An Introduction to Statistical Learning: Chapter 7

Due on January 15,2016 at 3:10pm

G. James et.al.

Tony Jiang

1.It was mentioned in the chapter that a cubic regression spline with one knot at can be obtained using a basis of the form $x, x^2, x^3, (x-\xi)^3_+$, where $(x-\xi)^3_+ = (x-\xi)^3$ if $x > \xi$ and equals 0 otherwise.
We will now show that a function of the form

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-\xi)^3_+$$

is indeed a cubic regression spline, regardless of the values of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$.
a) Find a cubic polynomial

$$f_1(x) = a_1 + b_1 x + c_1 x^2 + d_1 x^3$$

such that $f(x) = f_1(x)$ for all $x \leq \xi$. Express a_1, b_1, c_1, d_1 in terms of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$
(b) Find a cubic polynomial

$$f_2(x) = a_2 + b_2 x + c_2 x^2 + d_2 x^3$$

such that $f(x) = f_2(x)$ for all $x > \xi$. Express a_2, b_2, c_2, d_2 in terms of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$. We have now established that $f(x)$ is a piece wise polynomial. (c) Show that $f_1(\xi) = f_2(\xi)$. That is, $f(x)$ is continuous at ξ.
(d) Show that $f_1'(\xi) = f_2'(\xi)$. That is, $f'(x)$ is continuous at ξ. (e) Show that $f''(\xi) = f_2''(\xi)$. That is, $f''(x)$ is continuous at ξ. Therefore, $f(x)$ is indeed a cubic spline.
Hint: Parts (d) and (e) of this problem require knowledge of single variable calculus. As a reminder, given a cubic polynomial

$$f_1(x) = a_1 + b_1 x + c_1 x_2 + d_1 x_3,$$

the first derivative takes the form

$$f_1(x) = b_1 + 2c_1 x + 3d_1 x^2$$

and the second derivative takes the form

$$f_1''(x) = 2c_1 + 6d_1 x.$$

Answers:

(a). when $x \leq \xi$, the last term in $f(x)$ is eliminated. so $f_1(x) = f(x)$. i.e.

$$\begin{aligned} a_1 &= \beta_0 \\ b_1 &= \beta_1 \\ c_1 &= \beta_2 \\ d_1 &= \beta_3 \end{aligned}$$

(b).

$$
\begin{aligned}
&\text{for all x} \leq \xi \\
f(x) &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-\xi)_+^3 \\
&= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-\xi)^3 \\
&= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 \big(x^3 - 3\xi x^2 + 3\xi^2 x - \xi^3\big) \\
&= (\beta_0 - \beta_4\xi^3) + (\beta_1 + 3\beta_4\xi^2)x + (\beta_2 - 3\beta_4\xi)x^2 + (\beta_3 + \beta_4)x^3
\end{aligned}
$$

now we have an expression for$f_2(x) =$$a_2 + b_2 x + c_2 x^2 + d_2 x^3$in which

$$
\begin{aligned}
a_2 &= \beta_0 - \beta_4\xi^3 \\
b_2 &= \beta_1 + 3\beta_4\xi^2 \\
c_2 &= \beta_2 - 3\beta_4\xi \\
d_2 &= \beta_3 + \beta_4
\end{aligned}
$$

(c).

$$
\begin{aligned}
f_1(x=\xi) &= \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \\
&= \beta_0 + \beta_1 \xi + \beta_2 \xi^2 + \beta_3 \xi^3 \\
f_2(x=\xi) &= (\beta_0 - \beta_4\xi^3) + (\beta_1 + 3\beta_4\xi^2)x + (\beta_2 - 3\beta_4\xi)x^2 + (\beta_3 + \beta_4)x^3 \\
&= \beta_0 + \beta_1 \xi + \beta_2 \xi^2 + \beta_3 \xi^3 \\
f_1(x=\xi) &= f_2(x=\xi)
\end{aligned}
$$

(d).

$$
\begin{aligned}
f_1'(x=\xi) &= \beta_1 + 2\beta_2 x + 3\beta_3 x^2 \\
&= \beta_1 + 2\beta_2 \xi + 3\beta_3 \xi^2 \\
f_2'(x=\xi) &= \beta_1 + 3\beta_4\xi^2 + 2(\beta_2 - 3\beta 4\xi)x + 3(\beta_3 + \beta_4)x^2 \\
&= \beta_1 + 2\beta_2 \xi + 3\beta_3 \xi^2 \\
f_1'(x=\xi) &= f_2'(x=\xi)
\end{aligned}
$$

(e)

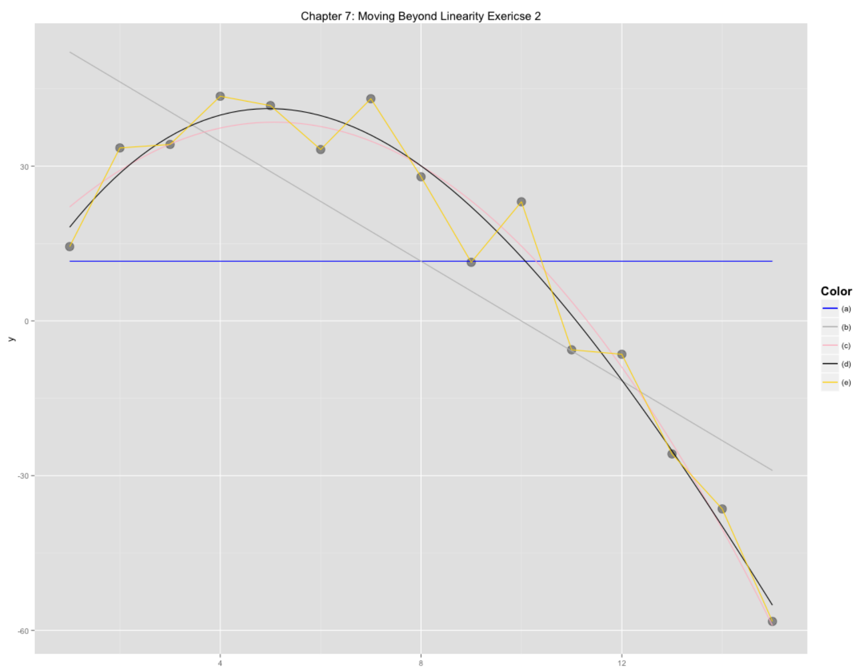
$$
\begin{aligned}
f_1''(x=\xi) &= 2\beta_2 + 6\beta_3 x \\
&= 2\beta_2\xi + 6\beta_3\xi \\
f_2''(x=\xi) &= 2(\beta_2 - 3\beta 4\xi) + 6(\beta_3 + \beta_4)x \\
&= 2\beta_2\xi + 6\beta_3\xi \\
f_2''(x=\xi) &= f_2''(x=\xi)
\end{aligned}
$$

2. Suppose that a curve g is computed to smoothly fit a set of n points using the following formula:

$$\hat{g} = arg\min_{g} \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int \left[g^{(m)}(x) \right]^2 dx \right)$$

where $g^{(m)}$ represents the mth derivative of g (and $g^{0)} = g$). Provide example sketches of $\hat{g}$ in each of the following scenarios.
(a) $\lambda = \infty, m = 0$.
(b) $\lambda = \infty, m = 1$.
(c) $\lambda = \infty, m = 2$.
(d) $\lambda = \infty, m = 3$.
(e) $\lambda = 0, m = 3$.

Answers:



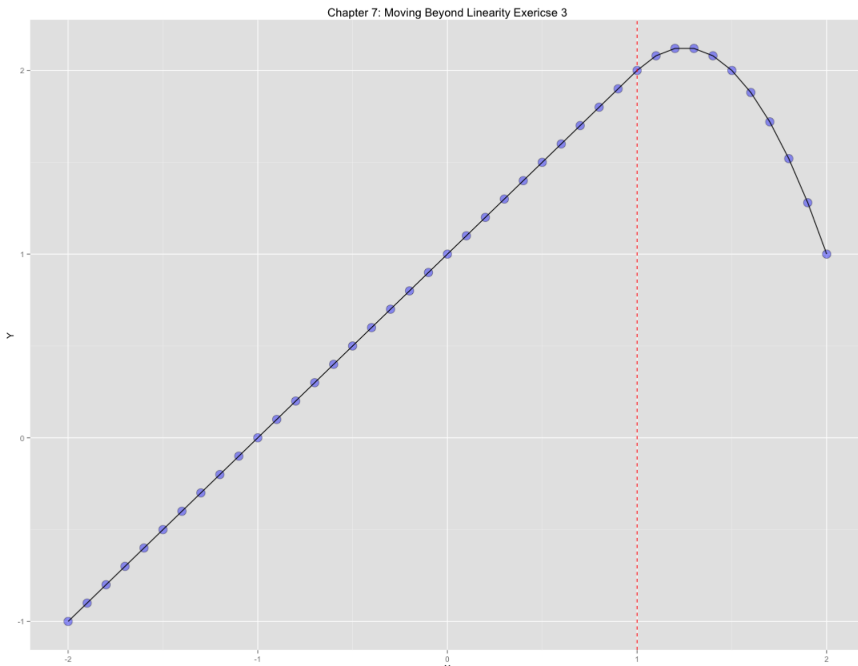
(a). the RSS term has no effect. $g(x) = k$ is the line parallel to x-axis passing through the mean of the data (order=0).
(b). the RSS term has no effect. $g'(x) = 0$ is the linear least square line fit to the data (order=1).
(c). the RSS term has no effect. $g''(x) = 0$ is the least square quadratic fit to the data(order=2).
(d). the RSS term has no effect. $g'''(x) = 0$ is the least square cubic fit to the data (order=3).
(e). the penalty term has no effect. g(x) will pass through each data point and can take any form.

3. Suppose we fit a curve with basis functions $b_1(X) = X, b_2(X) = (X-1)^2 I(X \geq 1)$. (Note that $I(X \geq 1)$ equals 1 for $X \leq 1$ and 0 otherwise.) We fit the linear regression model

$$Y = \beta_0 + \beta_1 b_1(X) + \beta_2 b_2(X) + \epsilon,$$

and obtain coefficient estimates $\hat{\beta}_0 = 1$, $\hat{\beta}_1 = 1$, $\hat{\beta}_2 = -2$, Sketch the estimated curve between $X = -2$ and $X = 2$. Note the intercepts, slopes, and other relevant information.

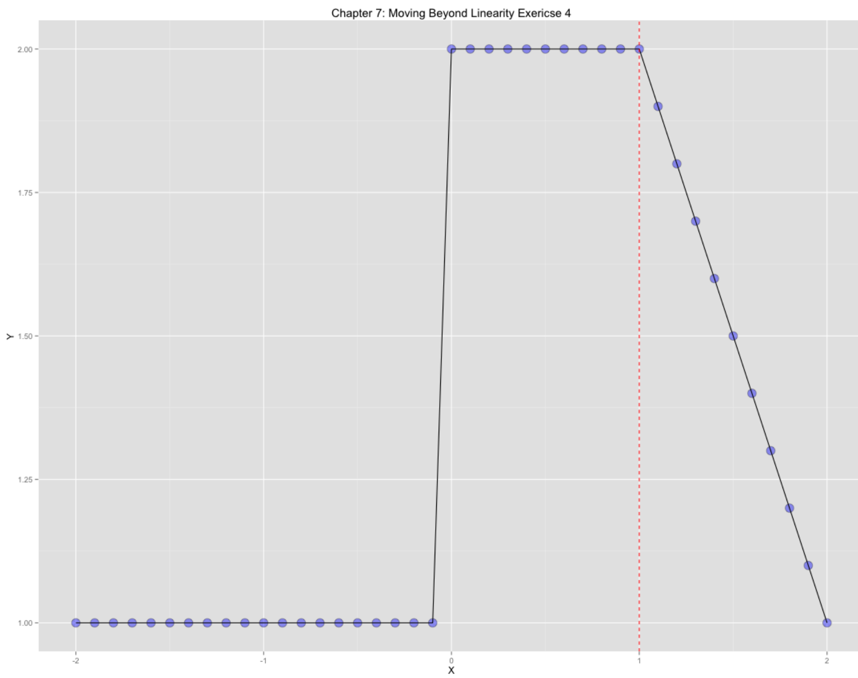
Answers:



4. Suppose we fit a curve with basis functions $b_1(X) = I(0 \leq X \leq 2) - (X-1)I(1 \leq X \leq 2), b_2(X) = (X-3)I(3 \leq X \leq 4) + I(4 \leq X \leq 5)$ We fit the linear regression model

$$Y = \beta_0 + \beta_1 b_1(X) + \beta_2 b_2(X) + \epsilon,$$

and obtain coefficient estimates $\hat{\beta}_0 = 1$, $\hat{\beta}_1 = 1$, $\hat{\beta}_2 = 3$. Sketch the estimated curve between X = 2 and X = 2. Note the intercepts, slopes, and other relevant information.
textbfAnswers:



5. Consider two curves, $\hat{g}_1$ and $\hat{g}_2$, defined by

$$\hat{g}_1 = arg\min_g \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int \left[g^{(3)}(x) \right]^2 dx \right)$$

$$\hat{g}_2 = arg\min_g \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int \left[g^{(4)}(x) \right]^2 dx \right)$$

where $g^{(m)}$ represents the mth derivative of g.
(a) As $\lambda \to \infty$, will $\hat{g}_1$ or $\hat{g}_2$ have the smaller training RSS?
(b) As $\lambda \to \infty$, will $\hat{g}_1$ or $\hat{g}_2$ have the smaller test RSS?
(c) For $\lambda = 0$, will $\hat{g}_1$ or $\hat{g}_2$ have the smaller training and test RSS?

Answers:
(a). $\hat{g}_2$ has more flexibility thus smaller training RSS.
(b). $\hat{g}_1$ has more flexibility thus smaller test RSS in most cases since $\hat{g}_2$ tends to overfit the data. However ,if the data really has a quartic item, then $\hat{g}_2$ is expected to have smaller test RSS as it is closer to the truth.
(c). when $\lambda = 0$, two curves will look exactly the same.