P9120 HW2 answer

Guojing Wu / UNI: gw2383

10/14/2019

1. Consider a two-class logistic regression problem with $x \in \mathbb{R}$. Characterize the maximum-likelihood estimates of the slope and intercept parameter if the sample x_i for the two classes are separated by a point $x_0 \in \mathbb{R}$. Generalize this result to (a) $x \in \mathbb{R}^p$ (see Figure 4.16), and (b) more than two classes.

For logistic regression, the log-likelihood looks like: $l(\beta) = \sum_{i=1}^{N}[y_i(\beta_0 + \beta_1 x_i) - log(1 + exp(\beta_0 + \beta_1 x_i))]$. And since the data are separated by a point $x_0 \in \mathbb{R}$, we define:

$$y_i = \begin{cases} 0, & x_i \le x_0 \\ 1, & x_i > x_0 \end{cases}$$

Then we can rewrite the log-likelihood as:

$$\begin{aligned} l(\beta) = & \sum_{i=1, x_i \le x_0}^{N} -log(1 + exp(\beta_0 + \beta_1 x_i)) \\ & + \sum_{i=1, x_i > x_0}^{N} [\beta_0 + \beta_1 x_i - log(1 + exp(\beta_0 + \beta_1 x_i))] \\ = & \sum_{i=1, x_i \le x_0}^{N} -log(1 + exp(\beta_0 + \beta_1 x_0 + \beta_1 (x_i - x_0)) \\ & + \sum_{i=1, x_i > x_0}^{N} [\beta_0 + \beta_1 x_0 + \beta_1 (x_i - x_0) - log(1 + exp(\beta_0 + \beta_1 x_0 + \beta_1 (x_i - x_0)))] \\ = & \sum_{i=1, x_i \le x_0}^{N} -log(1 + exp(\beta_1 (x_i - x_0)) + \sum_{i=1, x_i > x_0}^{N} [\beta_1 (x_i - x_0) - log(1 + exp(\beta_1 (x_i - x_0)))] \end{aligned}$$

Above we let $\beta_0 + \beta_1 x_0 = 0$ to simplify the equation.

For the first part of the equation $\sum_{i=1, x_i \le x_0}^{N} -log(1 + exp(\beta_1 (x_i - x_0)) \le 0$, when $\beta_1 \uparrow$, $exp(\beta_1 (x_i - x_0)) \downarrow$, so $-log(1 + exp(\beta_1 (x_i - x_0)) \uparrow$, this part is monotone increasing along with β_1. For the second part, we let $f(x) = x - log(1 + e^x)$, $f'(x) = 1 - \frac{e^x}{1+e^x} = \frac{1}{1+e^x} > 0$, so the second part is also monotone increasing along with β_1. Hence, the solution for one dimension separable logistic regression can be characterzied by $\beta_1 \to +\infty$ and $\beta_0 = -\beta_1 x_0 \to -sign(x_0)\infty$.

(a)

When generalize this result to $x \in \mathbb{R}^p$, all data are separated by a hyperplane $H_{\omega,b}$, the solution for p dimension separable logistic regression can be characterzied by $||\omega||_2^2 \to +\infty$.

(b)

When generalize this result to more than two classes, e.g., K classes, we then have K-1 hyperplanes H_{ω_i,b_i} that can separate all the data, the solution for K classes separable logistic regression can be characterzied by $||\omega_i||_2^2 \to +\infty$, for $i \in \{1, 2, ..., K-1\}$

2. Show that the truncated power basis functions in (5.3) represent a basis for a cubic spline with the two knots as indicated.

Based on all $h_i(X)$, we write $f(x) = \sum_{m=1}^{6} \beta_m h_m(x)$. Then we need to show the continuity of $f(x)$, $f'(x)$ and $f''(x)$ at knots ξ_1 and ξ_2.

(a) continuity of $f(x)$

left limit of $f(x)$ at ξ_1

$$\begin{aligned} &\forall h > 0 \\ f(\xi_1 - h) &= \beta_1 + \beta_2(\xi_1 - h) + \beta_3(\xi_1 - h)^2 + \beta_4(\xi_1 - h)^3 \\ &\quad + \beta_5(\xi_1 - h - \xi_1)_+^3 + \beta_6(\xi_1 - h - \xi_2)_+^3 \\ &= \beta_1 + \beta_2(\xi_1 - h) + \beta_3(\xi_1 - h)^2 + \beta_4(\xi_1 - h)^3 \\ lim_{h\to 0} f(\xi_1 - h) &= \beta_1 + \beta_2\xi_1 + \beta_3\xi_1^2 + \beta_4\xi_1^3 \end{aligned}$$

right limit of $f(x)$ at ξ_1

$$\begin{aligned} &\forall h > 0 \\ f(\xi_1 + h) &= \beta_1 + \beta_2(\xi_1 + h) + \beta_3(\xi_1 + h)^2 + \beta_4(\xi_1 + h)^3 \\ &\quad + \beta_5(\xi_1 + h - \xi_1)_+^3 + \beta_6(\xi_1 + h - \xi_2)_+^3 \\ &= \beta_1 + \beta_2(\xi_1 + h) + \beta_3(\xi_1 + h)^2 + \beta_4(\xi_1 + h)^3 + \beta_5 h^3 + \beta_6(\xi_1 + h - \xi_2)_+^3 \\ lim_{h\to 0} f(\xi_1 + h) &= \beta_1 + \beta_2\xi_1 + \beta_3\xi_1^2 + \beta_4\xi_1^3 \end{aligned}$$

Left limit = right limit, so $f(x)$ is continuous at ξ_1. In a similar way we can prove that $f(x)$ is also continuous at ξ_2

(b) continuity of $f'(x)$

left limit of $f'(x)$ at ξ_1

$$\begin{aligned} f'_-(\xi_1) &= lim_{h\to 0} \frac{f(\xi_1) - f(\xi_1 - h)}{h} \\ &= lim_{h\to 0} \frac{\beta_1 + \beta_2\xi_1 + \beta_3\xi_1^2 + \beta_4\xi_1^3 - \beta_1 - \beta_2(\xi_1 - h) - \beta_3(\xi_1 - h)^2 - \beta_4(\xi_1 - h)^3}{h} \\ &= lim_{h\to 0} \frac{\beta_2 h + 2\beta_3\xi_1 h + 3\beta_4\xi_1^2 h + O(h^2) + O(h^3)}{h} \\ &= \beta_2 + 2\beta_3\xi_1 + 3\beta_4\xi_1^2 \end{aligned}$$

right limit of $f'(x)$ at ξ_1

$$
\begin{aligned}
f'_+(\xi_1) &= lim_{h\to 0}\frac{f(\xi_1+h)-f(\xi_1)}{h} \\
&= lim_{h\to 0}\frac{\beta_1+\beta_2(\xi_1+h)+\beta_3(\xi_1+h)^2+\beta_4(\xi_1+h)^3+\beta_5 h^3+\beta_6(\xi_1+h-\xi_2)^3_+-\beta_1-\beta_2\xi_1-\beta_3\xi_1^2-\beta_4\xi_1^3}{h} \\
&= lim_{h\to 0}\frac{\beta_2 h+2\beta_3\xi_1 h+3\beta_4\xi_1^2 h+\beta_6(\xi_1+h-\xi_2)^3_+ +O(h^2)+O(h^3)}{h} \\
&= \beta_2+2\beta_3\xi_1+3\beta_4\xi_1^2
\end{aligned}
$$

Left limit = right limit, so $f'(x)$ is continuous at ξ_1. In a similar way we can prove that $f'(x)$ is also continuous at ξ_2

(c) continuity of $f''(x)$

In a similar way we can prove that $f''(x)$ is continuous at ξ_1 and ξ_2, $f''_-(\xi_1) = f''_+(\xi_1) = 6\beta_4\xi_1^2$.

Hence we proved that power basis functions in (5.3) represent a basis for a cubic spline with the two knots.

3. A simulation study

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 website. 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.

Appendix

```
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE, comment = "")
library(tidyverse)
options(knitr.table.format = "latex")
theme_set(theme_bw())
```