BIOSTAT 274 Spring 2021 Homework 2

Due by 11:59 PM, 05/14/2021

Zian ZHUANG

Theoretical Part

Q1.

(a)

Answer: For the training set, when the Bayes decision boundary is linear, QDA will perform better since it has higher flexibility and may fit the data better. As for the test set, LDA is likely to perform better because QDA may overfit the linearity of the Bayes decision boundary in training set.

(b)

Answer: When the Bayes decision boundary is non-linear, QDA is likely to perform better on both the training set and the test set because of its higher flexibility.

(c)

Answer: Generally speaking, QDA tends to improve prediction accuracy relative to LDA when the sample size n increases, since it has higher flexibility and may fit the data better. Increased variance in test set of QDA would be offset by the larger sample sizes

(d)

Answer: False : LDA is likely to fit a linear decision boundary better than QDA, so that it can provide a better test error rate. QDA may provide an better error rate on training set (over-fitting) but worse on the test set(due to higher variance), especially when sample size n is small.

Q2.

(a)

Answer:

According to the definition, we know that a cubic spline with the two knots at ξ_1 and ξ_2 follows,

$$f(x) = \begin{cases} a_1X^3 + b_1X^2 + c_1X + d_1, & X \leqslant \xi_1 \\ a_2X^3 + b_2X^2 + c_2X + d_2, & \xi_1 \leqslant X \leqslant \xi_2 \\ a_3X^3 + b_3X^2 + c_3X + d_3, & X \geqslant \xi_2 \end{cases}$$

Since function is continuous at the boundary, we also have

$$\begin{cases} a_1\xi_1^3 + b_1\xi_1^2 + c_1\xi_1 + d_1 - (a_2\xi_1^3 + b_2\xi_1^2 + c_2\xi_1 + d_2) = 0 \\ a_2\xi_2^3 + b_2\xi_2^2 + c_2\xi_2 + d_2 - (a_3\xi_2^3 + b_3\xi_2^2 + c_3\xi_2 + d_3) = 0 \end{cases}$$

Assume that these truncated power basis functions are a basis for the cubic spline, then we know that $f(x)$ can be formed by basis functions. It is easy to see that,

$$f(x) = \begin{cases} a_1h_4(X) + b_1h_3(X) + c_1h_2(X) + d_1h_1(X), & X \leqslant \xi_1 \\ a_2h_4(X) + b_2h_3(X) + c_2h_2(X) + d_2h_1(X), & \xi_1 \leqslant X \leqslant \xi_2 \\ a_3h_4(X) + b_3h_3(X) + c_3h_2(X) + d_3h_1(X), & X \geqslant \xi_2 \end{cases}$$

As for the region $\xi_1 \leqslant X \leqslant \xi_2$, define a coefficient θ for $h_5(X)$,
we can also represent it as,

$$\begin{aligned} & f(X) = a_1X^3 + b_1X^2 + c_1X + d_1 + \theta h_5(X), \quad \xi_1 \leqslant X \leqslant \xi_2 \\ \Longrightarrow & (a_1 + \theta)X^3 + (b_1 - 3\xi_1\theta)X^2 + (c_1 + 3\xi_1\theta)X + d_1 - \xi_1^3 \end{aligned}$$

Then we know that,

$$\begin{aligned} & a_1 + \theta = a_2 \Longrightarrow a_1 - a_2 = -\theta \\ & b_1 - 3\xi_1\theta = b_2 \Longrightarrow b_1 - b_2 = 3\xi_1\theta \\ & c_1 + 3\xi_1^2\theta = c_2 \Longrightarrow c_1 - c_2 = -3\xi_1^2\theta \\ & d_1 - \xi_1^3\theta = d_2 \Longrightarrow d_1 - d_2 = \xi_1^3 \end{aligned}$$

Then we consider equation at ξ_1, plug the results,

$$\begin{aligned} & -\theta\xi_1^3 + 3\xi_1\theta\xi_1^2 - 3\xi_1^2\theta\xi_1 + \xi_1^3\theta \\ = & -\theta\xi_1^3 + 3\theta\xi_1^3 - 3\theta\xi_1^3 + \theta\xi_1^3 \\ = & \xi_1^3(-\theta + 3\theta - 3\theta + \theta) \\ = & \xi_1^3 * 0 \\ = & 0 \end{aligned}$$

Then we can see that it meets boundary constraint of the cubic spline. We can further know that function $(a_1 + \theta)X^3 + (b_1 - 3\xi_1\theta)X^2 + (c_1 + 3\xi_1\theta)X + d_1 - \xi_1^3$ match the $f(x)$ when $\xi_1 \leqslant X \leqslant \xi_2$.

Similarly, considering the region $X \geqslant \xi_2$, we can define ϕ as a coefficient for $h_6(X)$ and get

$$\begin{aligned} & f(X) = a_2X^3 + b_2X^2 + c_2X + d_2 + \phi h_5(X), \quad X \geqslant \xi_2 \\ \Longrightarrow & (a_2 + \phi)X^3 + (b_2 - 3\xi_2\phi)X^2 + (c_2 + 3\xi_2\phi)X + d_2 - \xi_2^3 \end{aligned}$$

and then for the constraint equation at ξ_2,we have

$$\begin{aligned} & -\phi\xi_2^3 + 3\xi_2\phi\xi_2^2 - 3\xi_2^2\phi\xi_2 + \xi_2^3\phi \\ = & -\phi\xi_2^3 + 3\phi\xi_2^3 - 3\phi\xi_2^3 + \phi\xi_2^3 \\ = & \xi_2^3(-\phi + 3\phi - 3\phi + \phi) \\ = & \xi_2^3 * 0 \\ = & 0 \end{aligned}$$

Then we can see that it meets boundary constraint of the cubic spline. We can further know that function $(a_2 + \phi)X^3 + (b_2 - 3\xi_2\phi)X^2 + (c_2 + 3\xi_2\phi)X + d_2 - \xi_2^3$ match the $f(x)$ when $X \geqslant \xi_2$.

Since we know that we can form a weighted sum of the six basis functions to represent a cubic spline with two knots at ξ_1 and ξ_2, we can conclude that these truncated power basis functions are a basis for the cubic spline.

(b)

Answer:

Generalize (1) to K interior knots and use it represent a cubic splines fit, we have,

$$f(X) = \sum_{j=0}^{3} B_j X^j + \sum_{k=1}^{K} \theta_k (X - \xi_k)_+^3$$

Since Natural cubic splines (NCS) forces the second and third derivatives to be zero at the boundaries, as for the basis functions for cubic splines when $X \leqslant \xi_1$ (i.e. only contains $\sum_{j=0}^{3} B_j X^j$ part), we know that,

$$\frac{\partial}{\partial^2 X} \sum_{j=0}^{3} B_j X^j = 0 + 0 + 2B_2 + 6B_3 X$$
$$\frac{\partial}{\partial^3 X} \sum_{j=0}^{3} B_j X^j = 0 + 0 + 0 + 6B_3 X$$

It is easy to see that $B_2 = B_3 = 0$ since all second and third derivatives should be zero when $X \leqslant \xi_1$.

As for $X \geqslant \xi_k$, plug in $B_2 = B_3 = 0$ we have,

$$\begin{aligned} f(X) &= B_1 + B_2 X + \sum_{k=1}^{K} \theta_k (X - \xi_k)^3 \\ &= B_1 + B_2 X + \sum_{k=1}^{K} \theta_k (X^3 - 3X^2\xi_k + 3X\xi_k^2 - \xi_k^3) \end{aligned}$$

Since the function $f(X)$ should be linear when $X \geqslant \xi_k$, we know that both X^2 and X^3 terms should be zero, which means that,

$$\sum_{k=1}^{K} \theta_k X^3 = 0 \Longrightarrow X^3 \sum_{k=1}^{K} \theta_k = 0 \Longrightarrow \sum_{k=1}^{K} \theta_k = 0$$
$$\sum_{k=1}^{K} \theta_k 3X^2 \xi_k = 0 \Longrightarrow 3X^2 \sum_{k=1}^{K} \theta_k \xi_k = 0 \Longrightarrow \sum_{k=1}^{K} \theta_k \xi_k = 0$$

Thus we can conclude that, the natural boundary conditions for natural cubic splines imply the following linear constraints on the coefficients:

$$\begin{cases} B_2 = 0 \\ B_3 = 0 \\ \sum_{k=1}^{K} \theta_k = 0 \\ \sum_{k=1}^{K} \theta_k \xi_k = 0 \end{cases}$$