2
Método de Regressão Construtiva em Regiões Implícitas

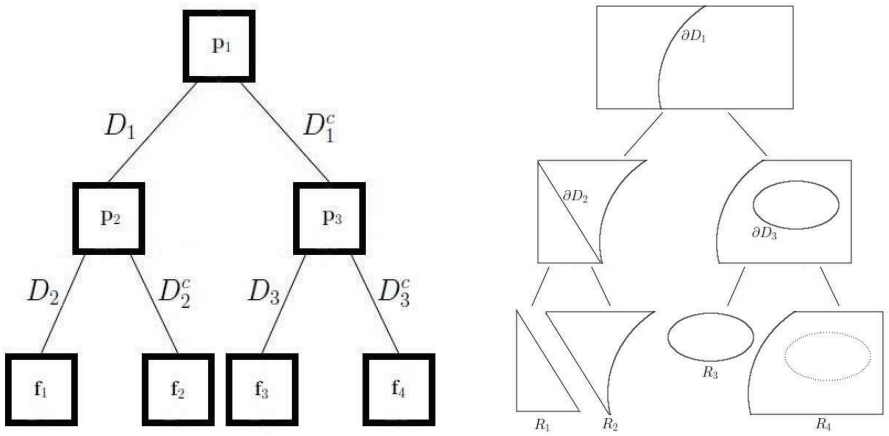
Esse capítulo tem como objetivo apresentar a estrutura e a construção da árvore do método de Regressão Construtiva em Regiões Implícitas (RCRI). De acordo com a introdução, o problema consiste em encontrar uma função f que descreva a relação entre $\mathbf{x}_i$ e y_i, a partir de uma amostra de tamanho N composta pelos dados

$$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \in \mathbb{R}^d \quad \text{e} \quad Y = \{y_1, y_2, \ldots, y_N\} \in \mathbb{R}.$$

2.1
Partição do Domínio e Caracterização da Árvore

A maioria dos trabalhos nesse tema divide o domínio apenas por retas paralelas aos eixos, como por exemplo os modelos CART (5), SUPPORT (9), STR-Tree (11), MARS (17), entre outros. Com isso as partições ficam restritas a conjuntos do tipo $\{\mathbf{x} \in \mathbb{R}^d \mid x_i \leq a\}$. A vantagem disso é que assim é possível realizar interpretações do modelo. Por outro lado, essas partições não são eficientes para separar dados cuja distribuição no domínio dependa de combinações das coordenadas. Além disso, para determinar em qual coordenada será feita a divisão, geralmente, é resolvido um problema de otimização para cada nó da árvore. Dessa forma a implementação é bem complexa e requer muito tempo de execução.

O método RCRI propõe algo diferente, onde as regiões podem assumir formas bem mais flexíveis, a fim de melhorar a regressão, com uma implementação razoavelmente simples. Ele é caracterizado por uma árvore binária onde cada nó interno define implicitamente o bordo de uma região em $\mathbb{R}^d$, através do conjunto de nível zero de uma função polinômio multivariada. Na verdade, o que cada nó interno guarda são os coeficientes de um polinômio $\mathbf{p}$. Dessa forma o bordo da região definida por ele pode ser expresso por $\partial D = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{p}(\mathbf{x}) = 0\}$. Observe que assim é realizada a seguinte partição do domínio: $\mathbb{R}^d = D \cup D^c$, onde $D = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{p}(x) \leq 0\}$, sub-árvore à esquerda, e $D^c = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{p}(x) > 0\}$, a sub-árvore à direita. Como ilustração veja a figura 2.1.



2.1(a): Árvore padrão. 2.1(b): Árvore por regiões.

Figura 2.1: Exemplo da árvore para o método RCRI.

Se for associada à cada folha k da árvore uma região R_k, definida pela interseção entre as regiões no caminho da raiz até a folha em questão, o conjunto de regiões associadas às folhas também será uma partição do domínio. Para o exemplo da figura 2.1 as regiões definidas nessa árvore são:

$$D_1 = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{p}_1(\mathbf{x}) \leq 0\} \quad D_1^c = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{p}_1(\mathbf{x}) > 0\}$$
$$D_2 = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{p}_2(\mathbf{x}) \leq 0\} \quad D_2^c = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{p}_2(\mathbf{x}) > 0\}$$
$$D_3 = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{p}_3(\mathbf{x}) \leq 0\} \quad D_3^c = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{p}_3(\mathbf{x}) > 0\}$$

$$R_1 = D_1 \cap D_2 \quad R_2 = D_1 \cap D_2^c \quad R_3 = D_1^c \cap D_3 \quad R_4 = D_1^c \cap D_3^c$$

2.2 Construção do Polinômio p

Nesse trabalho, como já foi comentado, a partição do domínio é definida implicitamente pelo bordo de uma região em $\mathbb{R}^d$, $\partial D = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{p}(\mathbf{x}) = 0\}$, onde $\mathbf{p}$ é um polinômio multivariado de grau g. Para encontrar os coeficientes desse polinômio $\mathbf{p}$ suponha, $\forall\ i, 1 \leq i \leq N$, a seguinte relação:

$$y_i = \bar{y} + \mathbf{p}(\mathbf{x}_i) + \varepsilon_i\ , \tag{2-1}$$

onde $\mathbf{p}$ é um polinômio de grau g, $\bar{y}$ é a média amostral dos valores em Y e ε_i são variáveis aleatórias iid, normais, com média zero e variância σ^2. Veja que $\mathbf{p}(\mathbf{x}_i)$ representa o quanto distante da média ficou y_i, a menos do erro ε_i. Se $D = \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{p}(\mathbf{x}) \leq 0\}$, D representa os pontos abaixo da média e D^c os pontos acima dela. Então a partição D e D^c tem como objetivo separar os pontos que ficaram acima da média daqueles que ficaram abaixo.

Exemplo 2.2.1 O gráfico apresentado na figura 2.2 mostra um exemplo em uma dimensão para um polinômio de grau 1, ou seja, $d = 1$ e $g = 1$.

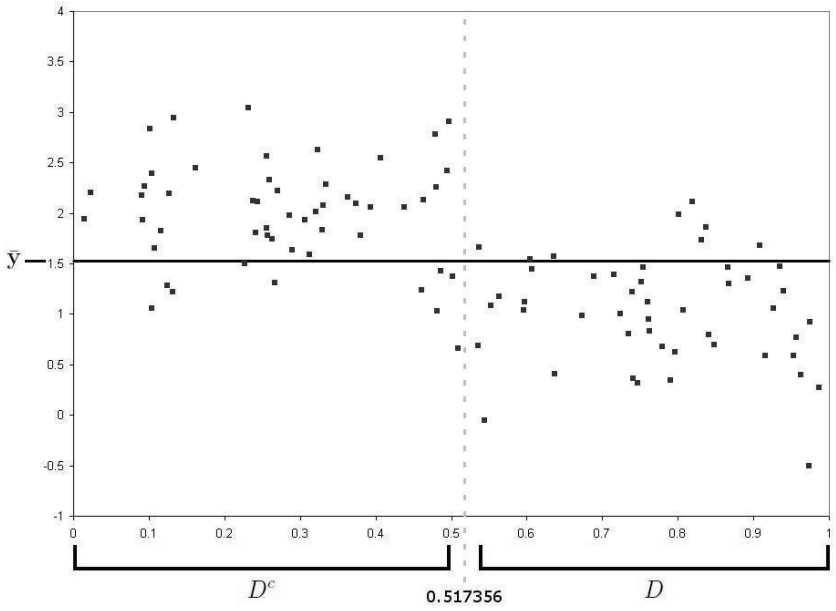


Figura 2.2: Exemplo de partição do domínio para uma dimensão.

A região D é definida a partir do polinômio $\mathbf{p}$. Em breve será explicado como encontrar os coeficientes desse polinômio, por enquanto, para esse exemplo, suponha $\mathbf{p}(x) = -1.59605x + 0.825725$. Então a região D será expressa por:

$$\begin{aligned} D &= \{x \in \mathbb{R} \mid \mathbf{p}(x) \leq 0\} \\ &= \{x \in \mathbb{R} \mid -1.59605x + 0.825725 \leq 0\} \\ &= \{x \in \mathbb{R} \mid x \geq \tfrac{-0.825725}{-1.59605}\} \\ &= \{x \in \mathbb{R} \mid x \geq 0.517356\} \end{aligned}$$

Número de Coeficientes do Polinômio

Antes de mostrar como serão determinados os coeficientes do polinômio $\mathbf{p}$ é preciso saber quantos coeficientes ele tem. Esse número depende do seu grau g e da dimensão d do seu domínio. Ele é expresso por uma equação de diferenças, como mostra a proposição a seguir.

Proposição 2.2.2 *Seja* $\mathbf{n}(g,d)$ *o número de coeficientes de um polinômio de grau* g*, homogeneo, com* d *variáveis. É verdade que:*

$$\begin{aligned}
&\mathbf{n}(1,d) = d+1\\
&\mathbf{n}(g,1) = g+1\\
&\mathbf{n}(g,d) = \mathbf{n}(g-1,d) + \mathbf{n}(g,d-1)
\end{aligned}$$

Demonstração:

Na demonstração será usada a seguinte nomenclatura: o coeficiente do monômio de grau zero será chamado de a; os coeficientes dos monômios de grau 1 serão chamados de a_i, $1 \le i \le d$; os coeficientes dos monômios de grau 2 serão chamados de a_{ij}, $1 \le i \le d$ e $i \le j \le d$; e assim por diante. O exemplo a seguir considera um polinômio $\mathbf{p}$ de grau 3 em $\mathbb{R}^3$, ou seja, $\mathbf{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$ e

$$\begin{aligned}
\mathbf{p}(\mathbf{x}) = \quad & a + && \text{grau 0}\\
& a_1x_1 + a_2x_2 + a_3x_3 + && \text{grau 1}\\
& a_{11}x_1^2 + a_{12}x_1x_2 + a_{13}x_1x_3 +\\
& \qquad a_{22}x_2^2 + a_{23}x_2x_3 +\\
& \qquad\qquad a_{33}x_3^2 + && \text{grau 2}\\
& a_{111}x_1^3 + a_{112}x_1^2x_2 + a_{113}x_1^2x_3 +\\
& \qquad a_{122}x_1x_2^2 + a_{123}x_1x_2x_3 +\\
& \qquad\qquad a_{133}x_1x_3^2 +\\
& \qquad a_{222}x_2^3 + a_{223}x_2^2x_3 +\\
& \qquad\qquad a_{233}x_2x_3^2 +\\
& \qquad\qquad\quad a_{333}x_3^3 && \text{grau 3}
\end{aligned}$$

Primeiro, a verificação das expressões $\mathbf{n}(1,d) = d+1$ e $\mathbf{n}(g,1) = g+1$ é imediata. Para mostrar que $\mathbf{n}(g,d) = \mathbf{n}(g-1,d) + \mathbf{n}(g,d-1)$ observe que $\mathbf{n}(g,d) - \mathbf{n}(g,d-1) =$ número de coeficientes adicionados quando o número de variáveis do polinômio, ou seja, a dimensão do domínio, passa de $d-1$ para d. Esses coeficientes a mais podem ser enumerados da seguinte forma:

$$\begin{aligned}
& a_d && \text{grau 1}\\
a_{1d},\ a_{2d},\ a_{3d},\ \ldots,\ a_{dd} && \text{grau 2}\\
a_{11d},\ a_{12d},\ a_{13d},\ \ldots,\ a_{1dd}\\
a_{22d},\ a_{23d},\ \ldots,\ a_{2dd}\\
\vdots\\
a_{ddd} && \text{grau 3}\\
& \vdots\\
& \vdots\\
a_{1\ldots11d},\ a_{1\ldots12d},\ a_{1\ldots13d},\ \ldots,\ a_{1\ldots1dd}\\
a_{1\ldots22d},\ a_{1\ldots23d},\ \ldots,\ a_{1\ldots2dd}\\
\vdots\\
a_{d\ldots dd} && \text{grau g}
\end{aligned}$$

Já os coeficientes de um polinômio de grau $g-1$ em $\mathbb{R}^d$, aqueles considerados em $\mathbf{n}(g-1,d)$, podem ser enumerados da seguinte forma:

$$\begin{array}{rl}
\left.\begin{array}{r} a \end{array}\right\} & \text{grau } 0 \\
\left.\begin{array}{r} a_1,\ a_2,\ a_3,\ \ldots,\ a_d \end{array}\right\} & \text{grau } 1 \\
\left.\begin{array}{r} a_{11},\ a_{12},\ a_{13},\ \ldots,\ a_{1d} \\ a_{22},\ a_{23},\ \ldots,\ a_{2d} \\ \vdots \\ a_{dd} \end{array}\right\} & \text{grau } 2 \\
\vdots & \\
\vdots & \\
\left.\begin{array}{r} a_{1\ldots11},\ a_{1\ldots12},\ a_{1\ldots13},\ \ldots,\ a_{1\ldots1d} \\ a_{1\ldots22},\ a_{1\ldots23},\ \ldots,\ a_{1\ldots2d} \\ \vdots \\ a_{d\ldots d} \end{array}\right\} & \text{grau } g-1
\end{array}$$

Observe que existe uma relação biunívoca entre os dois conjuntos de coeficientes. Se para os elementos do primeiro conjunto for desconsiderado o último d no índice dos coeficientes, serão obtidos exatamente os elementos do segundo conjunto. Dessa forma pode-se afirmar que a cardinalidade dos dois conjuntos é a mesma. Ou seja,

$$\mathbf{n}(g,d) - \mathbf{n}(g,d-1) = \mathbf{n}(g-1,d) \ \Rightarrow\ \mathbf{n}(g,d) = \mathbf{n}(g-1,d) + \mathbf{n}(g,d-1)$$

□

Estimativa dos Parâmetros

Agora, que já se sabe como determinar o número de coeficientes de um polinômio, é possível mostrar como esses coeficientes são estimados. Eles serão estimados através do estimador de mínimos quadrados lineares, apresentado no apêndice A. Para determinar o estimador basta resolver o sistema:

$$\mathbf{M^t M a} = \mathbf{M^t}\delta \tag{2-2}$$

onde $\mathbf{M}$, $\mathbf{a}$ e δ são definidos a seguir, considerando $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})$ e n o número de dados da amostra que chegaram ao nó (para o caso da raiz, $n = N$).

$$\delta = \begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix}_n \qquad M = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1d} & x_{11}^2 & x_{11}x_{12} & \ldots & x_{1d}^2 & x_{11}^3 & \ldots & x_{1d}^g \\ 1 & x_{21} & \ldots & x_{2d} & x_{21}^2 & x_{21}x_{22} & \ldots & x_{2d}^2 & x_{21}^3 & \ldots & x_{2d}^g \\ & & \vdots & & & & \vdots & & & \vdots & \\ 1 & x_{n1} & \ldots & x_{nd} & x_{n1}^2 & x_{n1}x_{n2} & \ldots & x_{nd}^2 & x_{n1}^3 & \ldots & x_{nd}^g \end{pmatrix}_{n \times \mathbf{n}(g,\ d)}$$

$$\mathbf{a^t} = \Big(\; a \; a_1 \; \ldots \; a_d \; a_{11} \; a_{12} \; \ldots \; a_{dd} \; a_{111} \; \ldots \; a_{dd\ldots d} \; \Big)_{\mathbf{n}(g,\; d)}$$

2.3
Critérios de Parada

Depois de determinar o polinômio $\mathbf{p}$, os dados de entrada X são particionados da seguinte forma:

$$\begin{array}{rclrcl} X_1 & = & \{\mathbf{x} \in X \mid \mathbf{p}(\mathbf{x}) \leq 0\} & X_2 & = & \{\mathbf{x} \in X \mid \mathbf{p}(\mathbf{x}) > 0\} \\ Y_1 & = & \{y_i \in Y \mid \mathbf{x}_i \in X_1\} & Y_2 & = & \{y_i \in Y \mid \mathbf{x}_i \in X_2\} \end{array}$$

Sejam n_1 e n_2 o número de elementos em X_1 e X_2 (ou em Y_1 e Y_2), respectivamente. Os critérios de parada a seguir servem para determinar se um nó deve ser dividido, dando origem a dois outros nós e virando um nó interno. Caso contrário, esse nó passa a ser uma folha.

1º Critério de Parada: Profundidade da Árvore

Esse é um critério bem simples, onde não será permitido que a árvore passe de uma profundidade pré-estabelecida. Dessa maneira procura-se evitar árvores muito profundas, que em geral resultam em estimativas ruins.

2º Critério de Parada: Número de Elementos

Esse critério serve para garantir que se $N \to \infty$, então $n \to \infty$, onde N = número total de dados na amostra e n = número de dados da amostra que chegam ao nó em questão. Dessa forma, a consistência do estimador, que é um comportamento assintótico, pode ser levada em consideração.

O objetivo deste critério é determinar um limite inferior, dependente do tamanho da amostra, para o número de elementos que chega a um nó qualquer. Seja $0 < q < 1$; com esse critério será garantido que $n > n_{lim} = qN$.

É importante ressaltar que a escolha de q influencia diretamente no tamanho da árvore. Quanto menor, mais profunda ela pode ser. A escolha desse valor será feita de acordo com os interesses e com os dados, mas q tem que ser tal que $n_{lim} = qN > \mathbf{n}(g, d)$, caso contrário não seria possível determinar os coeficientes do polinômio $\mathbf{p}$.

Então, escolhido um valor para q de forma que $n_{lim} = qN > \mathbf{n}(g, d)$, o critério será: Se depois da criação dos conjuntos X_1 e X_2, $n_1 \leq n_{lim}$ ou $n_2 \leq n_{lim}$, os conjuntos X_1 e X_2 serão desconsiderados, esse nó não será dividido e passa a ser uma folha; caso contrário, a divisão será mantida.

3º Critério de Parada: Teste com as Médias de Y_1 e Y_2

Esse teste procura verificar se, após a divisão em Y_1 e Y_2, é verdade que os dois grupos de elementos possuem médias diferentes. Com isso, a idéia é evitar que um grupo de dados de uma mesma variável aleatória seja dividido em dois. Essa verificação será feita através de um simples teste de hipótese, que pode ser encontrado em diversos livros de estatística, como por exemplo, Hoel, Port e Stone (22).

Suponha Y_1 e Y_2 duas amostras independentes, com distribuição normal, médias desconhecidas μ_1 e μ_2, e variância também desconhecida, porém iguais, $\sigma_1^2 = \sigma_2^2 = \sigma$. Suponha também $\overline{Y}_1$, $\overline{Y}_2$, S_1^2 e S_2^2 as médias e variâncias amostrais de Y_1 e Y_2, respectivamente.

A hipótese nula que se deseja testar é:

$$H_0 : \mu_1 = \mu_2 \ .$$

Sob H_0, a estatística de teste t_0 definida a seguir é uma variável aleatória com distribuição t de *Student* com $n_1 + n_2 - 2$ graus de liberdade.

$$t_0 = \frac{\overline{Y}_1 - \overline{Y}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \qquad S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \tag{2-3}$$

Seja t_n é uma variável aleatória com distribuição t de *Student* e $t_{a,n} \in \mathbb{R}$ tal que $P(t_n \geq t_{a,n}) = a$. Além disso, o nível de significância α é definido por: $\alpha = P($ erro tipo I $) = P($ rejeitar $H_0 \mid H_0$ verdade $)$.

Então, escolhido um nível de significância α e calculado o valor de t_0 de acordo com a expressão acima, esse critério de parada é feito da seguinte maneira: se $-t_{\frac{\alpha}{2},n_1+n_2-2} < t_0 < t_{\frac{\alpha}{2},n_1+n_2-2}$, não se pode rejeitar H_0 e assim esse nó não é dividido e passa a ser uma folha; caso contrário, H_0 é rejeitada e a divisão mantida.