3
Suavidade em Árvores de Regressão

Nesse capítulo será apresentado o conceito de suavidade e sua aplicação em árvores de regressão. Primeiro serão formuladas as definições relacionadas às funções de transição. Boa parte dessa teoria é adaptação do trabalho de Li, Griffiths e Ward (33), que utilizou a técnica de suavidade para ajustar curvas e superfícies à pontos dispersos. Em seguida, na seção 3.2, as árvores de regressão serão relacionadas às funções de transição e, por fim, na seção 3.3, a regressão no método RCRI é formalizada.

PUC-Rio - Certificação Digital Nº 0521445/CA

3.1
Funções de Transição

Definição 3.1.1 *Seja D um subconjunto de $\mathbb{R}^d$ e $\mathcal{N}_D$ uma vizinhança do bordo de D. Uma função $\mathbf{g}_D : \mathbb{R}^d \rightarrow [0,1]$ é chamada de função transição com relação à região D se $\mathbf{g}_D(\mathbf{x}) = 0 \ \forall \ \mathbf{x} \notin D \cup \mathcal{N}_D$, $0 < \mathbf{g}_D(\mathbf{x}) < 1 \ \forall \ \mathbf{x} \in \mathcal{N}_D$ e $\mathbf{g}_D(\mathbf{x}) = 1 \ \forall \ \mathbf{x} \in D - \mathcal{N}_D$.*

A figura 3.1 a seguir ilustra a definição acima, mostrando como serão os valores de $\mathbf{g}_D(\mathbf{x})$ em cada região de $\mathbb{R}^d$.

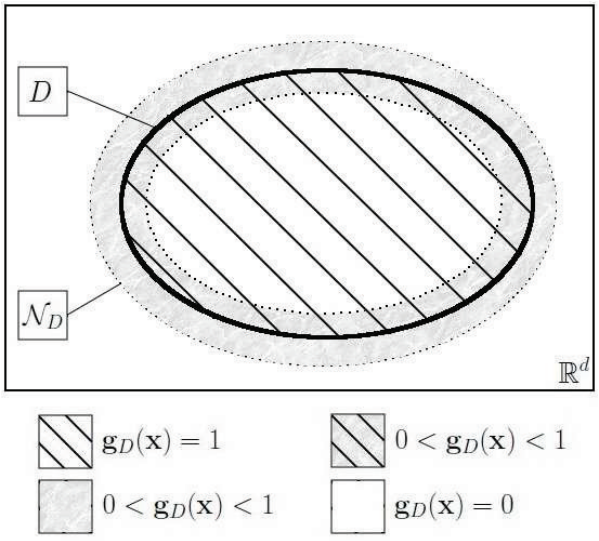


Figura 3.1: Valores da função de transição por regiões.

Proposição 3.1.2 *Sejam* $\mathbf{g}_A$ *e* $\mathbf{g}_B$ *funções de transição com relação às regiões* $A \in \mathbb{R}^d$ *e* $B \in \mathbb{R}^d$*, respectivamente. As funções* $\mathbf{g}_A\mathbf{g}_B$ *e* $1 - \mathbf{g}_A$ *definidas a seguir também são funções de transição com relação à regiões em* $\mathbb{R}^d$*.*

$$\begin{array}{rccc} \mathbf{g}_A\mathbf{g}_B: & \mathbb{R}^d & \longrightarrow & [0,1] \\ & \mathbf{x} & \longmapsto & \mathbf{g}_A(\mathbf{x})\mathbf{g}_B(\mathbf{x}) \end{array} \qquad \begin{array}{rccc} 1-\mathbf{g}_A: & \mathbb{R}^d & \longrightarrow & [0,1] \\ & \mathbf{x} & \longmapsto & 1-\mathbf{g}_A(\mathbf{x}) \end{array}$$

Demonstração:

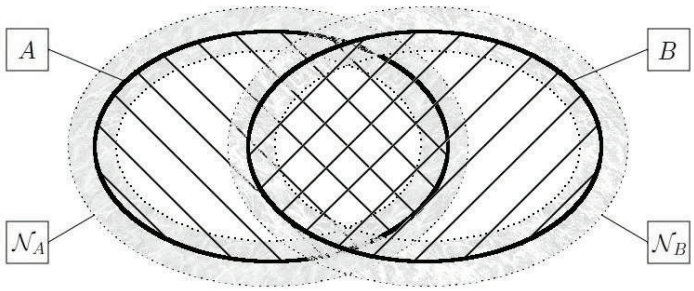


Figura 3.2: Diagrama das regiões definidas pelos conjuntos A e B.

Primeiro veja que $\mathbf{g}_A\mathbf{g}_B$ é função de transição com relação à $A \cap B$, onde $\mathcal{N}_{A\cap B} = (\mathcal{N}_A \cap (B \cup \mathcal{N}_B)) \cup (\mathcal{N}_B \cap (A \cup \mathcal{N}_A))$:

- Se $\mathbf{x} \notin (A \cap B) \cup \mathcal{N}_{A\cap B} \Rightarrow \mathbf{x} \notin \mathbf{A} \cup \mathcal{N}_A$ ou $\mathbf{x} \notin B \cup \mathcal{N}_B \Rightarrow$ $\mathbf{g}_A(\mathbf{x}) = 0$ ou $\mathbf{g}_B(\mathbf{x}) = 0 \Rightarrow \mathbf{g}_A\mathbf{g}_B(\mathbf{x}) = \mathbf{g}_A(\mathbf{x})\mathbf{g}_B(\mathbf{x}) = 0$;
- Se $\mathbf{x} \in \mathcal{N}_{A\cap B} \Rightarrow \mathbf{x} \in \mathcal{N}_A \cap (B \cup \mathcal{N}_B)$ ou $\mathbf{x} \in \mathcal{N}_B \cap (A \cup \mathcal{N}_A) \Rightarrow$ $(0 < \mathbf{g}_A(\mathbf{x}) < 1$ e $\mathbf{g}_B(\mathbf{x}) > 0)$ ou $(0 < \mathbf{g}_B(\mathbf{x}) < 1$ e $\mathbf{g}_A(\mathbf{x}) > 0) \Rightarrow$ $\mathbf{g}_A\mathbf{g}_B(\mathbf{x}) = \mathbf{g}_A(\mathbf{x})\mathbf{g}_B(\mathbf{x}) > 0$;
- Se $\mathbf{x} \in (A \cap B) - \mathcal{N}_{A\cap B} \Rightarrow \mathbf{x} \in \mathbf{A} - \mathcal{N}_A$ e $\mathbf{x} \in B - \mathcal{N}_B \Rightarrow$ $\mathbf{g}_A(\mathbf{x}) = 1$ e $\mathbf{g}_B(\mathbf{x}) = 1 \Rightarrow \mathbf{g}_A\mathbf{g}_B(\mathbf{x}) = \mathbf{g}_A(\mathbf{x})\mathbf{g}_B(\mathbf{x}) = 1$.

Para terminar, veja que $1 - \mathbf{g}_A$ é função de transição com relação à A^c, onde $\mathcal{N}_{A^c} = \mathcal{N}_A$:

- Se $\mathbf{x} \notin A^c \cup \mathcal{N}_{A^c} \Rightarrow \mathbf{x} \in \mathbf{A} - \mathcal{N}_{A^c} \Rightarrow \mathbf{x} \in \mathbf{A} - \mathcal{N}_A \Rightarrow \mathbf{g}_A(\mathbf{x}) = 1 \Rightarrow$ $(1 - \mathbf{g}_A)(\mathbf{x}) = 1 - \mathbf{g}_A(\mathbf{x}) = 1 - 1 = 0$;
- Se $\mathbf{x} \in \mathcal{N}_{A^c} \Rightarrow \mathbf{x} \in \mathcal{N}_A \Rightarrow 0 < \mathbf{g}_A(\mathbf{x}) < 1 \Rightarrow 0 > -\mathbf{g}_A(\mathbf{x}) > -1 \Rightarrow$ $1 > 1 - \mathbf{g}_A(\mathbf{x}) > 0 \Rightarrow 0 < (1 - \mathbf{g}_A)(\mathbf{x}) < 1$;
- Se $\mathbf{x} \in A^c - \mathcal{N}_{A^c} \Rightarrow \mathbf{x} \notin A \cup \mathcal{N}_A \Rightarrow \mathbf{g}_A(x) = 0 \Rightarrow$ $(1 - g_A)(\mathbf{x}) = 1 - \mathbf{g}_A(\mathbf{x}) = 1$.

□

PUC-Rio - Certificação Digital Nº 0521445/CA

Na aplicação em árvores de regressão, em particular, algumas funções de transição são mais interessantes que outras. Entre muitas maneiras de construí-las a apresentada a seguir é um exemplo que satisfaz as características desejadas nesse trabalho.

Seja ∂D o bordo de uma região D e $\mathbf{d}_D : \mathbb{R}^d \to \mathbb{R}$ uma função distância com relação ao bordo, contínua, com sinal e que satisfaz

$$\mathbf{d}_D(\mathbf{x}) \begin{cases} > 0, & \text{se } \mathbf{x} \in D - \partial D \\ = 0, & \text{se } \mathbf{x} \in \partial D \\ < 0, & \text{se } \mathbf{x} \in D^c \end{cases}$$

Se o bordo de D for definido implicitamente por uma função F, $\partial D = \{\mathbf{x} \in \mathbb{R}^d \mid F(\mathbf{x}) = 0\}$, um exemplo para $\mathbf{d}_D$ pode ser a distância algébrica, definida por $\mathbf{d}_D(\mathbf{x}) = F(\mathbf{x})$. Caso seja possível calcular a distância geométrica, ela também poderá ser usada para definir $\mathbf{d}_D$, mas nesse caso será necessário considerar o sinal positivo para os pontos em D e negativo para aqueles fora de D.

Seja $\mathbf{s} : \mathbb{R} \to [0, 1]$ uma função degrau, contínua, não-decrescente, tal que $\lim_{t\to-\infty} \mathbf{s}(t) = 0$, $\lim_{t\to\infty} \mathbf{s}(t) = 1$, $\mathbf{s}(0) = \frac{1}{2}$ e $\mathbf{s}(t; n) + \mathbf{s}(-t; n) = 1$. Um exemplo de função degrau com essas características é a função logística, definida por:

$$\mathbf{s_1}(t; \gamma) = \frac{1}{1 + \mathbf{e}^{-\gamma t}}. \tag{3-1}$$

Também pode ser citada como exemplo a função *Heaviside*, definida por:

$$\mathbf{s_2}(t; 0) = \begin{cases} 0, & \text{se } t < 0 \\ \frac{1}{2}, & \text{se } t = 0 \\ 1, & \text{se } t > 0 \end{cases} \qquad \begin{aligned} \mathbf{s_2}(t; n) = \tfrac{1}{2} \big(\big(1 + \tfrac{t}{n}\big)\, \mathbf{s_2}(t + 1; n - 1) + \\ \big(1 - \tfrac{t}{n}\big)\, \mathbf{s_2}(t - 1; n - 1)\big) \\ n = 1, 2, 3, \ldots \end{aligned} \tag{3-2}$$

Além das características já comentadas é interessante dizer que, de acordo com o trabalho de Li e Tian (34), a *Heaviside* também satisfaz:

- $\mathbf{s_2}(\,.\,; n)$ é de classe C^{n-1}, para $n > 1$;
- $\mathbf{s_2}(\,.\,; n)$ é uma função polinomial por partes $\forall\ n$;
- $\mathbf{s_2}(t; n) = 1$, se $t \geq n$; $\mathbf{s_2}(t; n) = 0$, se $t \leq -n$;

O último exemplo de função degrau é a função polinomial definida na equação (3-3) a seguir:

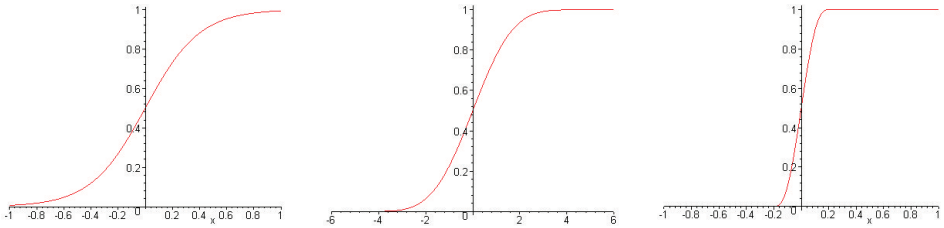
$$\mathbf{s_3}(t; \lambda) = \begin{cases} 0, & \text{se } t \leq -\frac{1}{\lambda} \\ 1, & \text{se } t \geq \frac{1}{\lambda} \\ \frac{1}{2} + \frac{15\lambda}{16} t - \frac{5\lambda^3}{8} t^3 + \frac{3\lambda^5}{16} t^5, & \text{caso contrário} \end{cases} \tag{3-3}$$

PUC-Rio - Certificação Digital Nº 0521445/CA

Esta função foi apresentada por Lage, Bordignon, Petronetto, Veiga, Tavares, Lewiner e Lopes em (30). Lá é possível verificar que s_3 também satisfaz as seguintes características desejadas:

- $\mathbf{s_3}(\ .\ ;\lambda)$ é contínua e tem derivada contínua $\forall\ \lambda$, ou seja, é de classe C^1;
- $\mathbf{s_3}(\ .\ ;\lambda)$ é uma função polinomial por partes $\forall\ \lambda$;
- $\mathbf{s_3}(t;\lambda) = 0 \Longleftrightarrow t \leq -\frac{1}{\lambda}$; $\mathbf{s_3}(t;\lambda) = 1 \Longleftrightarrow t \geq \frac{1}{\lambda}$;

A figura 3.3 a seguir ilustra os três exemplos de função degraus citados:



3.3(a): $\mathbf{s_1}$ com $\gamma = 5$. 3.3(b): $\mathbf{s_2}$ com $n = 5$. 3.3(c): $\mathbf{s_3}$ com $\lambda = 5$.

Figura 3.3: Gráficos dos exemplos citados de função degrau.

Proposição 3.1.3 *A composta* $\mathbf{g}_D = \mathbf{s} \circ \mathbf{d}_D$ *é uma função de transição contínua com relação à região* D.

Demonstração:

$$\begin{array}{rccccc} \mathbf{g}_D: & \mathbb{R}^d & \longrightarrow & \mathbb{R} & \longrightarrow & [0,1] \\ & \mathbf{x} & \longmapsto & \mathbf{d}_D(\mathbf{x}) & \longmapsto & \mathbf{s}(\mathbf{d}_D(\mathbf{x})) \end{array}$$

Seja $\mathcal{N}_D = \{\mathbf{x} \in \mathbb{R}^d \mid 0 < \mathbf{s}(\mathbf{d}_D(\mathbf{x})) < 1\}$ a vizinhança de ∂D.

- Se $\mathbf{x} \notin D \cup \mathcal{N}_D \ \Rightarrow\ \mathbf{x} \notin D$ e $\mathbf{x} \notin \mathcal{N}_D$.
  Como $\mathbf{x} \notin D \ \Rightarrow \mathbf{d}_D(\mathbf{x}) < 0 \ \Rightarrow\ \mathbf{s}(\mathbf{d}_D(\mathbf{x})) < \frac{1}{2}$.
  Além disso, $\mathbf{x} \notin \mathcal{N}_D \ \Rightarrow\ \mathbf{s}(\mathbf{d}_D(\mathbf{x})) = 0$ ou $\mathbf{s}(\mathbf{d}_D(\mathbf{x})) = 1$.
  Então, $\mathbf{s}(\mathbf{d}_D(\mathbf{x})) = 0 \ \Rightarrow\ \mathbf{g}_D(\mathbf{x}) = 0$.
- Se $\mathbf{x} \in \mathcal{N}_D \ \Rightarrow\ 0 < \mathbf{g}_D(\mathbf{x}) < 1$ pela própria definição de $\mathcal{N}_D$.
- Se $\mathbf{x} \in D - \mathcal{N}_D \ \Rightarrow\ \mathbf{x} \in D$ e $\mathbf{x} \notin \mathcal{N}_D$.
  Como $\mathbf{x} \in D \ \Rightarrow\ \mathbf{d}_D(\mathbf{x}) \geq 0 \ \Rightarrow\ \mathbf{s}(\mathbf{d}_D(\mathbf{x})) \geq \frac{1}{2}$.
  Além disso, $\mathbf{x} \notin \mathcal{N}_D \ \Rightarrow\ \mathbf{s}(\mathbf{d}_D(\mathbf{x})) = 0$ ou $\mathbf{s}(\mathbf{d}_D(\mathbf{x})) = 1$.
  Então, $\mathbf{s}(\mathbf{d}_D(\mathbf{x})) = 1 \ \Rightarrow\ \mathbf{g}_D(\mathbf{x}) = 1$.

Para finalizar, o grau de diferenciabilidade de $\mathbf{g}_D$ vai depender do grau de diferenciabilidade de $\mathbf{s}$ e $\mathbf{d}_D$. Como, pelas definições, $\mathbf{s}$ e $\mathbf{d}_D$ são contínuas, $\mathbf{g}_D$ também será uma função contínua, uma vez que ela é a composta de duas funções contínuas.

□

PUC-Rio - Certificação Digital Nº 0521445/CA

3.2
Aplicação em Árvores

Suponha a estrutura de árvore binária definida no capítulo anterior, onde cada nó interno define implicitamente o bordo de uma região em $\mathbb{R}^d$ por $\partial D = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{p}(\mathbf{x}) = 0\}$. Sejam $D = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{p}(\mathbf{x}) \leq 0\}$ e $D^c = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{p}(\mathbf{x}) > 0\}$ as regiões que caracterizam os filhos à esquerda e à direita desse nó. Para cada uma dessas regiões é construída uma função de transição, $\mathbf{g}_D$ e $\mathbf{g}_{D^c}$. A construção de $\mathbf{g}_D$ e $\mathbf{g}_{D^c}$ é feita de acordo com a seção anterior, onde as funções distância são as distâncias algébricas definidas por $\mathbf{d}_D(\mathbf{x}) = -\mathbf{p}(\mathbf{x})$, $\mathbf{d}_{D^c}(\mathbf{x}) = \mathbf{p}(\mathbf{x})$ e $\mathbf{s}$ está definida em (3-3).

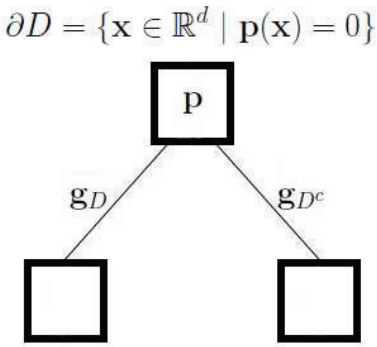
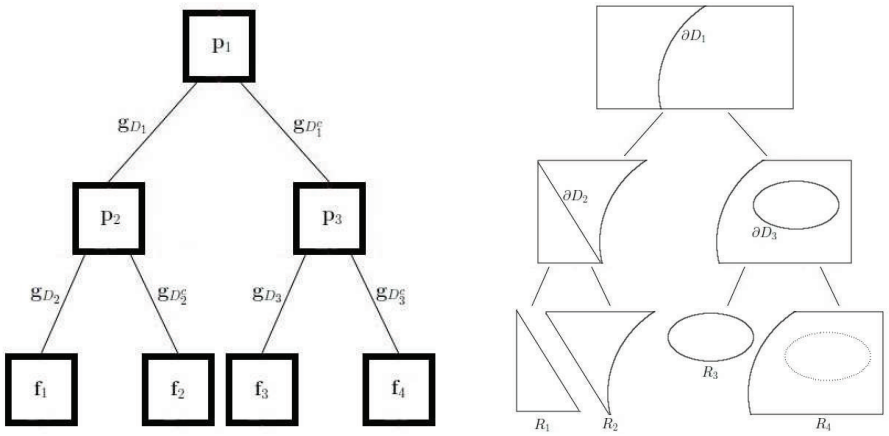


Figura 3.4: Criação de cada nó interno da árvore.

Como já foi comentado, seja R_k a região definida pela folha k, que é a interseção entre as regiões definidas pelo caminho entre a raiz e a folha em questão. A função de transição $\mathbf{G}_k$, associada à essa região, pode ser definida pelo produto das funções de transição pelo caminho da raiz até essa folha. Como na figura 2.1 do capítulo anterior, a figura 3.5 a seguir apresenta um exemplo de estrutura de árvore e faz sua relação com as funções de transição. Para esse exemplo,

$$\begin{aligned}
\mathbf{G}_1(\mathbf{x}) &= \mathbf{g}_{D_1}(\mathbf{x})\mathbf{g}_{D_2}(\mathbf{x}) \\
\mathbf{G}_2(\mathbf{x}) &= \mathbf{g}_{D_1}(\mathbf{x})\mathbf{g}_{D_2^c}(\mathbf{x}) = \mathbf{g}_{D_1}(\mathbf{x})(1 - \mathbf{g}_{D_2}(\mathbf{x})) \\
\mathbf{G}_3(\mathbf{x}) &= \mathbf{g}_{D_1^c}(\mathbf{x})\mathbf{g}_{D_3}(\mathbf{x}) = (1 - \mathbf{g}_{D_1}(\mathbf{x}))\mathbf{g}_{D_3}(\mathbf{x}) \\
\mathbf{G}_4(\mathbf{x}) &= \mathbf{g}_{D_1^c}(\mathbf{x})\mathbf{g}_{D_3^c}(\mathbf{x}) = (1 - \mathbf{g}_{D_1}(\mathbf{x}))(1 - \mathbf{g}_{D_2}(\mathbf{x}))
\end{aligned}$$

PUC-Rio - Certificação Digital Nº 0521445/CA



3.5(a): Árvore padrão. 3.5(b): Árvore por regiões.

Figura 3.5: Exemplo da aplicação de suavidade em árvores de regressão.

De forma geral, para uma árvore qualquer,

$$\mathbf{G}_k(\mathbf{x}) = \prod_{\mathbf{g}_i \in \mathcal{G}_k} \mathbf{g}_i(\mathbf{x}) \tag{3-4}$$

é a função de transição associada a região definida pela folha k, onde $\mathcal{G}_k =$ conjunto das funções de transição pelo caminho da raiz até a folha k.

Proposição 3.2.1 *Seja uma árvore com l folhas. Então, $\sum_{k=1}^{l} \mathbf{G}_k(\mathbf{x}) = 1$ para todo $\mathbf{x} \in \mathbb{R}^d$.*

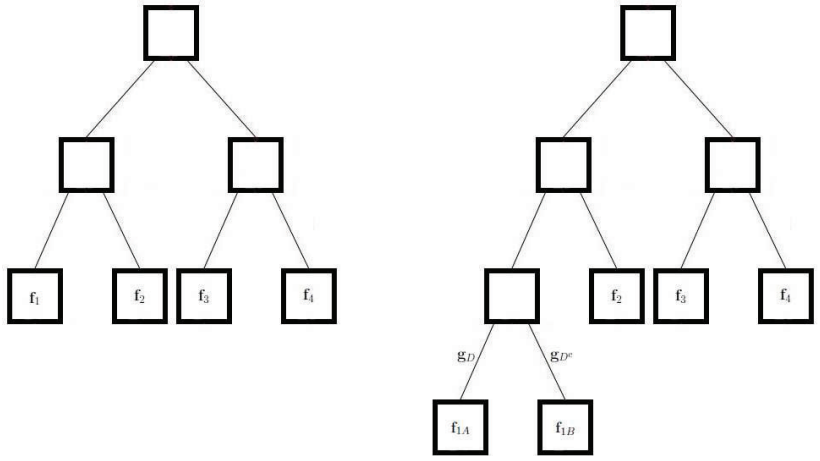
Demonstração:

A proposição será demonstrada por indução. Primeiro o caso base: uma árvore com um único nó interno e duas folhas. Sejam D e D^c as regiões definidas pelo nó interno. Nesse caso, $\mathbf{G}_1(\mathbf{x}) = \mathbf{g}_D(\mathbf{x})$ e $\mathbf{G}_2(\mathbf{x}) = \mathbf{g}_{D^c}(\mathbf{x}) = 1 - \mathbf{g}_D(\mathbf{x})$. Então, $\mathbf{G}_1(\mathbf{x}) + \mathbf{G}_2(\mathbf{x}) = \mathbf{g}_D(\mathbf{x}) + \mathbf{g}_{D^c}(\mathbf{x}) = \mathbf{g}_D(\mathbf{x}) + 1 - \mathbf{g}_D(\mathbf{x}) = 1$.

Agora suponha que seja verdade para um árvore qualquer com l folhas. Ou seja, $\sum_{k=1}^{l} \mathbf{G}_k(\mathbf{x}) = 1$.

Para incrementar essa árvore basta substituir uma folha por um nó interno mais suas duas folhas descendentes. Sem perdas, substitua a folha $\mathbf{f}_1$ por um nó interno e adicione as folhas $\mathbf{f}_{1A}$ e $\mathbf{f}_{1B}$, como mostra a figura 3.6. Assim a árvore passa a ter $l + 1$ folhas: $1\mathrm{A}, 1B, 2, \ldots, l$. Sejam D e D^c as regiões definidas pelo novo nó interno. Dessa forma, $\mathbf{G}_{1A}(\mathbf{x}) = \mathbf{g}_D(\mathbf{x})\mathbf{G}_1(\mathbf{x})$ e $\mathbf{G}_{1B}(\mathbf{x}) = \mathbf{g}_{D^c}(\mathbf{x})\mathbf{G}_1(\mathbf{x}) = (1 - \mathbf{g}_D(\mathbf{x}))\mathbf{G}_1(\mathbf{x})$.

PUC-Rio - Certificação Digital Nº 0521445/CA



3.6(a): Árvore com l folhas. 3.6(b): Árvore com $l+1$ folhas.

Figura 3.6: Ilustração do incremento no número de folhas.

Então, para essa nova árvore, com $l+1$ folhas, a soma de todas as funções de transição associadas as regiões definidas pelas folhas é:

$$\begin{aligned}
\mathbf{G}_{1A}(\mathbf{x}) + \mathbf{G}_{1B}(\mathbf{x}) + \textstyle\sum_{k=2}^{l} \mathbf{G}_k(\mathbf{x}) &= \\
\mathbf{g}_D(\mathbf{x})\mathbf{G}_1(\mathbf{x}) + \mathbf{g}_{D^c}(\mathbf{x})\mathbf{G}_1(\mathbf{x}) + \textstyle\sum_{k=2}^{l} \mathbf{G}_k(\mathbf{x}) &= \\
\mathbf{g}_D(\mathbf{x})\mathbf{G}_1(\mathbf{x}) + (1 - \mathbf{g}_D(\mathbf{x}))\mathbf{G}_1(\mathbf{x}) + \textstyle\sum_{k=2}^{l} \mathbf{G}_k(\mathbf{x}) &= \\
\mathbf{G}_1(\mathbf{x})(\mathbf{g}_D(\mathbf{x}) + 1 - \mathbf{g}_D(\mathbf{x})) + \textstyle\sum_{k=2}^{l} \mathbf{G}_k(\mathbf{x}) &= \\
\mathbf{G}_1(\mathbf{x}) + \textstyle\sum_{k=2}^{l} \mathbf{G}_k(\mathbf{x}) &= \\
\textstyle\sum_{k=1}^{l} \mathbf{G}_k(\mathbf{x}) &= 1
\end{aligned}$$

□

3.3 Regressão Construtiva nas Regiões Implícitas

Seguindo ainda a estrutura de árvore definida no capítulo anterior, suponha que cada folha k guarda uma regressão em $R^k \subset \mathbb{R}^d$, isto é, uma função $\mathbf{f}_k : \mathbb{R}^d \to \mathbb{R}$. A estimativa final dada pela árvore de regressão do método RCRI pode ser expressa por:

$$\widehat{\mathbf{F}}(\mathbf{x}) = \sum_{k=1}^{l} \mathbf{G}_k(\mathbf{x})\mathbf{f}_k(\mathbf{x}) \ , \tag{3-5}$$

onde a função de transição $\mathbf{G}_k$ está definida na equação (3-4).

Esse estimador tem diversas vantagens, entre elas o fato de ser um estimador de classe C^m, onde m é a menor entre as classes de $\mathbf{G}_k$ e $\mathbf{f}_k$, $k = 1, \ldots, l$. Além disso, como $\{\mathbf{G}_k\}_{k=1}^{l}$ representa uma partição da unidade,

PUC-Rio - Certificação Digital Nº 0521445/CA

isto é, $\sum_{k=1}^{l} \mathbf{G}_k(\mathbf{x}) = 1$, o erro global pode ser controlado através do controle do erro local, aplicando a proposição 3.3.1 a seguir.

Proposição 3.3.1 *Seja* $\mathbf{F}$ *uma função real definida em* $\mathbb{R}^d$ *e* $\{D_k\}_{k=1}^{l}$ *uma partição de* $\mathbb{R}^d$. *Para cada* $k = 1, \ldots, l$ *sejam* $\mathbf{g}_k$ *a função de transição com relação à região* D_k *e* $\mathcal{N}_k$ *a vizinhança de* D_k *definida por* $\mathbf{g}_k$, *ou seja,* $\mathcal{N}_k = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{x} \notin D_k \text{ e } \mathbf{g}_k(\mathbf{x}) \neq 0\}$.

Se para cada $k = 1, \ldots, l$ $\mathbf{f}_k$ *é uma aproximação local de* $\mathbf{F}$ *tal que* $\mid \mathbf{F}(\mathbf{x}) - \mathbf{f}_k(\mathbf{x}) \mid < \varepsilon,\ \forall\ \mathbf{x} \in D_k \cup \mathcal{N}_k$ *e* $\sum_{k=1}^{l} \mathbf{g}_k(\mathbf{x}) = 1$, *então a aproximação global definida por* $\widehat{\mathbf{F}}(\mathbf{x}) = \sum_{k=1}^{l} \mathbf{g}_k(\mathbf{x})\mathbf{f}_k(\mathbf{x})$ *satisfaz a seguinte propriedade:*

$$\left| \mathbf{F}(\mathbf{x}) - \widehat{\mathbf{F}}(\mathbf{x}) \right| < \varepsilon\ \forall\ \mathbf{x} \in \mathbb{R}^d$$

Demonstração:

$$\begin{aligned}
\left| \mathbf{F}(\mathbf{x}) - \widehat{\mathbf{F}}(\mathbf{x}) \right| &= \left| \mathbf{F}(\mathbf{x}) - \textstyle\sum_{k=1}^{l} \mathbf{g}_k(\mathbf{x})\mathbf{f}_k(\mathbf{x}) \right| = \\
&\left| \left(\textstyle\sum_{k=1}^{l} \mathbf{g}_k(\mathbf{x})\right) \mathbf{F}(\mathbf{x}) - \textstyle\sum_{k=1}^{l} \mathbf{g}_k(\mathbf{x})\mathbf{f}_k(\mathbf{x}) \right| = \\
&\left| \textstyle\sum_{k=1}^{l} \mathbf{g}_k(\mathbf{x})\mathbf{F}(\mathbf{x}) - \textstyle\sum_{k=1}^{l} \mathbf{g}_k(\mathbf{x})\mathbf{f}_k(\mathbf{x}) \right| = \\
&\left| \textstyle\sum_{k=1}^{l} \mathbf{g}_k(\mathbf{x}) \left(\mathbf{F}(\mathbf{x}) - \mathbf{f}_k(\mathbf{x})\right) \right| \leq \\
&\textstyle\sum_{k=1}^{l} \mid \mathbf{g}_k(\mathbf{x}) \left(\mathbf{F}(\mathbf{x}) - \mathbf{f}_k(\mathbf{x})\right) \mid = \\
&\textstyle\sum_{k=1}^{l} \mathbf{g}_k(\mathbf{x}) \mid \mathbf{F}(\mathbf{x}) - \mathbf{f}_k(\mathbf{x}) \mid\ <\ \textstyle\sum_{k=1}^{l} \mathbf{g}_k(\mathbf{x})\varepsilon\ =\ \varepsilon
\end{aligned}$$

□

Nada ainda foi dito sobre o formato dessas regressões, porém o fato da equação (2-1) incluir a média amostral $\bar{y}$ indica uma tendência para que essa regressão seja uma função constante. Se for desejado que a regressão seja linear, ou de qualquer outro formato, basta fazer algumas alterações na equação (2-1) de forma que a partição definida pelo polinômio $\mathbf{p}$ separe quem está super-estimado de quem esta sub-estimado, em vez de separar quem está acima ou abaixo da média.

Logo, cada $\mathbf{f}_k$ é uma função constante, $\mathbf{f}_k(\mathbf{x}) = \mathbf{c}_k$, e o estimador gerado pela árvore é dado por:

$$\widehat{\mathbf{F}}(\mathbf{x}) = \sum_{k=1}^{l} \mathbf{G}_k(\mathbf{x})\mathbf{c}_k\ . \tag{3-6}$$

PUC-Rio - Certificação Digital Nº 0521445/CA

A seguir são apresentadas duas maneiras diferentes de determinar os valores de c_k. A primeira delas mostra a forma local, onde para estimar c_k só são levados em consideração os pontos dentro da região $R_k \cup \mathcal{N}_{R_k}$. A segunda mostra uma estimativa global, onde todos os valores c_k, $k = 1, 2, \ldots, l$, são estimados ao mesmo tempo e por isso todos os pontos da amostra são levados em consideração.

Continuando com a notação $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^d$ e $Y = \{y_1, y_2, \ldots, y_N\} \subset \mathbb{R}$ para os dados da amostra, suponha que já foi feito todo o processo descrito na seção 2.2, ou seja, a estrutura da árvore já está pronta, já se sabe quantos nós, quais os polinômios em cada nó, e quantas folhas ela tem. Também considere o vetor de coeficientes $\mathbf{c} = (c_1, \ldots, c_l)^{\mathbf{t}}$, que guarda os elementos a serem estimados nessa seção.

Estimativa Local

Como já foi comentado, o modelo local para estimar c_k leva em consideração apenas os pontos na região definida pela folha k e sua vizinhança, isto é, $R_k \cup \mathcal{N}_{R_k}$. Para ser mais preciso, os pontos $\mathbf{x}$ tais que $G_k(\mathbf{x}) \neq 0$.

O estimador de cada elemento do vetor $\mathbf{c}$ é simplesmente a média ponderada pela função de transição $\mathbf{G}_k$. Dessa forma $\widehat{\mathbf{c}} = (\widehat{c_1}, \ldots, \widehat{c_l})$ e

$$\widehat{c_k} = \frac{\sum_{i=1}^{N} \mathbf{G}_k(\mathbf{x}_i) y_i}{\sum_{i=1}^{N} \mathbf{G}_k(\mathbf{x}_i)} \tag{3-7}$$

Se for garantido, para toda folha k, um erro local menor que δ, ou seja, se $\forall\ 1 \leq k \leq l$ e $\forall\ i$ tal que $\mathbf{x}_i \in R_k \cup \mathcal{N}_{R_k}$ tem-se $\text{erro}_i = |y_i - \widehat{c_k}| < \gamma$, então, pela aplicação da proposição 3.3.1, é garantido que o erro global também será menor que γ. Dessa forma, se for interessante definir um limite superior para o erro global, basta adicionar um critério de parada que controle o erro local.

Estimativa Global

A estimativa global se difere da local principalmente pelo fato de que nela $\mathbf{c}$ é estimado de forma a minimizar o erro total da amostra. A primeira hipótese a ser feita é que $\widehat{\mathbf{F}}$ representa a relação entre $\mathbf{x}_i$ e y_i, isto é:

$$y_i = \widehat{\mathbf{F}}(\mathbf{x}_i) + \varepsilon_i = \sum_{k=1}^{l} \mathbf{G}_k(\mathbf{x}_i)\mathbf{c}_k + \varepsilon_i \qquad \forall\ 1 \leq i \leq N$$

onde ε_i é uma variável aleatória normal, de média nula e variância σ^2, que representa o erro na estimativa.

PUC-Rio - Certificação Digital Nº 0521445/CA

O vetor de coeficientes $\mathbf{c} = (c_1, \ldots, c_l)^{\mathbf{t}}$ é estimado por mínimos quadrados, como descreve o apêndice A. Para isso é resolvido o seguinte sistema, apresentado em A-3,

$$\mathbf{G^t G c} = \mathbf{G^t y} \tag{3-8}$$

onde $\mathbf{G}$, $\mathbf{c}$ e $\mathbf{y}$ são definidos por:

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}_1(\mathbf{x}_1) & \mathbf{G}_2(\mathbf{x}_1) & \ldots & \mathbf{G}_l(\mathbf{x}_1) \\ \mathbf{G}_1(\mathbf{x}_2) & \mathbf{G}_2(\mathbf{x}_2) & \ldots & \mathbf{G}_l(\mathbf{x}_2) \\ \vdots & \vdots & & \vdots \\ \mathbf{G}_1(\mathbf{x}_N) & \mathbf{G}_2(\mathbf{x}_N) & \ldots & \mathbf{G}_l(\mathbf{x}_N) \end{pmatrix} \quad \mathbf{c} = \begin{pmatrix} c_1 \\ \vdots \\ c_l \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

Dessa forma,

$$\widehat{\mathbf{c}} = (\mathbf{G^t G})^{-1}\, \mathbf{G^t y}$$

é o estimador dos parâmetros em $\mathbf{c}$. Mais detalhes sobre suas condições de existência, tendência e consistência podem ser encontrados no apêndice A.

Exemplo 3.3.2 Pode-se determinar a regressão dada pela árvore a partir dos dados da amostra do exemplo 2.2.1, exposto no capítulo anterior. A figura 3.7 a seguir mostra o resultado usando tanto a regressão local (em azul) quanto a regressão global (em vermelho). Apenas com esse exemplo não é possível afirmar que uma forma é mais interessante que a outra. O que se pode dizer é que as duas estimativas são similares e identificam os diferentes comportamentos da amostra dentro do domínio.

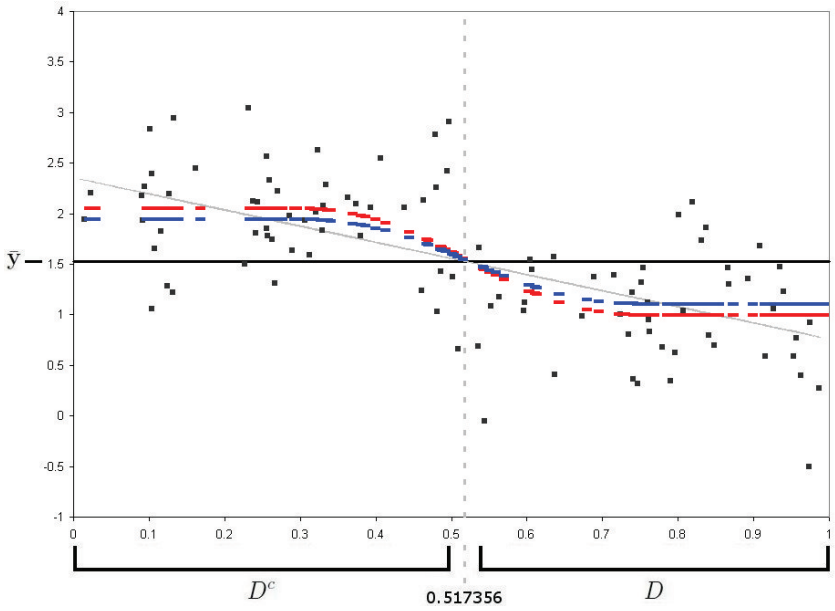


Figura 3.7: Exemplo das regressões local e global para uma dimensão.

PUC-Rio - Certificação Digital Nº 0521445/CA

Suavizar a estimativa de uma árvore de regressão por meio de funções de transição, ou outras funções semelhantes, é uma prática comum. Por exemplo, como já foi comentado, da Rosa, Veiga e Medeiros (11) utilizaram essa técnica na criação do STR-Tree. A construção da função de transição deles é bem semelhante à descrita nesse trabalho, porém eles escolheram como $\mathbf{s}$ a função logística. Já Chaudhuri, Huang, Loh e Yao criam em (9) o modelo SUPPORT, onde a suavidade é feita por meio de uma função, que não é uma função de transição, chamada por eles de "média ponderada".

A primeira vantagem em usar a função (3-3) no lugar da logística é que sua forma é polinomial, assim ela é mais rápida de ser avaliada, pois funções exponenciais são caras computacionalmente. Além disso, o que é ainda mais interessante, é que no caso da função definida em (3-3) o conjunto $\{t \in \mathbb{R} \mid \mathbf{s}(t) \neq 0 \text{ e } \mathbf{s}(t) \neq 1\}$ é limitado, o que para um modelo recursivo pode significar mais eficiência, já que só serão levados em consideração as folhas que definem regiões próximas a $\mathbf{x}$.

PUC-Rio - Certificação Digital Nº 0521445/CA