34.7 \end{bmatrix}$$

$$w_{ridge} = \left(X^T X + N\lambda \mathbf{I}'\right)^{-1} X^T y$$

$$= \left(\begin{bmatrix} 1 & 6.6 & 1 & 4.0 \\ 1 & 6.4 & 2 & 5.0 \\ 1 & 7.2 & 2 & 5.0 \\ 1 & 6.4 & 2 & 5.0 \\ 1 & 7.2 & 2 & 5.0 \end{bmatrix}^T \begin{bmatrix} 1 & 6.6 & 1 & 4.0 \\ 1 & 6.4 & 2 & 5.0 \\ 1 & 7.2 & 2 & 5.0 \\ 1 & 6.4 & 2 & 5.0 \\ 1 & 7.2 & 2 & 5.0 \end{bmatrix} + 5 \cdot 0.1 \cdot \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 6.6 & 1 & 4.0 \\ 1 & 6.4 & 2 & 5.0 \\ 1 & 7.2 & 2 & 5.0 \\ 1 & 6.4 & 2 & 5.0 \\ 1 & 7.2 & 2 & 5.0 \end{bmatrix}^T \begin{bmatrix} 24.0 \\ 21.6 \\ 34.7 \\ 21.6 \\ 34.7 \end{bmatrix}$$

$$= \left(\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 6.6 & 6.4 & 7.2 & 6.4 & 7.2 \\ 1 & 2 & 2 & 2 & 2 \\ 4.0 & 5.0 & 5.0 & 5.0 & 5.0 \end{bmatrix} \begin{bmatrix} 1 & 6.6 & 1 & 4.0 \\ 1 & 6.4 & 2 & 5.0 \\ 1 & 7.2 & 2 & 5.0 \\ 1 & 6.4 & 2 & 5.0 \\ 1 & 7.2 & 2 & 5.0 \end{bmatrix} + 5 \cdot 0.1 \cdot \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 6.6 & 6.4 & 7.2 & 6.4 & 7.2 \\ 1 & 2 & 2 & 2 & 2 \\ 4.0 & 5.0 & 5.0 & 5.0 & 5.0 \end{bmatrix} \begin{bmatrix} 24.0 \\ 21.6 \\ 34.7 \\ 21.6 \\ 34.7 \end{bmatrix}$$

7

We need to account for the L_2 term in the gradient. Constants were "absorbed" by the learning rate α.

$$\mathbf{w} = \mathbf{w} - \alpha \left(\frac{1}{N} X^T \left(X\mathbf{w} - \mathbf{y} \right) + \lambda \mathbf{I}' \mathbf{w} \right)$$

Repeat for a certain number of iterations.