

P9120- Homework # 2

Assigned: October 3, 2019

Due: October 17, 2019 (in class)

Maximum points that you can score in this Homework is 20.

Please include all R or matlab code you used to complete this homework.

1. Ex. 4.5 of [ESL]
2. Ex. 5.1 of [ESL]
3. (A simulation study).
 - (a) Generate a vector \mathbf{x} consisting of 50 points drawn at random from Uniform $[0, 1]$.
 - (b) Generate 100 training sets. Each training set consists of 50 pairs of (X, Y) , with $(X_1, \dots, X_{50}) = \mathbf{x}$ and $Y_i = \sin^3(2\pi X_i^3) + \epsilon_i$ for $i = 1, \dots, 50$, where ϵ_i is drawn from the standard normal distribution. For each training set, do following.
 - i. Fit the data with methods/models listed below.
 - OLS with linear model: $\beta_0 + \beta_1 X$.
 - OLS with cubic polynomial model: $\beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$.
 - Cubic spline (or B-spline) with 2 knots at 0.33 and 0.66.
 - Natural cubic spline with 5 knots at 0.1, 0.3, 0.5, 0.7, 0.9.
 - Smoothing spline with tuning parameter chosen by GCV.
 - ii. Compute the vector of fitted value $\hat{\mathbf{y}}$ obtained from each method/model.
 - (c) Now for each method/model, you obtain a matrix of fitted values, with the i -th row and j -th column value \hat{y}_{ij} representing the fitted value at $X = x_i$ from the j -th training set.
 - (d) For each method/model, compute the pointwise variance of fitted values across the 100 training sets. This gives you a vector of pointwise variance.

Plot the pointwise variance curves (against \mathbf{x}) for each method/model. (Note: Your plot would be similar to Figure 5.3 of [ESL].)

4. The South African heart disease data is described on page 122 of the textbook. This data set can be found on the text book web site:

<https://web.stanford.edu/~hastie/ElemStatLearn/>.

Divide the dataset into a training set consisting of the first 300 observations, and a test set consisting of the remaining observations. Apply logistic regression, LDA and QDA on the training set. For each method, report the test error and its standard error over the test set. Briefly discuss your results.