3.
Árvore de Decisão

A árvore de decisão consiste de uma hierarquia de nós internos e externos que são conectados por ramos. O nó interno, também conhecido como nó decisório ou nó intermediário, é a unidade de tomada de decisão que avalia através de teste lógico qual será o próximo nó descendente ou filho. Em contraste, um nó externo (não tem nó descendente), também conhecido como folha ou nó terminal, está associado a um rótulo ou a um valor.

Em geral, o procedimento de uma árvore de decisão é o seguinte: apresenta-se um conjunto de dados ao nó inicial (ou nó raiz que também é um nó interno) da árvore; dependendo do resultado do teste lógico usado pelo nó, a árvore ramifica-se para um dos nós filhos e este procedimento é repetido até que um nó terminal é alcançado. A repetição deste procedimento caracteriza a recursividade da árvore de decisão.

No caso das árvores de decisão binária, cada nó intermediário divide-se exatamente em dois nós descendentes: o nó esquerdo e o nó direito. Quando os dados satisfazem o teste lógico do nó intermediário seguem para o nó esquerdo e quando não satisfazem seguem para o nó direito. Logo, uma decisão é sempre interpretada como verdadeira ou falsa. Deve ser mencionado que, restringimos a nossa descrição de divisão para árvores binárias, pois estas serão empregadas nesta tese. Contudo, na literatura há árvore de decisão com várias divisões e sua descrição pode ser encontrada em Zighed (2000).

A Figura 1 é uma representação gráfica de uma árvore de decisão binária.

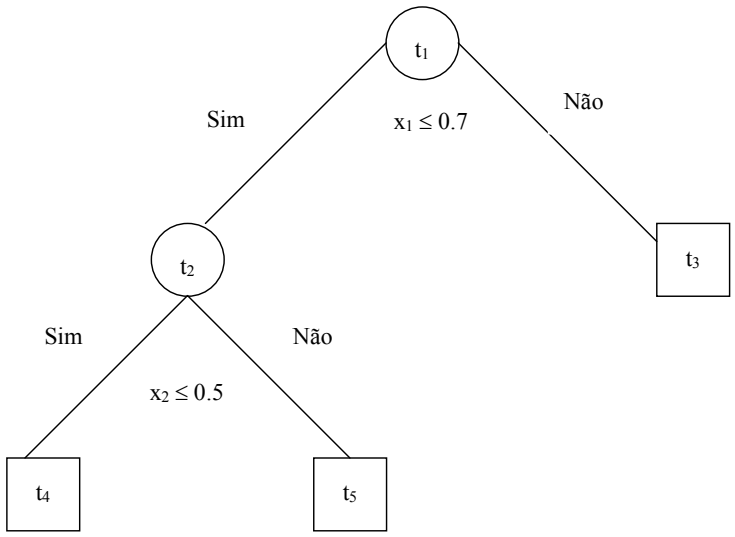

Figura 1 - Árvore de Decisão Binária

Na figura anterior, os círculos representam os nós internos (intermediários ou decisórios); os quadrados representam os nós folhas ou terminais; as linhas representam os ramos que interligam dois nós; e x_1 e x_2 representam as variáveis decisórias. Chama-se de variável decisória a variável de entrada que levará a uma nova divisão da árvore de decisão, em relação a um possível valor.

A interpretação da representação gráfica da árvore de decisão ilustrada anteriormente é descrita da seguinte forma:

- Quando a condição é satisfeita (por exemplo, $x_1 \leq 0.7$), os dados seguem para o nó esquerdo (SIM) e, caso contrário, os dados seguem para o nó direito (NÃO);
- Numeração dos nós:

  $t_1 \Rightarrow$ nó 1, $t_2 \Rightarrow$ nó 2, $t_3 \Rightarrow$ nó 3, $t_4 \Rightarrow$ nó 4 e $t_5 \Rightarrow$ nó 5
- Níveis da árvore de decisão:

  Nível 0 : nó 1

  Nível 1 : nós 2 e 3

  Nível 2 : nós 4 e 5

O aprendizado de uma árvore de decisão é supervisionado, ou seja, o método aproxima funções-alvo de valor discreto, na qual a função aprendida é representada por uma árvore de decisão. As árvores treinadas podem ser representadas como um conjunto de regras "Se-Então" para melhoria da

compreensão e interpretação.

As árvores de decisão são estudadas em vários campos de pesquisa como ciências sociais, estatística, engenharia e inteligência artificial. Atualmente, elas têm sido aplicadas, com sucesso, em um enorme campo de tarefas desde diagnóstico de casos médicos até avaliação de risco de crédito de requerentes de empréstimo.

Árvores de decisão usadas para problemas de classificação são chamadas de Árvores de Classificação. Em algumas referências bibliográficas (Torgo, 1997), a árvore de classificação pode ser denominada, simplesmente, como árvore de decisão.

Nas árvores de classificação, cada nó terminal ou folha contém um rótulo que indica a classe predita para um determinado conjunto de dados. Neste tipo de árvore pode existir dois ou mais nós terminais com a mesma classe.

Para ilustrar uma árvore de classificação, encontra-se na Figura 2 a representação gráfica deste tipo de árvore para duas classes.

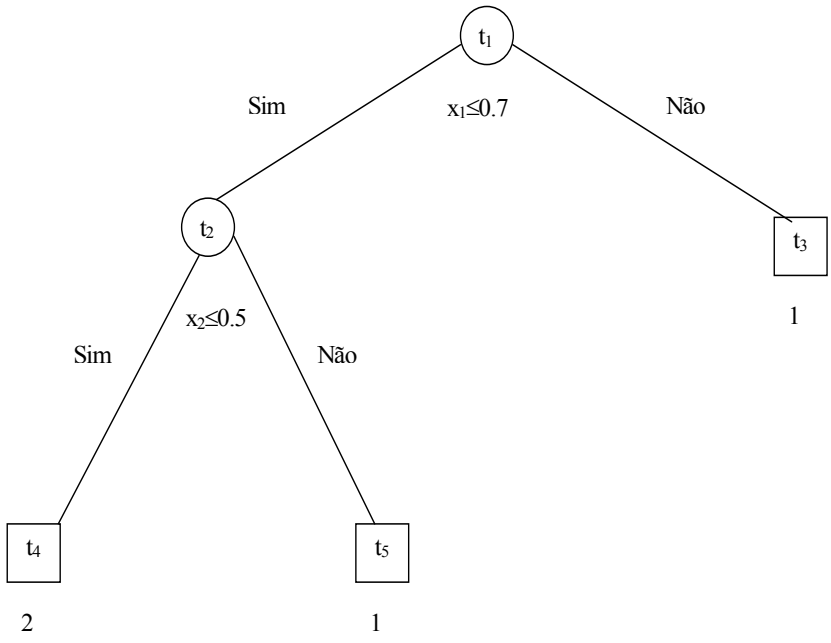

Figura 2 - Árvore de Classificação

Na árvore de classificação ilustrada na figura anterior as classes formadas são Classe 1, representada pelos nós 3 e 5 , e a Classe 2, representada pelo nó 4. As regras obtidas após a árvore treinada são:

Regra para Classe 1 Se ($x_1 > 0.7$) ou Se ($x_1 \leq 0.7$ e $x_2 > 0.5$)

Regra para Classe 2 Se ($x_1 \leq 0.7$ e $x_2 \leq 0.5$)

Árvores de decisão usadas para problemas de regressão são chamadas de Árvores de Regressão. Nas árvores de regressão, cada nó terminal ou folha contém uma constante (geralmente, uma média) ou uma equação para o valor previsto de um determinado conjunto de dados.

Empregando a mesma representação gráfica da árvore de classificação (Figura 2), temos para cada nó terminal um modelo linear.

$$Y = \begin{cases} \sum \beta_i^{(3)} x_i + \varepsilon^{(3)} & \text{se } x \in t_3 \\ \sum \beta_i^{(4)} x_i + \varepsilon^{(4)} & \text{se } x \in t_4 \\ \sum \beta_i^{(5)} x_i + \varepsilon^{(5)} & \text{se } x \in t_5 \end{cases} \quad (1)$$

onde:

$\beta_i^{(k)}$: i-ésimo parâmetro β do modelo linear do k-ésimo nó;

$\varepsilon^{(k)}$: ruído do modelo linear do k-ésimo nó;

x : dados de entrada

Y : dados de saída

Existem dois aspectos que merecem destaques em uma árvore de decisão, o crescimento e a poda, que serão abordados na seção 3.1.

Por fim, um dos mais conhecidos e mais completos algoritmos de árvore de decisão é o CART - "Classification and Regression Tree" - que foi proposto por Breiman (1984). Como este algoritmo será empregado em uma das etapas da modelagem proposta nesta tese, é conveniente realizar uma breve descrição do CART na seção 3.2.

3.1. Crescimento e Poda

As árvores de decisão são construídas usando um algoritmo de partição recursiva. Este algoritmo constrói uma árvore por divisões recursivas binárias que começa no nó raiz e desce até os nós folhas. Têm-se dois fatores principais no algoritmo de partição: a forma para selecionar uma divisão para cada nó intermediário (Crescimento) e uma regra para determinar quando um nó é terminal (Poda).

O problema chave, no algoritmo de partição recursiva, é a confiabilidade das estimativas do erro usado para selecionar as divisões. As escolhas da divisão em níveis maiores da árvore produzem, freqüentemente, estatísticas não-confiáveis apesar da estimativa do "erro de resubstituição" (estimativa obtida com os dados de treinamento usado durante o crescimento da árvore) manter-se decrescendo. Com isto, a precisão das estimativas do erro é fortemente dependente da qualidade da amostra. Como o algoritmo divide recursivamente o conjunto de dados de treinamento original, as divisões estão sendo avaliadas com amostras cada vez menores. Isto significa que as estimativas de erro têm menos confiabilidade à medida que crescemos a árvore. Com intuito de minimizar este problema e evitar o superajustamento dos dados de treinamento com árvores muito complexas, tem-se a estratégia conhecida como método de podagem.

Há dois procedimentos alternativos para podagem da árvore de decisão: a pós-podagem e a pré-podagem.

A pós-podagem é o processo pelo qual uma árvore é crescida ao tamanho máximo e então métodos de evolução confiáveis são usados para selecionar a árvore podada de tamanho certo desde o modelo inicial. Este algoritmo considera a podagem como um processo de "dois-estágios". No primeiro estágio, um conjunto de árvores podadas de Tmax (árvore de tamanho máximo) é gerado de acordo com algum critério, enquanto no segundo estágio uma dessas árvores é selecionada como o modelo final.

Os métodos de pós-podagem podem ser computacionalmente ineficientes, no sentido que não é usual achar domínios onde uma árvore extremamente grande (por exemplo, com milhares de nós) é pós-podada em poucas centenas de nós - isto parece um desperdício computacional. Uma alternativa de parada no

procedimento de crescimento da árvore é interromper o crescimento tão logo a divisão seja considerada não-confiável. Isto é conhecido como a pré-podagem da árvore.

O método de pré-podagem usa um procedimento "passo único". Este algoritmo corre através dos nós da árvore ou "de baixo para cima" ou "de cima para baixo", decidindo para cada nó, se é para podar de acordo com algum critério de avaliação.

Os métodos de pré-podagem também apresentam um ponto negativo no seu algoritmo. A pré-podagem corre o risco de selecionar uma árvore subótima ao interromper o crescimento da árvore (Breiman, 1984).

Breiman (1984) descreveu duas alternativas para a seleção da árvore final baseada nas estimativas dos erros obtidos. Ou seleciona a árvore com menor erro estimado ou escolhe a menor árvore na seqüência, cujo erro estimado está dentro do intervalo: $Err_b + SE(Err_b)$, onde Err_b é o menor erro estimado e $SE(Err_b)$ é o erro padrão desta estimativa. Mas tarde, este método será conhecido como a regra "1-SE". Para maiores detalhes sobre essas alternativas consultar Breiman (1984) ou Zighed (2000).

Destaca-se que para árvores de classificação a podagem é em função da complexidade do custo mínimo (erro de resubstituição) e para árvores de regressão, a podagem é em função da complexidade do erro mínimo.

3.2. CART

A metodologia do modelo CART (Breiman, 1984) é tecnicamente conhecida como partição recursiva binária. O processo é binário porque os nós pais são sempre divididos exatamente em dois nós filhos e recursivamente porque o processo pode ser repetido tratando cada nó filho como um nó pai. As principais características do CART são: definir o conjunto de regras para dividir cada nó da árvore; decidir quando a árvore está completa; associar cada nó terminal a uma classe ou a um valor preditivo no caso da regressão.

Para dividir um nó em dois nós filhos, o algoritmo sempre faz perguntas que tem apenas um "sim" ou um "não" como resposta. Por exemplo, as questões podem ser: a idade é <=55? ou o crédito é <=600?

O próximo passo é ordenar cada regra de divisão com base no critério de qualidade de divisão. O critério padrão usado para *classificação* é o Índice de Gini que tem por base o cálculo da entropia (Zighed, 2000 e Lamas, 2000)

$$\phi(p_1,...,p_2) = -\sum_j p_j \log p_j \quad (2)$$

onde p é a frequência encontrada de cada classe j, e o processo de divisão da árvore de *regressão* procura minimizar R(T).

$$R(T) = \frac{1}{N} \sum_{t \in T} \sum_{x \in t} (y - \bar{y}(t))^2 \quad (3)$$

sendo t o identificador de cada nó da árvore e R(T) o valor esperado da soma dos erros quadráticos da regressão utilizando uma constante como modelo preditivo (a média). Como pode-se notar na equação 3, o CART não apresenta na árvore de regressão, um modelo linear em seus nós terminais e sim uma média.

Uma vez encontrada a melhor divisão, repete-se o processo de procura para cada nó filho, continuamente até que a divisão seja impossível ou interrompida.

No procedimento do CART, ao invés de determinar quando um nó é terminal ou não, continua-se proporcionando o crescimento da árvore até que não seja mais possível fazê-lo, como por exemplo ao atingir um número mínimo de dados na amostra. Depois que todos os nós terminais foram encontrados, é definida a árvore como maximal, ou seja, a árvore de tamanho máximo.

Após encontrar a árvore maximal, começa-se a podar alguns ramos da mesma árvore de modo a aumentar o poder de generalização. Algumas sub-árvores, obtidas através da poda de alguns ramos desta árvore, são examinadas testando taxas de erros e a melhor delas é escolhida.