

Introduction to Machine Learning

Homework 3: Model Order Selection

Prof. Sundeep Rangan

Fall 2017

1. For each of the following pairs of true functions $f_0(\mathbf{x})$ and model classes $f(\mathbf{x}, \boldsymbol{\beta})$ determine:
(i) if the model class is linear; (ii) if there is no under-modeling; and (iii) if there is no under-modeling, what is the true parameter?

(a) $f_0(x) = 1 + 2x$, $f(x, \boldsymbol{\beta}) = \beta_0 + \beta_1 x + \beta_2 x^2$

(b) $f_0(x) = 1 + 1/(2 + 3x)$, $f(x, a_0, a_1, b_0, b_1) = (a_0 + a_1 x)/(b_0 + b_1 x)$.

(c) $f_0(x) = (x_1 - x_2)^2$ and

$$f(\mathbf{x}, a, b_1, b_2, c_1, c_2) = a + b_1 x_1 + b_2 x_2 + c_1 x_1^2 + c_2 x_2^2.$$

2. In this problem, we will see how to calculate the bias when there is undermodeling. Suppose that training data (x_i, y_i) , $i = 1, \dots, n$ is fit using a simple linear model of the form,

$$\hat{y} = f(x, \boldsymbol{\beta}) = \beta_0 + \beta_1 x.$$

However, the true relation between x and y is given

$$y = f_0(x), \quad f_0(x) = \beta_{00} + \beta_{01}x + \beta_{02}x^2,$$

where the “true” function $f_0(x)$ is quadratic and $\boldsymbol{\beta}_0 = (\beta_{00}, \beta_{01}, \beta_{02})$ is the vector of the true parameters. There is no noise.

- (a) Write an expression for the least-squares estimate $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$ in terms of the training data (x_i, y_i) , $i = 1, \dots, n$. These expressions will involve multiple steps. You do not need to simplify the equations. Just make sure you state clearly how one would compute $\hat{\boldsymbol{\beta}}$ from the training values.
- (b) Using the fact that $y_i = f_0(x_i)$ in the training data, write the expression for $\boldsymbol{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ in terms of the values x_i and the true parameter values $\boldsymbol{\beta}_0$. Again, you do not need to simplify the equations. Just make sure you state clearly how one would compute $\hat{\boldsymbol{\beta}}$ from the true parameter vector $\boldsymbol{\beta}_0$ and \mathbf{x} .
- (c) Suppose that the true parameters are $\boldsymbol{\beta}_0 = (1, 2, -1)$ and the model is trained using 10 values x_i uniformly spaced in $[0, 1]$. Write a short python program to compute the estimate parameters $\hat{\boldsymbol{\beta}}$. Plot the estimated function $f(x, \hat{\boldsymbol{\beta}})$ and true function $f_0(x)$ for $x \in [0, 3]$.

- (d) For what value x in this range $x \in [0, 3]$ is the bias $\text{Bias}^2(x) = (f(x, \hat{\beta}) - f_0(x))^2$ largest?
3. A medical researcher wishes to evaluate a new diagnostic test for cancer. A clinical trial is conducted where the diagnostic measurement y of each patient is recorded along with attributes of a sample of cancerous tissue from the patient. Three possible models are considered for the diagnostic measurement:
- Model 1: The diagnostic measurement y depends linearly only on the cancer volume.
 - Model 2: The diagnostic measurement y depends linearly on the cancer volume and the patient's age.
 - Model 3: The diagnostic measurement y depends linearly on the cancer volume and the patient's age, but the dependence (slope) on the cancer volume is different for two types of cancer – Type I and II.
- (a) Define variables for the cancer volume, age and cancer type and write a linear model for the predicted value \hat{y} in terms of these variables for each of the three models above. For Model 3, you will want to use one-hot coding.
- (b) What are the numbers of parameters in each model? Which model is the most complex?
- (c) Since the models in part (a) are linear, given training data, we should have $\hat{\mathbf{y}} = \mathbf{A}\boldsymbol{\beta}$ where $\hat{\mathbf{y}}$ is the vector of predicted values on the training data, \mathbf{A} is a feature matrix and $\boldsymbol{\beta}$ is the vector of parameters. To test the different models, data is collected from 100 patients. The records of the first three patients are shown below:

Patient ID	Measurement y	Cancer type	Cancer volume	Patient age
12	5	I	0.7	55
34	10	II	1.3	65
23	15	II	1.6	70
\vdots	\vdots	\vdots	\vdots	\vdots

Based on this data, what would be the values of first three rows of the three \mathbf{A} matrices be for the three models in part (a)?

- (d) To evaluate the models, 10-fold cross validation is used with the following results.

Model	Mean training RSS	Mean test RSS	Test RSS std deviation
1	2.0	2.01	0.03
2	0.7	0.72	0.04
3	0.65	0.70	0.05

All RSS values are per sample, and the last column is the (biased) standard deviation – not the standard error. Which model should be selected based on the “one standard error rule”?