Eric Perkerson Date: March 23, 2020

Monotone Splines for Regression

1 The Regression Problem

Given some statistical data $\{(x_i, y_i)\}_{i=1}^n$ where $x_i,\ y_i \in \mathbb{R}$, we the regression problem is the problem of estimating the function $f(x) = E(y|x)$ where $f\colon \mathbb{R} \to \mathbb{R}$. The primary difficulty here is to avoid overfitting, that is, to construct an estimate $\widehat{f}$ for the unkown function f that is close in some sense, not merely one that is close for the observed data points. This is the generalization problem. A good estimate $\widehat{f}$ is close to f even for non-observed values of x. To find such an estimate, we have make some prior assumptions about $\hat{f}$. These are typically enforced by either restricting the function class V from which we choose $\widehat{f}$ or by penalizing certain functions. Linear regression is an example of an extremely strong limitation on the class of functions used, viz. $V = \{f\colon f(x) = ax + b\}$. Here we use monotone splines (M-splines), integrated splines (I-splines), and convex splines (C-splines) as ways of restricting the class of functions to avoid overfitting.

2 Monotone Splines

We first a degree k for the spline function $\widehat{f}$ and a closed interval $[a, b]$ for the domain of $\widehat{f}$. This can be done, for instance, by using the minimum and maximum value of the set $\{x_i\}$. We then choose m interior break points $t_{k+1}, \dots, t_{k+m} \in (a, b)$. Lastly, define k additional break points at each end point $t_1 = \cdots = t_k = a$ and $t_{m+k+1} = \cdots = t_{m+2k} = b$ for notational simplicity.

The monotone splines or M-splines of order k are given by the recursive formula

$$
\begin{aligned}
M_i^{(1)}(x) &= \frac{\mathbb{1}_{[t_i, t_{i+1}]}}{t_{i+1} - t_i} \text{ if } t_{i+1} \neq t_i \\
M_i^{(1)}(x) &= 0 \text{ if } t_{i+1} = t_i \\
M_i^{(k)}(x) &= \frac{k[(x - t_i)M_i^{(k-1)}(x) + (t_{i+k} - x)M_{i+1}^{(k-1)}(x)]}{(k-1)}
\end{aligned}
$$

and then integrated and convex splines are defined by

$$
\begin{aligned}
I_i^{(k)}(x) &= \int_a^x M_i^{(k)}(t)dt \\
C_i^{(k)}(x) &= \int_a^x I_i^{(k)}(t)dt
\end{aligned}
$$

3 Degree $k = 1$

t

4 Degree $k = 2$

For $i = 1$:

$$M_1^{(2)}(x) = \begin{cases} 0, & x \notin [t_i, t_{i+2}] \\ \frac{2(x-t_i)}{(t_{i+2}-t_i)(t_{i+1}-t_i)}, & x \in [t_i, t_{i+1}] \\ \frac{-2(x-t_{i+2})}{(t_{i+2}-t_i)(t_{i+2}-t_{i+1})}, & x \in [t_{i+1}, t_{i+2}] \end{cases}$$

For $i = 2, \ldots, k + m - 1$:

$$M_i^{(2)}(x) = \begin{cases} 0, & x \notin [t_i, t_{i+2}] \\ \frac{2(x-t_i)}{(t_{i+2}-t_i)(t_{i+1}-t_i)}, & x \in [t_i, t_{i+1}] \\ \frac{-2(x-t_{i+2})}{(t_{i+2}-t_i)(t_{i+2}-t_{i+1})}, & x \in [t_{i+1}, t_{i+2}] \end{cases}$$

5 Degree $k = 3$