3
Árvores de Classificação e Regressão

3.1
Introdução

É crescente o uso de métodos estruturados por árvores de decisão como alternativas aos tradicionais modelos de classificação e regressão. As Árvores de Classificação são utilizadas quando a variável dependente é categórica, enquanto Árvores de Regressão tratam problemas em que esta é contínua.

A identificação de interação entre variáveis explicativas nas Ciências Sociais foi o problema que motivou o surgimento dos modelos estruturados por árvores. Morgan & Sonquist [73] desenvolveram o algoritmo AID (Automatic Interaction Identification) para identificação automática de interações, que posteriormente originaria o algoritmo CHAID (Chi-squared Automatic Detection), bastante utilizado no novo ramo da ciência que convencionou-se chamar de Aprendizado de Máquina.

Dentro da abordagem de Aprendizado de Máquina, o algoritmo ID3 (Inductive Dichotomizer 3) desenvolvido por Quinlan ([76]) é importante referência histórica na construção de árvores de decisão com a finalidade de classificação. Posteriormente, os algoritmos C4.5 e C5 ([78])aperfeiçoaram esta idéia inicial.

Entretanto, não há dúvidas de que o principal marco na utilização destes modelos foi a monografia CART [23] que unificou todos os desenvolvimentos, feitos até então, sobre este assunto. A partir deste trabalho, as árvores de decisão começaram a ganhar maior visibilidade como um procedimento estatístico.

Uma das razões para o sucesso dos modelos estruturados em árvores está na filosofia de utilizar modelos mais simples para subamostras dos dados, dividindo de forma conveniente o problema em partes. Por este motivo, as árvores de decisão também são conhecidas como métodos de particionamento recursivo, com larga aplicação nas Ciências da Saúde e Biologia (ver [100]).

As Árvores de Classificação e Regressão são utilizadas de forma mais frequente como métodos não-paramétricos aplicados à problemas de classificação e

regressão, pois não assumem a existência de modelos probabilísticos, não fazem suposições sobre componentes aleatórios e a forma funcional do modelo.

Estes métodos também encontram espaço, mais reduzido, na literatura de modelos não-lineares, dentro da classe dos modelos lineares por partes ou dos modelos aditivos descritos em Hastie & Tibshirani [46].

Algo de incomum nestes métodos, se comparado com a tradicional análise de regressão, é que o modelo ajustado é apresentado através de um gráfico em formato de uma árvore que cresce da raiz em direção as folhas, que também são chamadas de nós terminais. A raiz, também denominada de nó inicial, contém todas as observações no conjunto de dados e um teste lógico sobre o conjunto de variáveis explanatórias que só admite resposta no conjunto binário $\{$sim,não$\}$.

Após a aplicação do teste à cada observação, a raiz dará origem a dois novos nós que conterão parte das observações, em função de suas respostas ao teste. Se a resposta for ”sim”, por convenção, a observação é alocada dentro do nó esquerdo, caso contrário, dentro do nó direito. O mesmo procedimento é utilizado recursivamente nas observações dentro dos nós criados, através de novos testes lógicos. Deste modo, os nós geram dois filhos, cada filho gera mais dois e assim por diante, até atingir um ponto em que não há ganho em efetuar divisões, para melhorar a qualidade da predição, explicação do modelo ou a avaliação de uma especificada função perda. Quando o nó é estéril, ou seja, não gera dois novos nós, ele é classificado como folha ou nó terminal.

Outro importante elemento dentro da terminologia ilustrada na Figura 3.1 é a noção de profundidade, pois esta medirá quantos ancestrais determinado nó da árvore possui. A quantidade de ancestrais, em termos de modelagem estatística, dimensionará a complexidade da interação entre as variáveis explicativas presentes no modelo final.

Deste ponto em diante é adotada, de acordo com a Figura 3.2, uma forma padrão para enumerar os nós da árvore que será útil na parametrização dos modelos a serem apresentados nesta tese. À raiz, atribui-se o número 0 e depois na seqüência da esquerda para a direita, números inteiros de forma crescente.

É importante salientar que quando o nó não é gerado, salta-se o seu número e prossegue-se a enumeração a partir do próximo inteiro. Para exemplificar esta situação, os nós 4 e 5 não aparecem na Figura 3.2. Esta notação é semelhante à utilizada em [32] que identificam a posição dos nós gerados a partir da posição do nó gerador e, em conseqüência, é possível desenhar a árvore a partir da equação final do modelo.

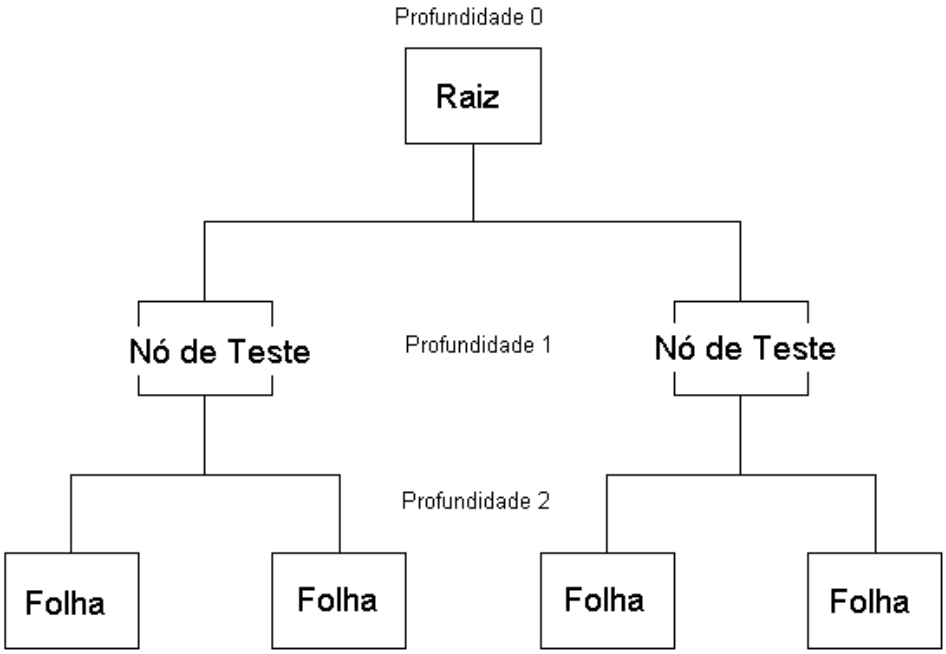

Figura 3.1: Terminologia de um Modelo Estruturado por Árvore

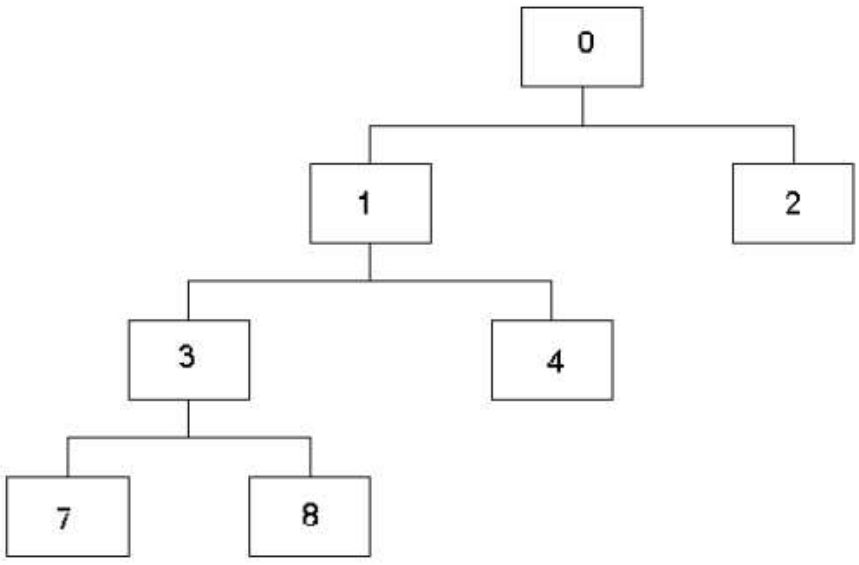

Figura 3.2: Enumeração dos Nós de uma Árvore

3.2 Formulação Matemática

Considere $\mathbf{x}_t = (x_{1t}, \ldots, x_{mt})' \in \mathbb{X} \subseteq \mathbb{R}^m$ um vetor que contém m variáveis explanatórias para uma resposta univariada contínua $y_t \in \mathbb{R}$.

A relação entre y_t e $\mathbf{x}_t$ segue o modelo de regressão:

$$y_t = f(\mathbf{x}_t) + \varepsilon_t, \tag{3-1}$$

onde a forma funcional $f(\cdot)$ é desconhecida e não há suposições sobre a distribuição do termo aleatório ε_t.

Seguindo [62], um modelo de árvore de regressão com K folhas é um modelo de particionamento recursivo que aproxima $f(\cdot)$ por uma função geral não-linear $H(\mathbf{x}_t; \boldsymbol{\psi})$ de $\mathbf{x}_t$ e definida pelo vetor de parâmetros $\boldsymbol{\psi} \in \mathbb{R}^r$ onde r é o número total de parâmetros.

Usualmente, $H(\cdot)$ é uma função constante por partes definida por K subregiões $k_i(\boldsymbol{\theta}_i)$, $i = 1, \ldots, K$, de algum domínio $\mathbb{K} \subset \mathbb{R}^m$.

Cada região é determinada pelo vetor de parâmetros $\boldsymbol{\theta}_i$, $i = 1, \ldots, K$, de forma que

$$f(\mathbf{x}_t) \approx H(\mathbf{x}_t; \boldsymbol{\psi}) = \sum_{i=1}^{K} \beta_i I_i(\mathbf{x}_t; \boldsymbol{\theta}_i), \tag{3-2}$$

onde

$$I_i(\mathbf{x}_t; \boldsymbol{\theta}_i) = \begin{cases} 1 & \text{se } \mathbf{x}_t \in k_i(\boldsymbol{\theta}_i); \\ 0 & \text{cc.} \end{cases} \tag{3-3}$$

Note que $\boldsymbol{\psi} = (\beta_1, \ldots, \beta_K, \boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_K)'$. Condicionalmente ao conhecimento das subregiões, a relação entre y_t e $\mathbf{x}_t$ em (3-1) é aproximada por uma regressão linear em um conjunto de K variáveis to tipo *dummy*.

A Figura 3.3 ilustra as características de um modelo gerado por uma árvore de regressão que explica a relação entre a variável resposta y e um conjunto de $m = 2$ variáveis explanatórias (preditoras) x_1 e x_2.

Os valores preditos para y são obtidos através de um cadeia de sentenças lógicas que dividem o conjunto de dados em quatro subconjuntos que particionam $\mathbb{R}^2$.

O conjunto de sentenças lógicas ou regras de predição associados ao modelo da Figura 3.3 é formado por:

Regra 1 Se $x_1 > 11$, então a melhor predição para y é 6.

Regra 2 Se $(x_1 < 11)$ & $(x_2 < 5.3)$, então a melhor predição para y é 1.8.

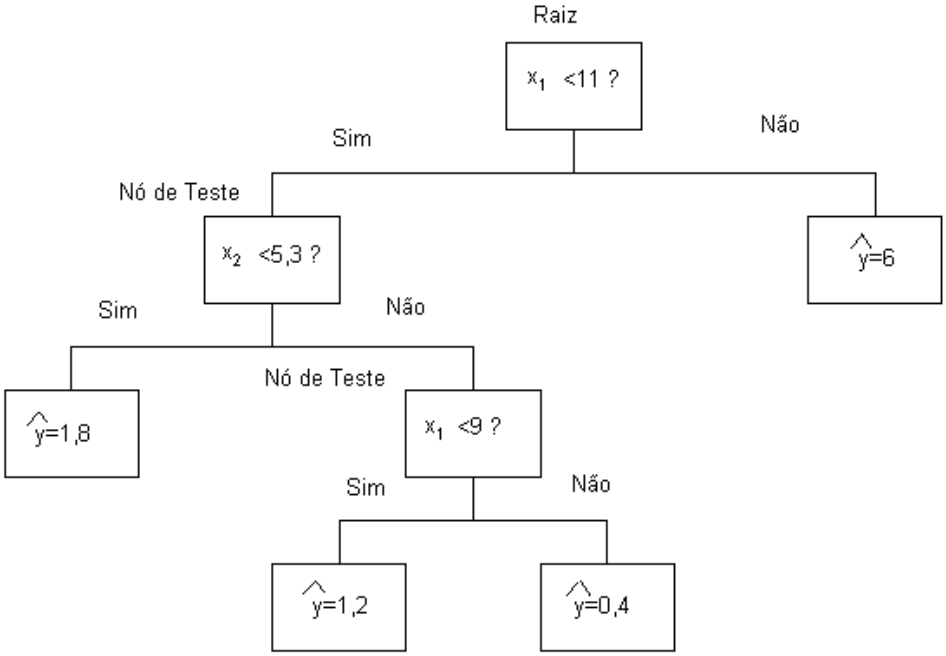

Figura 3.3: Saída Gráfica de uma Árvore de Regressão.

Regra 3 Se $(x_1 < 9)$ & $(x_2 > 5.3)$, então a melhor predição para y é 1.2.

Regra 4 Se $(9 < x_1 < 11)$ & $(x_2 > 5.3)$, então a melhor predição para y é 0.4.

A referência mais importante sobre árvores de regressão é abordagem CART discutida em [23]. Neste contexto, é usual definir subregiões k_i, $i = 1, \ldots, K$, em (3-2) através de hiperplanos que são ortogonais aos eixos das variáveis preditoras; veja, por exemplo, a Figura 3.3.

Considere a estrutura mais simples com $K = 2$ folhas e profundidade $d = 1$ como ilustrado na Figura 3.4.

A função desconhecida $f(\mathbf{x}_t)$ em (3-1) pode ser aproximada por um modelo constante em cada folha e, assim, escrita como

$$y_t = \beta_1 I(\mathbf{x}_t; s_0, c_0) + \beta_2 \left[1 - I(\mathbf{x}_t; s_0, c_0)\right] + \varepsilon_t, \tag{3-4}$$

onde

$$I(\mathbf{x}_t; s_0, c_0) = \begin{cases} 1 & \text{if } x_{s_0 t} \leq c_0; \\ 0 & \text{caso contrário,} \end{cases} \tag{3-5}$$

e $s_0 \in \mathbb{S} = \{1, 2, \ldots, m\}$.

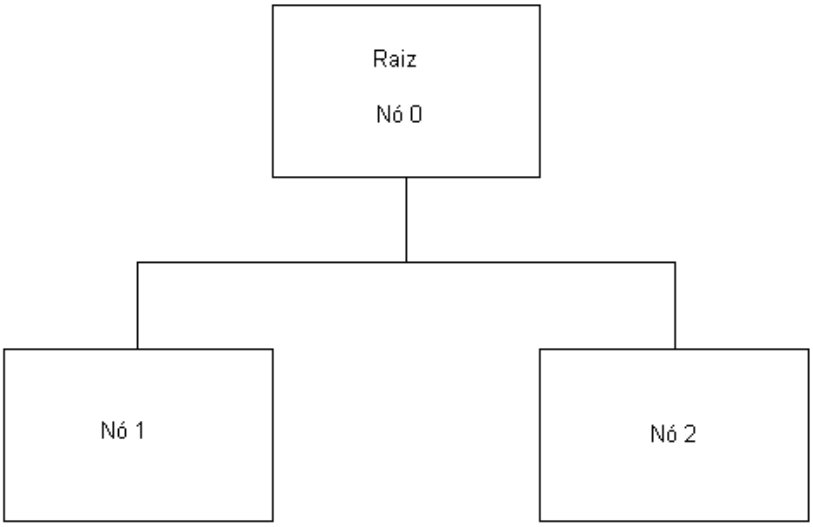

Figura 3.4: Arquitetura da Árvore de Regressão Mais Simples .

Considere agora uma árvore com N nós intermediários. As variáveis x_{s_j}, $j = 1, \ldots, N$ são usualmente chamadas *varíaveis de limiar ou divisão.*

Árvores mais complexas são mostradas nos seguintes exemplos.

Exemplo 3 *Considere uma árvore de regressão definida por*

$$\begin{aligned} y_t = &\{\beta_3 I(\mathbf{x}_t; s_1, c_1) + \beta_4 [1 - I(\mathbf{x}_t; s_1, c_1)]\} I(\mathbf{x}_t; s_0, c_0) + \\ &\beta_2 [1 - I(\mathbf{x}_t; s_0, c_0)] + \varepsilon_t. \end{aligned} \tag{3-6}$$

A representação gráfica de (3-6) é ilustrada na Figura 3.5. A árvore induzida por (3-6) possui 2 nós intermediários, 3 nós terminais, e profundidade igual a 2.

Exemplo 4 *Considere a seguinte árvore de regressão:*

$$\begin{aligned} y_t = &\{\beta_3 I(\mathbf{x}_t; s_1, c_1) + \beta_4 [1 - I(\mathbf{x}_t; s_1, c_1)]\} I(\mathbf{x}_t; s_0, c_0) + \\ &\{\beta_5 I(\mathbf{x}_t; s_2, c_2) + \beta_6 [1 - I(\mathbf{x}_t; s_2, c_2)]\} [1 - I(\mathbf{x}_t; s_0, c_0)] + \varepsilon_t. \end{aligned} \tag{3-7}$$

A representação gráfica de (3-7) é ilustrada na Figura 3.6. O modelo(3-7) tem 3 nós intermediários, 4 folhas , e profundidade igual a 2.

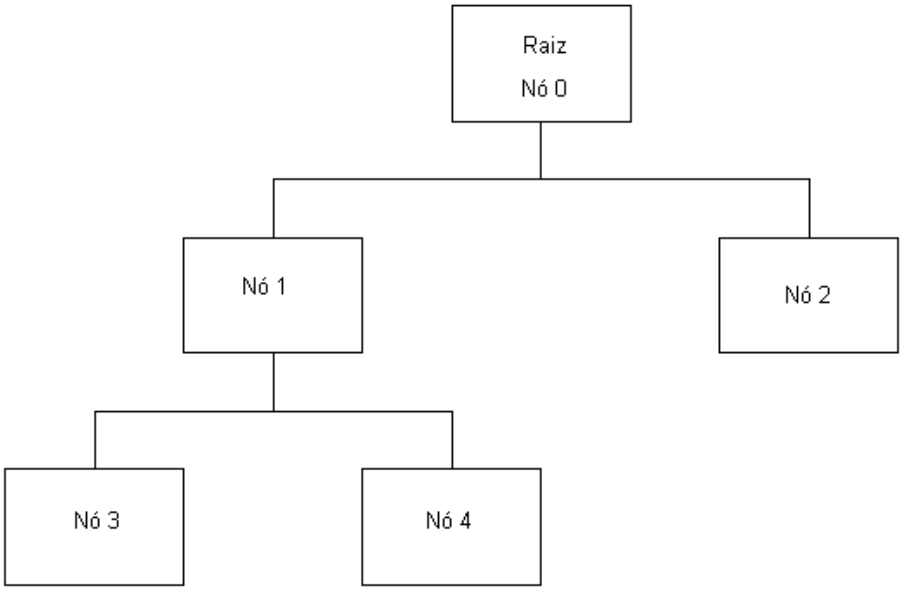

Figura 3.5: Árvore de Regressão com 3 Folhas Representando (3-6).

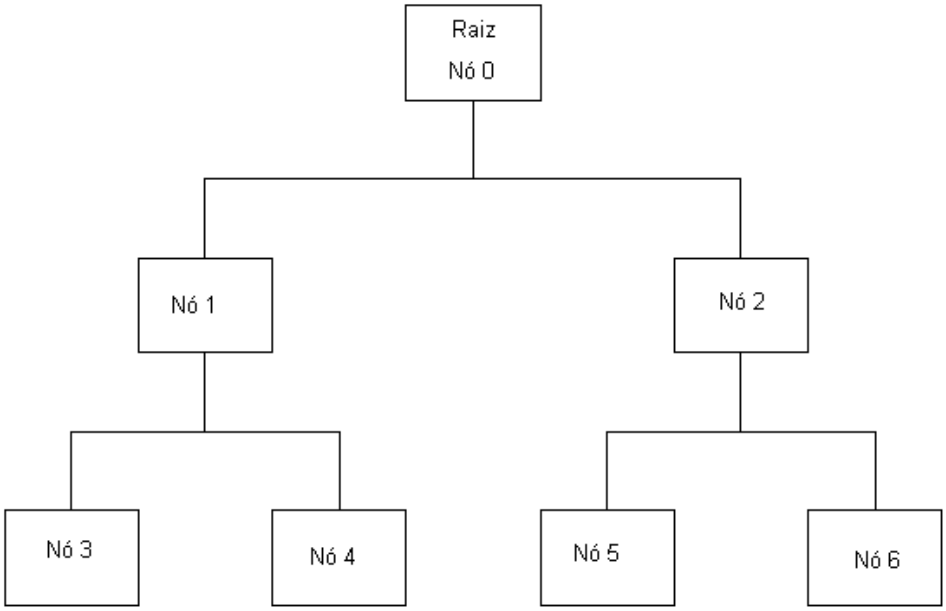

Figura 3.6: Árvore de Regressão com 4 Folhas Representando(3-7).

3.3 O Algoritmo CART

A referência mais importante sobre modelos estruturados por árvores é o livro seminal CART [23], que também empresta a denominação para um algoritmo e um software. Outros algoritmos tais como M5 [77] e RETIS [59] foram desenvolvidos com a mesma finalidade, entretanto a literatura sobre o CART, assim como o seu uso, ainda encontram-se mais difundidos.

O CART pode ser aplicado a problemas de classificação e regressão que envolvam variáveis explicativas de todos os tipos: nominal, ordinal, discreta e contínua. As árvores de classificação são utilizadas quando a variável resposta é nominal. No presente trabalho o interesse é direcionado para problemas de regressão, concentrando-se o texto subseqüente neste assunto.

O crescimento da árvore de regressão, segundo a metodologia CART, pode ser visto como um ciclo iterativo e recursivo que especifica a cada passo uma ”arquitetura”para a árvore e estima os parâmetros. A cada iteração devem ser especificados: um nó para ser dividido (particionado), uma variável de divisão e o limiar desta divisão. Após a especificação, são estimados os parâmetros dos modelos locais para as observações alocadas dentro dos nós gerados pela divisão. Este procedimento é repetido recursivamente até que atinja-se um ponto a partir do qual não haja ganho em efetuar subdivisões na árvore.

O modelo final é avaliado por medidas de custo-complexidade ou por sua capacidade preditiva, podendo ser re-especificado através do corte de alguns ramos da árvore, uma técnica que é chamada de podagem (prunning). O ciclo começa com a especificação do modelo mais simples (Figura 3.4), através da seleção de uma variável de divisão x_{s_1} e seu correspondente limiar c_1. Posteriormente é feita a estimação dos modelos localmente constantes representados pelos parâmetros β’s . Comumente, a seleção e estimação são realizadas simultaneamente pela busca exaustiva do par (x_{s_0}, c_0) que minimiza a soma dos erros quadráticos:

$$SQ^{Arv_1} = \sum_{i=1}^{T} \{y_t - \beta_1 I(\mathbf{x}_t; s_0, c_0) - \beta_2 [1 - I(\mathbf{x}_t; s_0, c_0)]\}^2 \qquad (3\text{-}8)$$

ou, alternativamente, a soma dos desvios absolutos,

$$SDA^{Arv_1} = \sum_{i=1}^{T} |y_t - \beta_1 I(\mathbf{x}_t; s_0, c_0) - \beta_2 [1 - I(\mathbf{x}_t; s_0, c_0)]| \qquad (3\text{-}9)$$

O índice Arv_1 caracteriza a árvore de regressão mais simples. A árvore obtida através da minimização de (3-8) é chamada árvore de regressão MQ (Mínimos Quadrados). Condicionalmente ao par (x_{s_0}, c_0), é direta a verificação que as médias amostrais da variável resposta dentro dos nós gerados são os melhores estimadores para β_1 e β_2 , pois minimizam (3-8).

$$\widehat{\beta_1}^{MQ} = \frac{\sum_{t=1}^{T} y_t I(\mathbf{x}_t; s_0, c_0)}{\sum_{i=1}^{T} I(\mathbf{x}_t; s_0, c_0)} \tag{3-10}$$

$$\widehat{\beta_2}^{MQ} = \frac{\sum_{t=1}^{T} y_t (1 - I(\mathbf{x}_t; s_0, c_0))}{\sum_{i=1}^{T} (1 - I(\mathbf{x}_t; s_0, c_0))} \tag{3-11}$$

A árvore obtida pela minimização da soma dos desvios absolutos é chamada de árvore de regressão MDA (Mínimos Desvios Absolutos). Pode ser mostrado que a mediana amostral da variável resposta dentro do nó gerado é o melhor estimador para este critério. Mais sobre regressão através da minimização dos desvios absolutos pode ser encontrado em [74] e [16].

Após a divisão da raiz (nó 0) que implica a estimação do vetor $(s_0, c_0, \beta_1, \beta_2)$, se o critério de quadrados mínimos for utilizado, o modelo pode ser re-especificado pela seleção da arquitetura mostrada na Figura 3.5 o que implica na seleção do par (s_1, c_1) que minimize:

$$\begin{aligned} SQ^{Arv_2} = \sum_{i=1}^{T} \{ & y_t - \beta_2 [1 - I(\mathbf{x}_t; s_0, c_0)] - \\ & (\beta_3 I(\mathbf{x}_t; s_1, c_1) + \beta_4 [1 - I(\mathbf{x}_t; s_1, c_1)])(I(\mathbf{x}_t; s_0, c_0))\}^2. \end{aligned} \tag{3-12}$$

No algoritmo CART, a minimização de SQ^{Arv_2} é equivalente a maximizar:

$$R(Arv_2) = SQ(Arv_1) - SQ(Arv_2) \tag{3-13}$$

onde o argumento Arv_2 é uma referência à primeira árvore especificada após à mais simples, Arv_1. Esta árvore naturalmente possuirá 3 nós terminais e pode ser especificada de duas formas; divindo o nó 2 conforme a Figura 3.5 ou então o nó 3. O processo especificação-estimação continuará maximizando a diminuição na soma total dos erros quadráticos, isto é, maximizando $R(Arv_3)$, depois $R(Arv_4)$ e assim por diante.

O procedimento descrito acima conduz o modelo a um superajuste dos dados pois força, através de seguidas partições, a diminuição da soma dos erros quadráticos. Assim, é necessário estabelecer um critério de parada que possa verificar se um nó gerado será dividido recursivamente ou declarado como termi-

nal. A sugestão apresentada em [23] é de que um nó que contenha 5 observações ou menos seja declarado como terminal. Para reduzir a complexidade da árvore, uma última verificação pode ser feita por uma técnica de podagem.

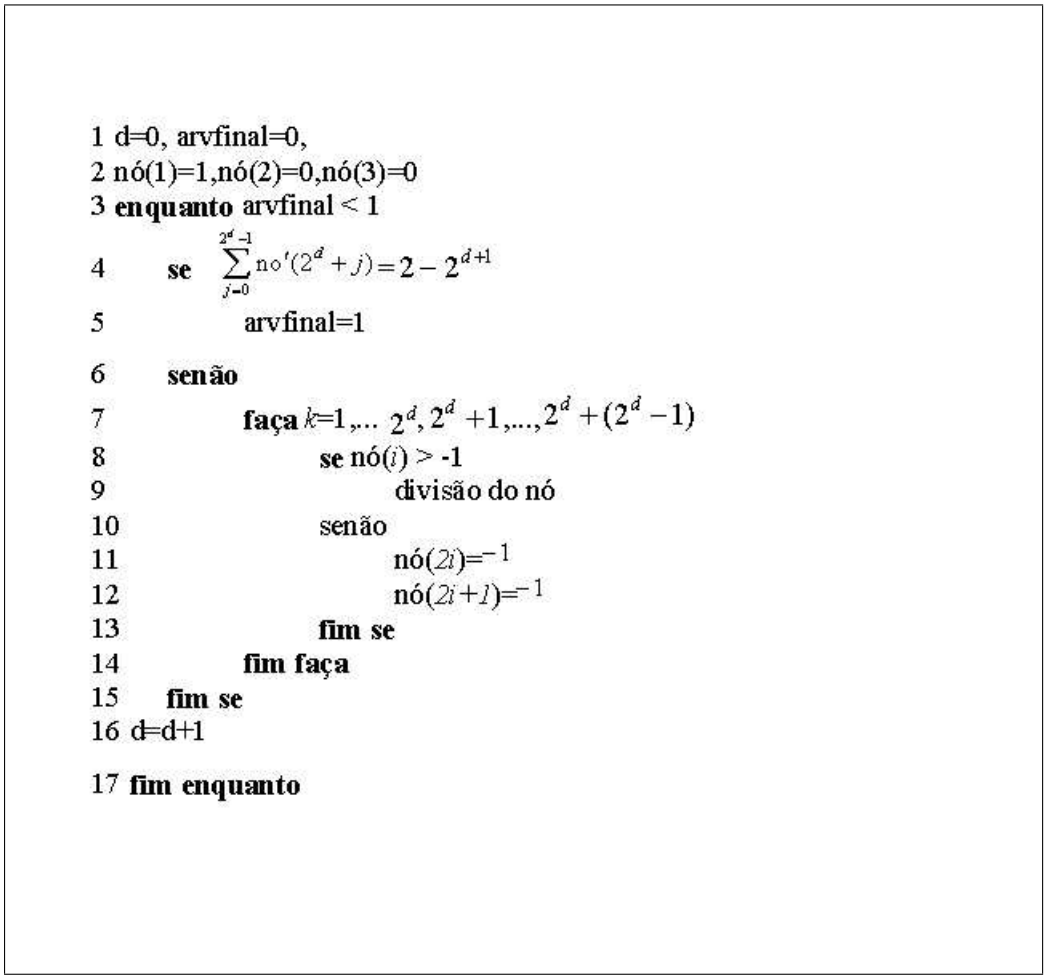

Figura 3.7: Algoritmo de Crescimento da Árvore

As Figuras 3.7 e 3.8 contém o que seria esboço mais simples do algoritmo computacional para construção do modelo CART. Neste algoritmo há uma parte principal que é responsável por controlar os estados dos nós presentes na árvore.

Basicamente, o nó pode assumir 3 estados: nó de teste (0), nó terminal (1) e nó inexistente (-1). O último destes estados é apenas colocado como um artifício para que a proposta de numeração estabelecida na Figura 3.2 possa ser seguida.

Na Figura 3.8 há a parte principal do algoritmo que é responsável pela decisão sobre a divisão de um nó. Dois importantes passos para a decisão são: a escolha da função perda ($L(.)$) e a regra de parada, que na Figura 3.8 é configurada pelo número mínimo de 5 observações dentro do nó terminal.

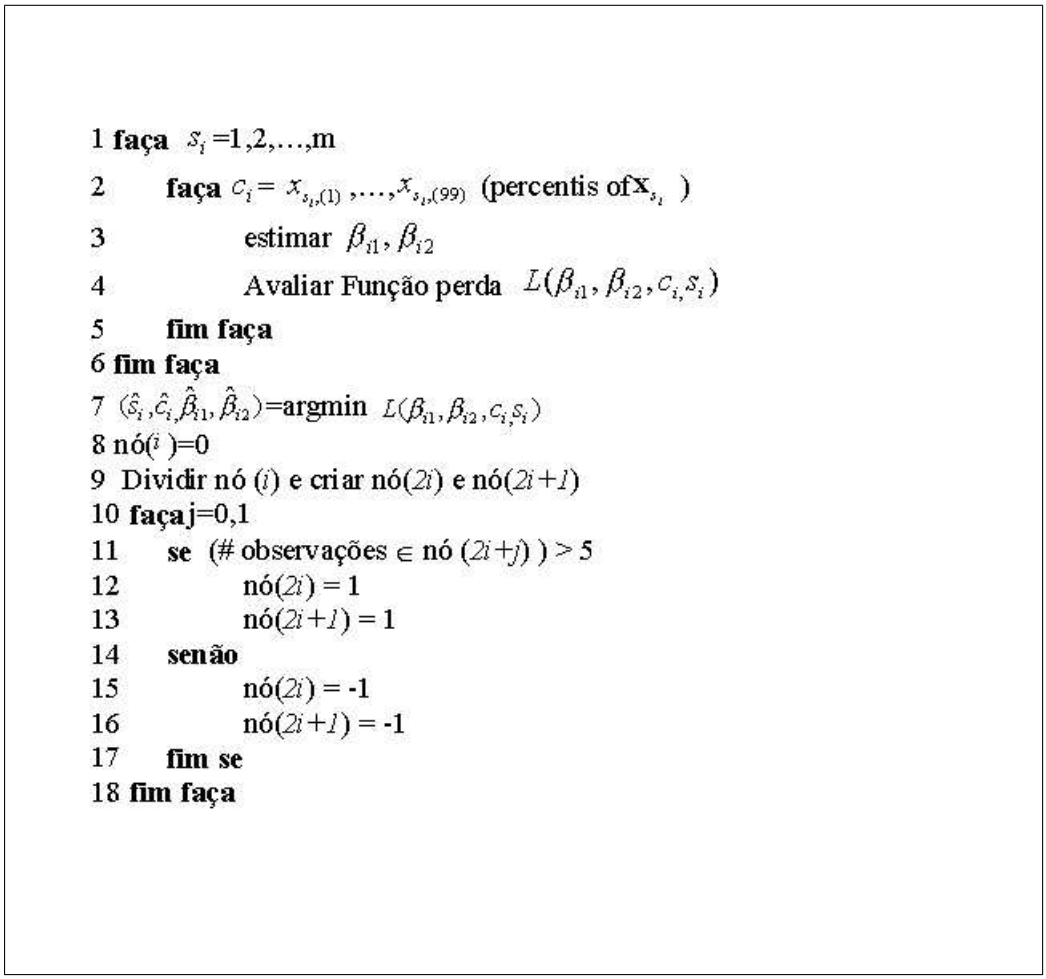

Figura 3.8: Algoritmo de Divisão do Nó

3.4 Podagem

A podagem pode ser considerada como uma forma de diminuir a complexidade das árvores trazendo vantagens em termos de compreensão e interpretação de modelos e também um modo de melhorar a acurácia preditiva em muitos casos [90]. Dentro da metodologia CART, a metodologia estabelecida é a podagem por custo-complexidade. Suponha uma medida de qualidade de ajuste R_i calculada dentro da i-ésima folha e que árvore possua N folhas . Em [23] há a sugestão para avaliar o custo complexidade através de uma função:

$$R^*(N, \alpha) = \sum_{i=1}^{N} R_i + |\alpha|N \tag{3-14}$$

onde α é um parâmetro que penaliza a árvore pelo seu tamanho.

Outra forma clássica de realizar a podagem é utilizar um experimento de validação cruzada. Esta técnica consiste em particionar o conjunto de dados em

dois subconjuntos, utilizando um destes, o chamado conjunto de treinamento, para ajustar a árvore e o restante (conjunto de teste) para avaliar a sua capacidade preditiva. A forma clássica de conduzir este experimento segue os seguintes passos:

1. Particionar o conjunto em 10 partes de tamanhos aproximadamente iguais;
2. Selecionar uma das parte para conjunto de teste e utilizar 9/10 restantes como conjunto de treinamento.
3. Selecionar, sem reposição, outra parte e repetir o procedimento do passo anterior

A execução dos passos acima é chamada de validação cruzada com 10-dobras e, inclusive, é utilizada para obter estimativas de medidas de custo complexidade tais como a citada em (3-14).

3.5 Outros Desenvolvimentos

O uso de estruturas de árvore para modelagem tem crescido consideravelmente desde a unificação destes métodos em [23].

Ciampi [27] apresenta uma proposta, dentro do contexto dos modelos lineares generalizados (GLM), para a construção de árvores de regressão . O autor utiliza a razão entre *deviances* para decisões sobre a formação dos nós. A conjunção de métodos de análise de dados longitudinais com estruturação por árvores é feita em [82]. Modelos nas folhas da regressão no contexto de análise de sobrevivência são utilizados em [1] e os resultados identificam efeitos locais (nas folhas) das covariáveis sobre o tempo de sobrevivência de pacientes que realizaram transplante de coração. Também, sobre análise de sobrevivência, pode ser citado o trabalho de [29].

Em [90], há a proposta de utilizar modelos de regressão local com o objetivo de tornar o modelo mais suave e diminuir a variância dos estimadores. Este procedimento consiste em ajustar um modelo de regressão linear local dentro das folhas. Três alternativas são sugeridas: utilizar o modelo de regressão local para induzir o crescimento da árvore, construir a árvore através de procedimentos padrões e utilizar o modelo local durante a podagem, ou construir a árvore e realizar a podagem com procedimentos usuais e utlilizar a regressão local somente para predição.

Na análise de séries temporais, uma aplicação da metodologia de árvores de regressão em modelos lineares por partes foi proposta por [50]. O desenvolvimento foi feito para séries univariadas e o limiar, ao contrário da proposta de Tong que utiliza um limiar univariado, é construído através da combinação de variáveis presentes no vetor de estado do modelo, que desempenham o papel de variáveis independentes na análise de regressão. Os autores atribuem a este modelo a denominação PCAR (Piecewise Constant AutoRegressive). [28] propõe o uso de uma árvore de regressão inspirada no algoritmo CART para capturar diferentes níveis na evolução de uma série temporal associada com o ciclo de negócios nos EUA. Neste desenvolvimento, o conjunto de possíveis variáveis de divisão é formado pelo índice de tempo e defasagens da série analisada. Loh propõe, também no contexto de séries temporais, uma metodologia que conjuga os modelos TAR com divisões binárias feitas por uma árvore. A partição recursiva do espaço das variáveis pode ser feita pela aplicação do teste t aos resíduos ou por busca exaustiva. Nestes modelos, medidas de custo complexidade são utilizadas para podagem a posteriori conforme [97].

Seleção Bayesiana da arquitetura da árvore pode ser econtrada em [32]. Os autores utilizam um algoritmo do tipo RJMCMC (Reversible Jump Monte Carlo Markov Chain) descrito em [40], para selecionar a arquitetura da árvore. Neste artigo, o enfoque principal recai sobre o problema de classificação, muito embora os autores apontem o caminho para implementação em problemas de regressão.