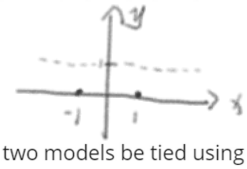
作业四

2021年12月10日 20:44

5. You are given the data points: $(-1, 0), (\rho, 1), (1, 0),\ \rho \geq 0$, and a choice between two models:

- constant $h_0(x) = b_0$ and
- linear $h_1(x) = a_1 x + b_1$.



For which value of ρ would the two models be tied using leave-one-out cross-validation with the squared error measure?

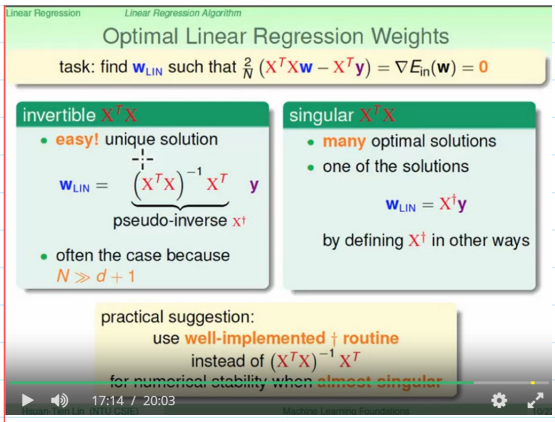
$E_{LOOCV_0} = 0.5^2 + 1^2 + 0.5^2 = 1.5$

11. For Questions 11-12, consider linear regression with virtual examples. That is, we add K virtual examples $(\tilde{\mathbf{x}}_1, \tilde{y}_1), (\tilde{\mathbf{x}}_2, \tilde{y}_2), \ldots, (\tilde{\mathbf{x}}_K, \tilde{y}_K)$ to the training data set, and solve

$$\min_{\mathbf{w}} \frac{1}{N+K} \left(\sum_{n=1}^{N} (y_n - \mathbf{w}^T \mathbf{x}_n)^2 + \sum_{k=1}^{K} (\tilde{y}_k - \mathbf{w}^T \tilde{\mathbf{x}}_k)^2 \right).$$

We will show that using some "special" virtual examples, which were claimed to be a possible way to combat overfitting in Lecture 9, is related to regularization, another possible way to combat overfitting discussed in Lecture 10. Let $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1 \tilde{\mathbf{x}}_2 \ldots \tilde{\mathbf{x}}_K]^T$, and $\tilde{\mathbf{y}} = [\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_K]^T$.

What is the optimal $\mathbf{w}$ to the optimization problem above, assuming that all the inversions exist?



$X = \begin{pmatrix} X_0 \\ \tilde{X} \end{pmatrix} \quad Y = \begin{pmatrix} Y_0 \\ \tilde{Y} \end{pmatrix}$

$w = (X^TX)^{-1}X^TY$

$= \left[\begin{pmatrix} X_0^T & \tilde{X}^T \end{pmatrix} \begin{pmatrix} X_0 \\ \tilde{X} \end{pmatrix} \right]^{-1} \begin{pmatrix} X_0^T & \tilde{X}^T \end{pmatrix} \begin{pmatrix} Y_0 \\ \tilde{Y} \end{pmatrix}$

$= \left[X_0^T X_0 + \tilde{X}^T \tilde{X} \right]^{-1} \left[X_0^T Y_0 + \tilde{X}^T \tilde{Y} \right]$

12. For what $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$ will the solution of the linear regression problem above equal to

$$\mathbf{w}_{\text{reg}} = \text{argmin}_{\mathbf{w}} \frac{\lambda}{N}\|\mathbf{w}\|^2 + \frac{1}{N}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2?$$

$E_{\text{aug}} = \frac{1}{N}\|Xw - y\|^2 + \frac{\lambda}{N}\|w\|^2$

$= E_{\text{sqr}}(w) + \frac{\lambda}{N} w^T w$

weight decay

Regularization

- ○ $\tilde{\mathbf{X}} = \mathbf{I}, \tilde{\mathbf{y}} = \mathbf{0}$
- ✓ $\tilde{\mathbf{X}} = \sqrt{\lambda}\mathbf{I}, \tilde{\mathbf{y}} = \mathbf{0}$
- ○ $\tilde{\mathbf{X}} = \lambda\mathbf{I}, \tilde{\mathbf{y}} = \mathbf{1}$
- ○ $\tilde{\mathbf{X}} = \sqrt{\lambda}\mathbf{X}, \tilde{\mathbf{y}} = \mathbf{y}$
- ○ none of the other choices

Regularization — Weight Decay Regularization

Augmented Error

- if oracle tells you $\lambda > 0$, then

  solving $\nabla E_{\text{in}}(\mathbf{w}_{\text{REG}}) + \frac{2\lambda}{N}\mathbf{w}_{\text{REG}} = \mathbf{0}$

  $\frac{2}{N}\left(\mathbf{Z}^T\mathbf{Z}\mathbf{w}_{\text{REG}} - \mathbf{Z}^T\mathbf{y}\right) + \frac{2\lambda}{N}\mathbf{w}_{\text{REG}} = \mathbf{0}$

- optimal solution:

  $\mathbf{w}_{\text{REG}} \leftarrow (\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{I})^{-1}\mathbf{Z}^T\mathbf{y}$

  —called ridge regression in Statistics

$\tilde{X}^T\tilde{X} = \lambda I \quad \tilde{X}^T Y = 0$

8. For Questions 8-10, please read the following story first. In our credit card example, the bank starts with some vague idea of what constitutes a good credit risk. So, as customers $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ arrive, the bank applies its vague idea to approve credit cards for some of these customers based on a formula $a(\mathbf{x})$. Then, only those who get credit cards are monitored to see if they default or not.

For simplicity, suppose that the first $N = 10000$ customers were given credit cards by the credit approval function $a(\mathbf{x})$. Now that the bank knows the behavior of these customers, it comes to you to improve their algorithm for approving credit. The bank gives you the data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$. Before you look at the data, you do mathematical derivations and come up with a credit approval function. You now test it on the data and, to your delight, obtain perfect prediction.

What is M, the size of your hypothesis set? No data snooping

- ◉ 1
- ○ N
- ○ 2^N
- ○ N^2

10. You assure the bank that you have a got a system g for approving credit cards for new customers, which is nearly error-free. Your confidence is given by your answer to the previous question. The bank is thrilled and uses your g to approve credit for new customers. To their dismay, more than half their credit cards are being defaulted on. Assume that the customers that were sent to the old credit approval function and the customers that were sent to your g are indeed i.i.d. from the same distribution, and the bank is lucky enough (so the "bad luck" that "the true error of g is worse than 1%" does not happen). Which of the following claim is true?

- By applying $a(\mathbf{x})$ NOR $g(\mathbf{x})$ to approve credit for new customers, the performance of the overall credit approval system can be improved with guarantee provided by the previous problem.
- By applying $a(\mathbf{x})$ NAND $g(\mathbf{x})$ to approve credit for new customers, the performance of the overall credit approval system can be improved with guarantee provided by the previous problem.
- By applying $a(\mathbf{x})$ OR $g(\mathbf{x})$ to approve credit for new customers, the performance of the overall credit approval system can be improved with guarantee provided by the previous problem.
- **(selected)** By applying $a(\mathbf{x})$ AND $g(\mathbf{x})$ to approve credit for new customers, the performance of the overall credit approval system can be improved with guarantee provided by the previous problem.
- none of the other choices

Origin Data $\xrightarrow[\text{if } a(x)=1]{}$ $(x_1, y_1), \cdots, (x_n, y_n) \xrightarrow[g(x)]{}$ perfect prediction