Uma introdução aos modelos uni e multivariados de classificação e regressão por árvores

Cesar Augusto Taconeli
Departamento de Estatística - UFPR

Sumário

1. Árvores de Classificação e Regressão - CART)

- Principal referência: Breiman et al (1984);
- Modelagem não paramétrica;
- Execução de sucessivas partições binárias de uma amostra, buscando a constituição de sub-amostras menos heterogêneas.
- Variável dependente:
  - Numérica – Árvore de Regressão
  - Categórica – Árvore de Classificação

1. Árvores de Classificação e Regressão - CART)

- Alternativa ou complemento a procedimentos estatísticos de classificação e regressão como:
  - Regressão linear múltipla;
  - Regressão logística;
  - Análise de sobrevivência;
  - Análise discriminante;
  - Análise de agrupamentos, dentre outros.

1.1 Atrativos

- Procedimento de simples aplicação;
- Possibilidade de modelar dados com estruturas complexas:
  - Dados desbalanceados;
  - Dados faltantes;
  - Grande número de variáveis independentes.
- Detecção de interações de ordens elevadas;
- Ausência de pressuposições paramétricas;
- Produção de resultados facilmente interpretáveis.

1.2 Representação

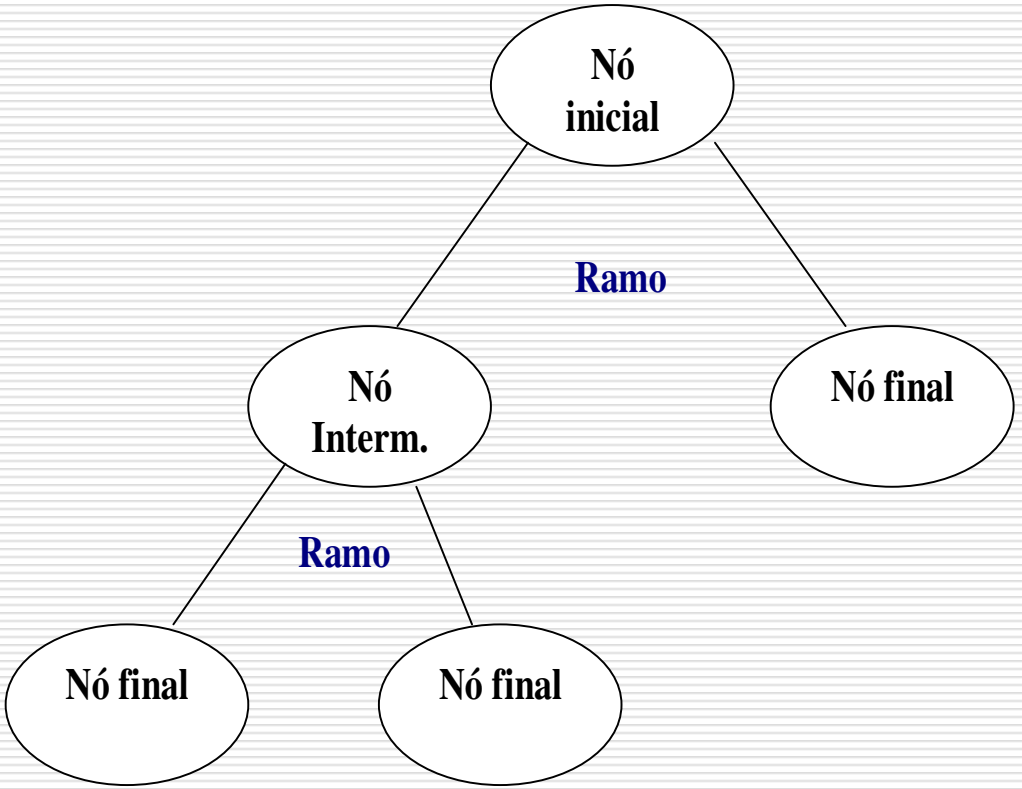

Figura 1 – Ilustração de uma árvore de regressão/classificação

1.3 Construção das árvores

- Definição e execução de um critério de partição;
- Poda;
- Seleção do modelo;
- Caracterização dos nós finais.

1.3.1 Definição e execução de um critério de partição

As partições devem ser realizadas com base nos resultados das co-variáveis.

- Seja $\{Y_j, \mathbf{x}_j\}$, $j = 1, 2, \ldots, n$ observações de uma variável dependente Y e de um vetor p-dimensional de variáveis independentes $\mathbf{x}$. Deve-se partir a amostra original em duas, agrupando observações de acordo com respostas a questões do tipo:

1.3.1 Definição e execução de um critério de partição

- Para covariáveis numéricas: "$X_{ij} \leq \tau$?"
  - X_{ij} : valor da $i-ésima$ variável no elemento j ;
  - τ : qualquer valor amostrado $i-ésima$ variável.
- Para covariáveis categorizadas: "$x_{ij} \in A$?"
  - A : qualquer categoria (ou subconjunto de categorias) de X_i .

1.3.1 Definição e execução de um critério de partição

- **Questão:** Qual das possíveis partições deve ser executada?
  - Aquela que melhor explicar a variação da resposta, constituindo sub-amostras pouco heterogêneas.
  - Quantifica-se a heterogeneidade das sub-amostras constituídas por meio de alguma ***medida de impureza***.

1.3.1 Definição e execução de um critério de partição

- Medidas de impureza
  - Para árvores de classificação: índice de entropia.

Considere um nó t qualquer. Dispõe-se, por exemplo, da seguinte medida de impureza:

$$\phi(t) = -\sum_{k} p\left(k \mid t\right) \log\left(p\left(k \mid t\right)\right)$$

$p\left(k \mid t\right)$: proporção de observações pertencentes ao nó *t* e à classe *k*.

1.3.1 Definição e execução de um critério de partição

- Medidas de impureza
  - Para árvores de regressão: índice ANOVA.

$$\phi(t) = \sum_i \{y(j \mid t) - \bar{y}(t)\}^2$$

$y(j \mid t)$: observação j em t;

$\bar{y}(t)$: média das observações no nó t.

1.3.1 Definição e execução de um critério de partição

- Variação da impureza

Considere um nó t dividido em dois novos nós, t_L e t_R baseado em uma partição s . A redução da impureza produzida pela partição é calculada como:

$$\Delta_\phi(s,t) = \phi(t) - \frac{n_L}{n}\phi(t_L) - \frac{n_R}{n}\phi(t_R)$$

- Executa-se s que maximiza $\Delta_\phi(s,t)$.
- Procede-se igualmente em relação às sub-amostras até a constituição de uma árvore com reduzido número de observações em cada nó final

1.3.2 Poda

- Objetivo: Eliminar da árvore partições que pouco contribuem para a explicação da variável resposta.

- Método: Baseado nos valores de uma função de custo-complexidade:

1.3.2 Poda

- Baseada na seguinte função de custo-complexidade:

$$R_\alpha(T) = R(T) + \alpha\left|\tilde{T}\right|$$

$R(T) = \sum_{t \in \tilde{T}} \phi(t)$: custo associado à taxa de má-classificação da árvore;

$\left|\tilde{T}\right|$: número de nós finais da árvore;

$\alpha \geq 0$: parâmetro de complexidade.

- Aumentando o valor de α a partir de zero obtém-se uma seqüência aninhada de árvores de tamanho decrescente, cada uma ótima para seu tamanho.

1.3.3 Seleção do modelo

- Construção de um gráfico de custo-complexidade, representando as árvores da seqüência aninhada com custos estimados por validação cruzada;

- Seleção da árvore pela regra do desvio padrão (*1-se rule* – Breiman et al, 1984).

1.3.3 Seleção do modelo

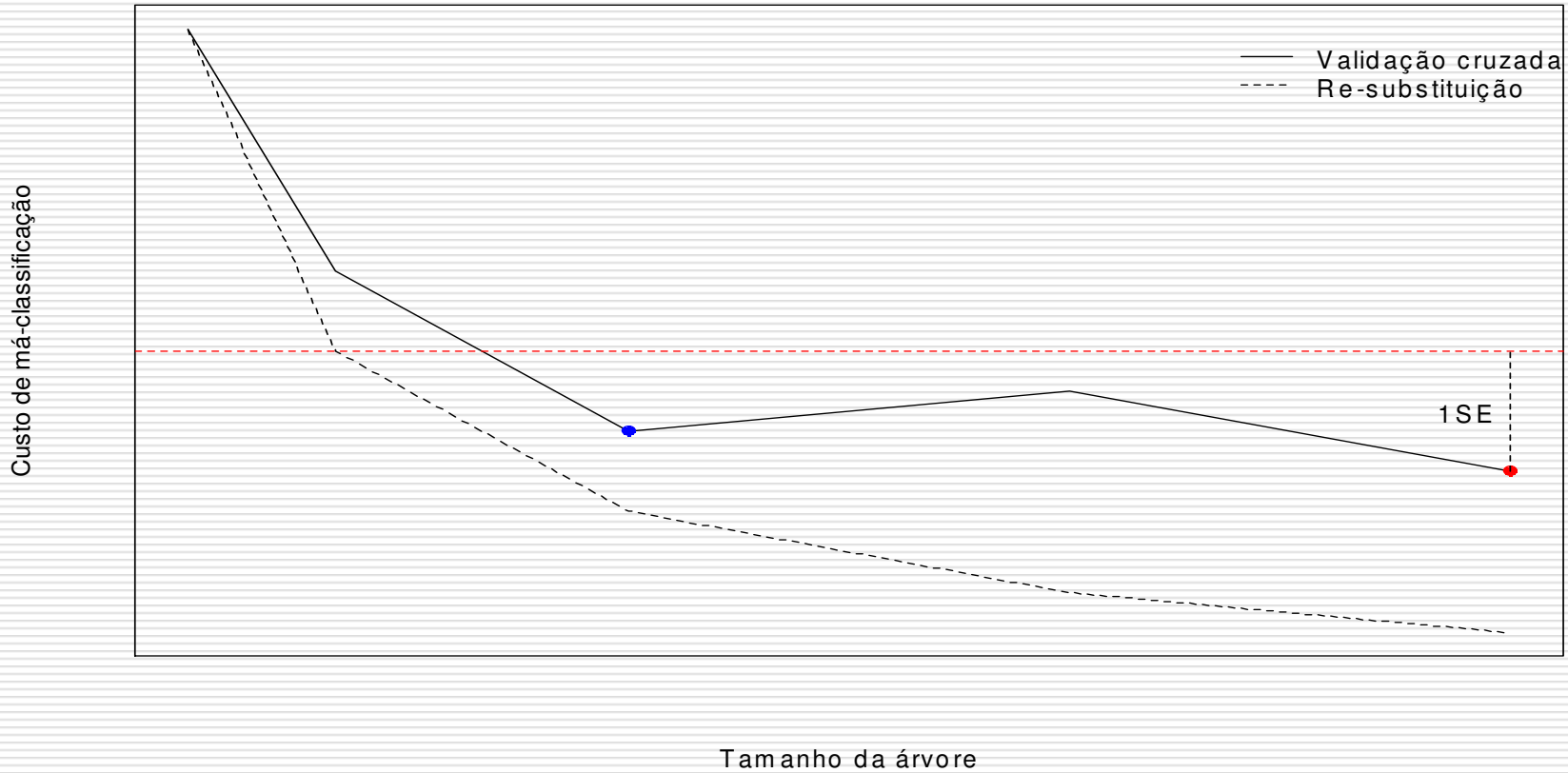

Figura 2 - Curva de custo complexidade.

1.3.4 Caracterização dos nós finais

- **Árvores de classificação:** por meio das proporções de ocorrências de cada uma das classes;
- **Árvores de regressão:** com a média das observações que formam o nó.
- **Predição:** Realizada conduzindo cada nova observação pela árvore e inferindo o valor da resposta de acordo com o valor característico do nó final ao qual foi alocada.

1.4 Exemplo

- **Dados:** Distribuição de 12 espécies de aranhas caçadoras capturadas em armadilhas em dunas holandesas (Van de Art e Smeeck Enserinck, 1975). Foram amostradas 28 locações.

- **Variáveis respostas:**

  - Abundâncias – tomadas as raízes quadrada

1.4 Exemplo

- **Variáveis ambientais:**
  - Mseca: logaritmo da porcentagem de matéria seca no solo;
  - Areia: logaritmo da porcentagem de cobertura com areia;
  - Galhos: logaritmo da porcentagem de cobertura com galhos e folhas;
  - Musgos: logaritmo da porcentagem de cobertura com musgos;
  - Capim: logaritmo da porcentagem de cobertura com capim;
  - Ref: reflexão da superfície do solo com o céu encoberto.
- Nota: Amplitude das variáveis ambientais divididas em 10 classes, correspondentes aos valores inteiros entre zero e nove.

1.4 Exemplo

Figura 2 – Aranhas caçadoras

1.4 Exemplo

Espécie: *Alopecosa accentuata*

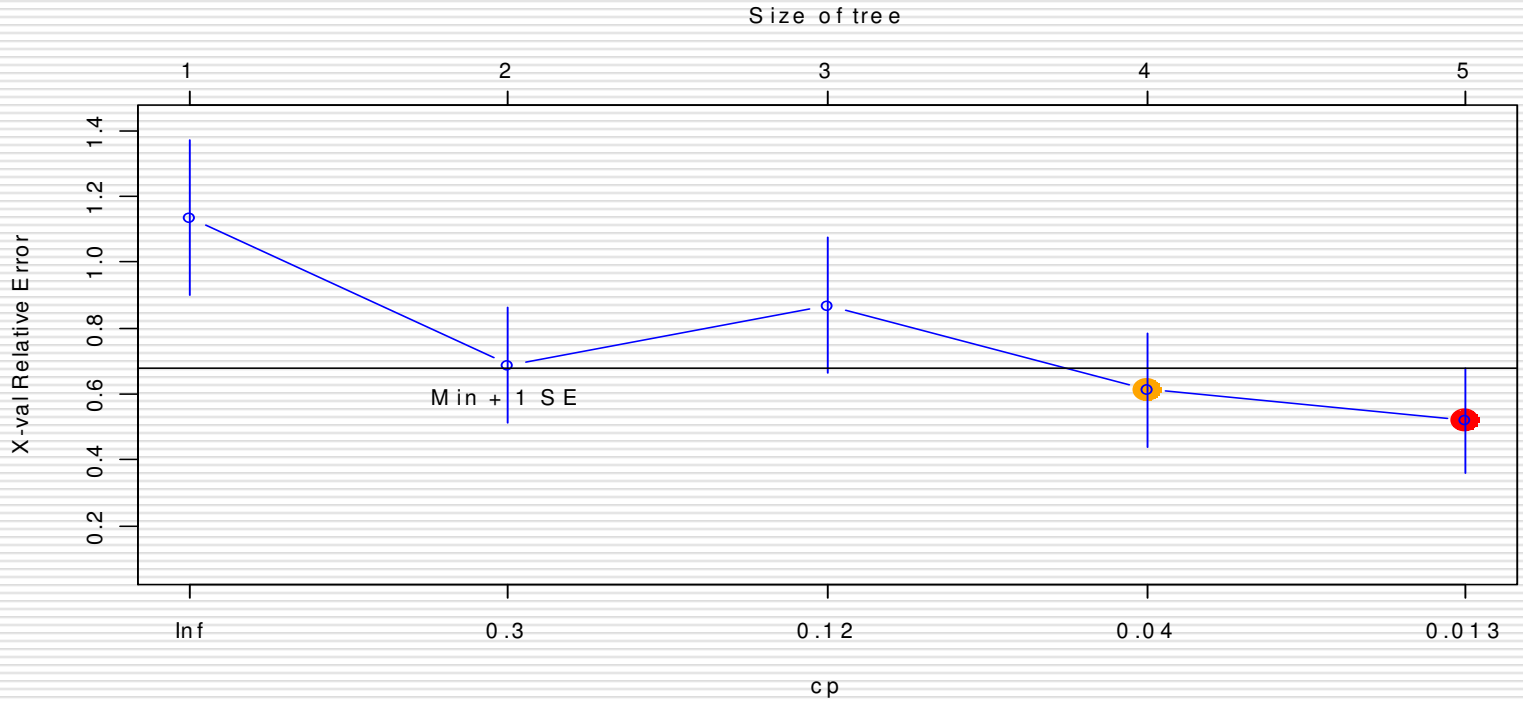

Figura 3 – Gráfico de custo-complexidade

1.4 Exemplo

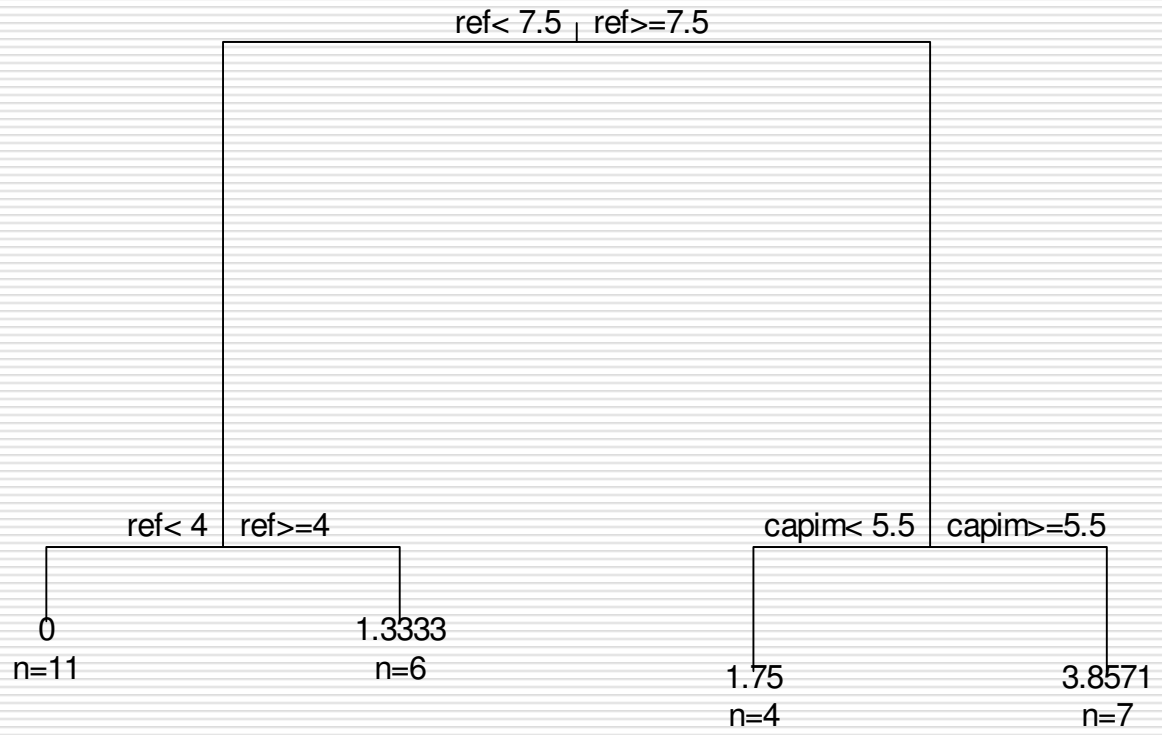

Figura 4 – Árvore de regressão para *Alopecosa accentuata*

1.4 Exemplo

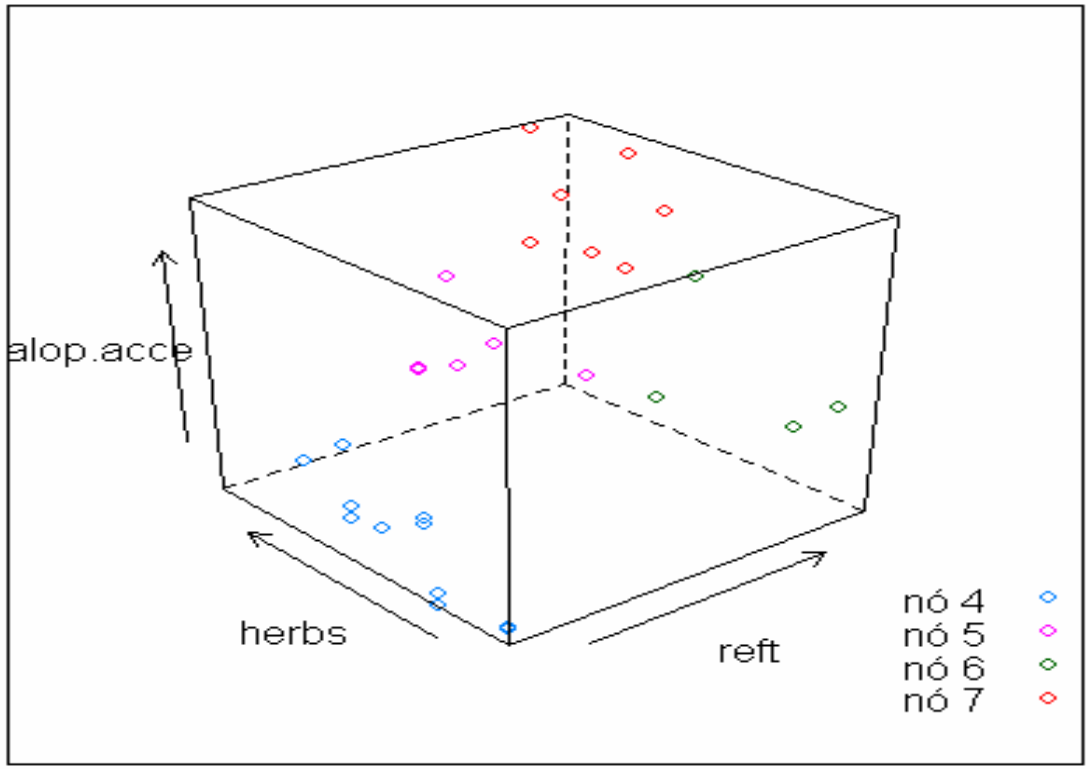

Figura 5 – Partições

1.4 Exemplo

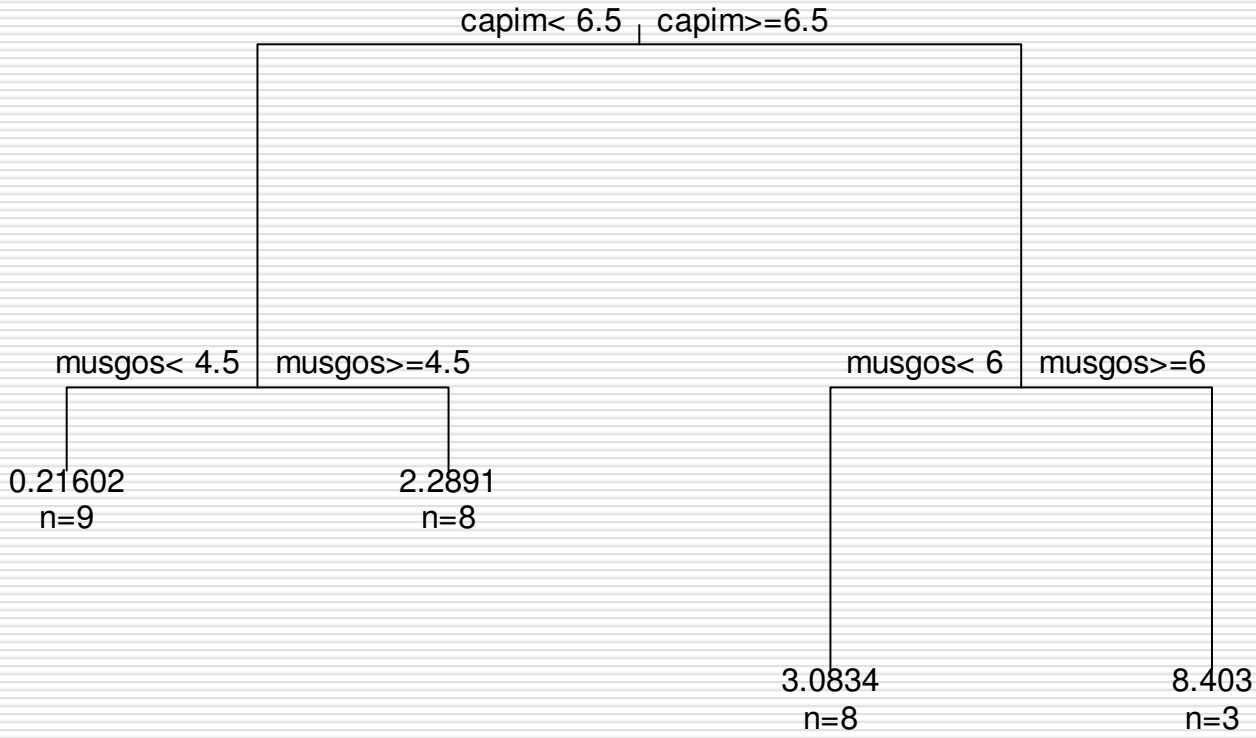

Figura 6 – Árvore de regressão para *Pardosa monticola*

1.4 Exemplo

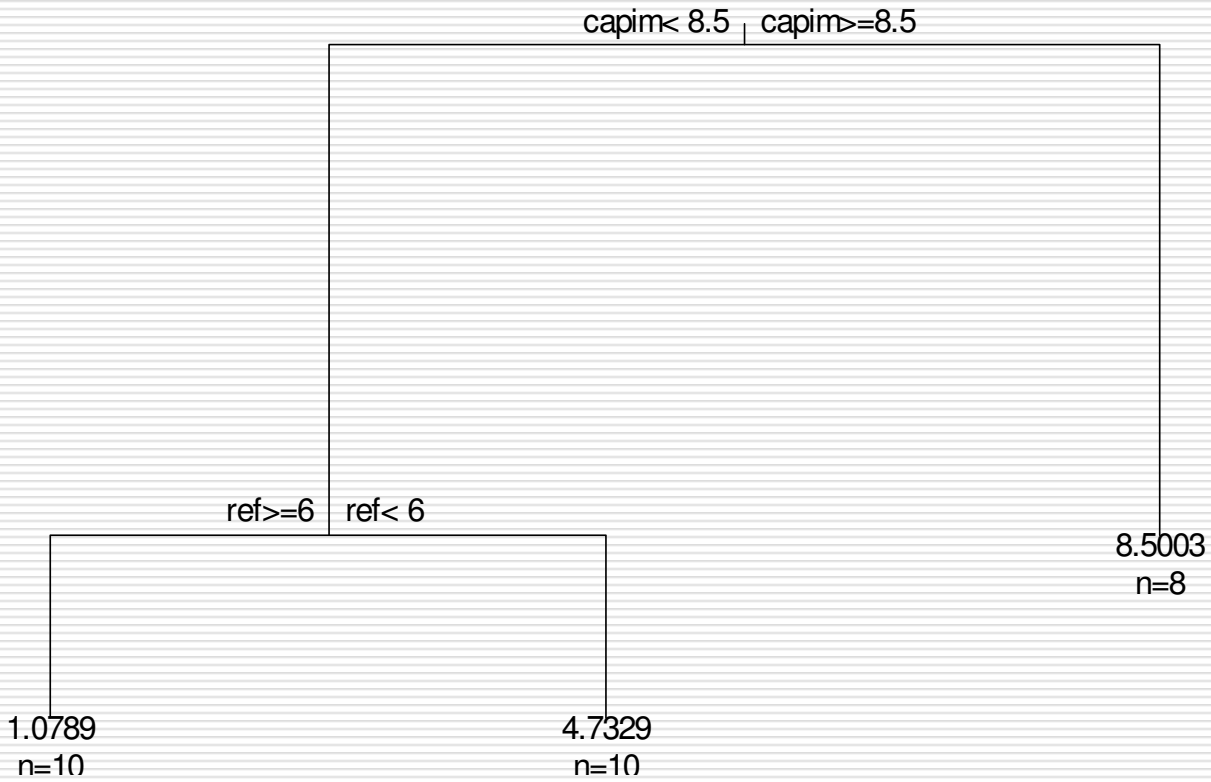

Figura 7 – Árvore de regressão para *Trochosa terricola*

1.4 Exemplo

Figura 8 – Árvore de regressão para *Pardosa lugubris*

1.4 Exemplo

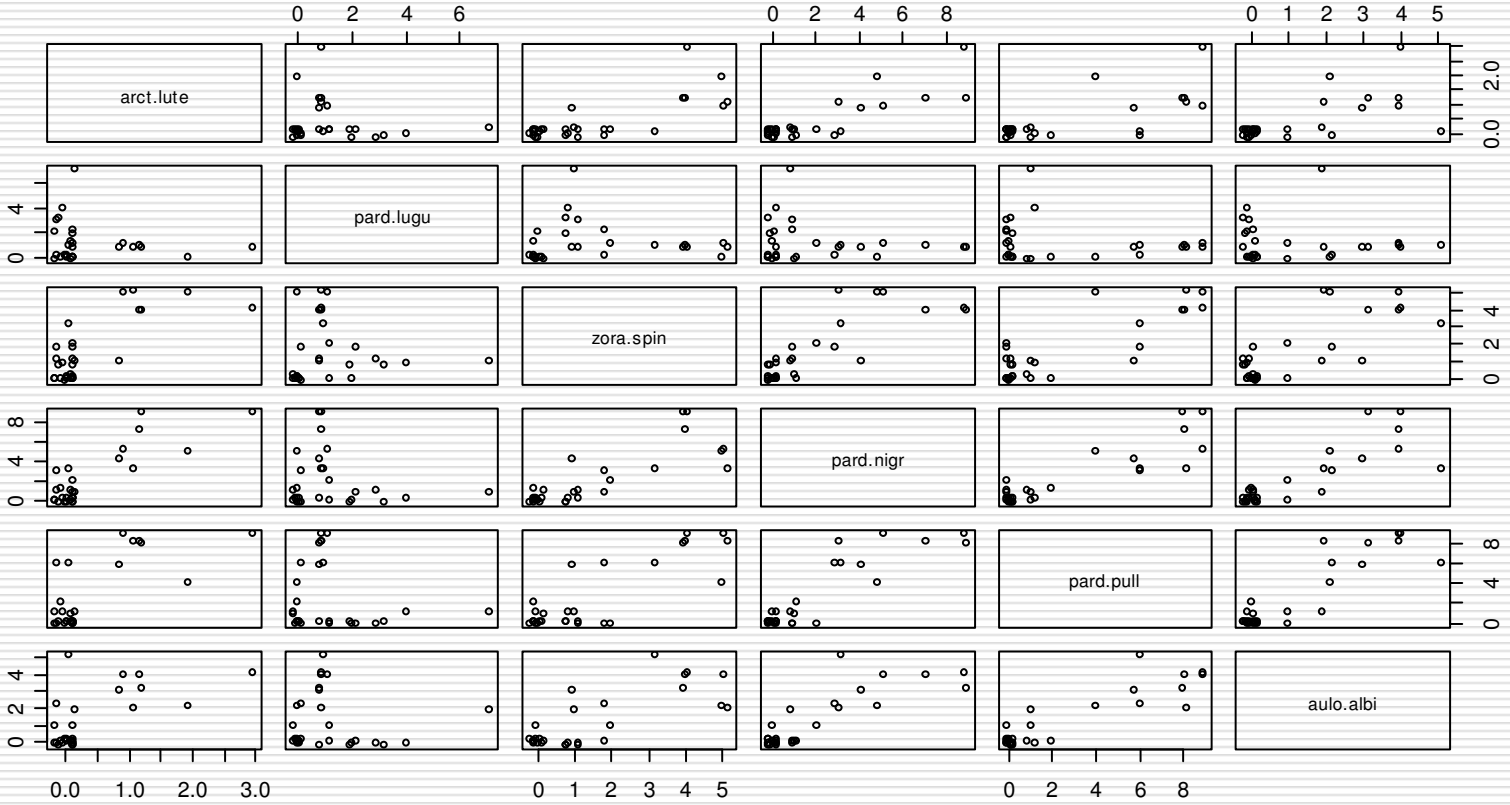

Figura 9 – Gráficos de dispersão

1.4 Exemplo

- Problemas:
  - Elevado número de espécies;
  - Correlação entre abundâncias das diferentes espécies.
- Solução:
  - Análise multivariada

2. Árvores de Regressão multivariadas (De Ath, 2002)

- Estudo da relação espécies/ambiente através da construção de árvores de regressão multivariadas.

- Objetivos :
  - Detectar quais fatores ambientais (ou combinações dos mesmos) são responsáveis pela distribuição espacial das 12 espécies de aranhas caçadoras.
  - Identificar e analisar a co-existência ou predominância de determinadas espécies em locações com diferentes características.

2. Árvores de Regressão multivariadas

Tabela 2 - Alternativas de medidas de impureza (construção da árvore) e de erro de predição (poda):

Descrição	Impureza	Erro de predição
Soma multivariada dos quadrados dos desvios em relação à média.	$\sum_{i,j}\left(y_{ij}-\bar{y}_j\right)^2$	$\sum_{j}\left(y^*-\bar{y}_j\right)^2$
Soma multivariada dos desvios absolutos em relação à mediana.	$\sum_{i,j}\left\|y_{ij}-\tilde{y}_j\right\|$	$\sum_{j}\left\|y^*-\tilde{y}_j\right\|$
Medidas de distância	$\sum_{i>k,k} d_{ik}^2$	$\sum_{i}\frac{d_i^{*2}}{n}-\sum_{i>k,k}\frac{d_{ik}^2}{n^2}$

2. Árvores de Regressão multivariadas

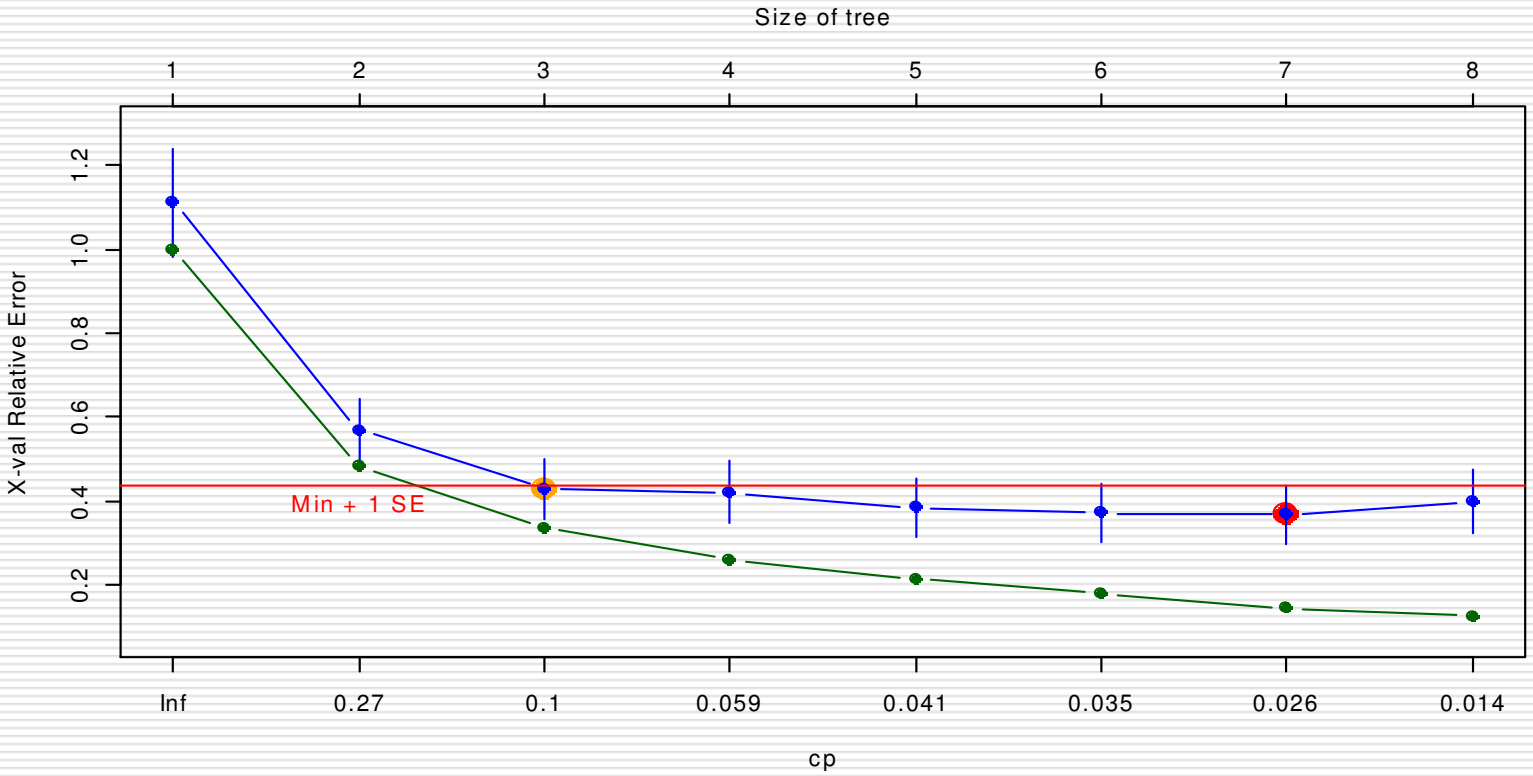

Figura 10 – Gráfico de complexidade para a árvore de regressão multivariada

2. Árvores de Regressão multivariadas

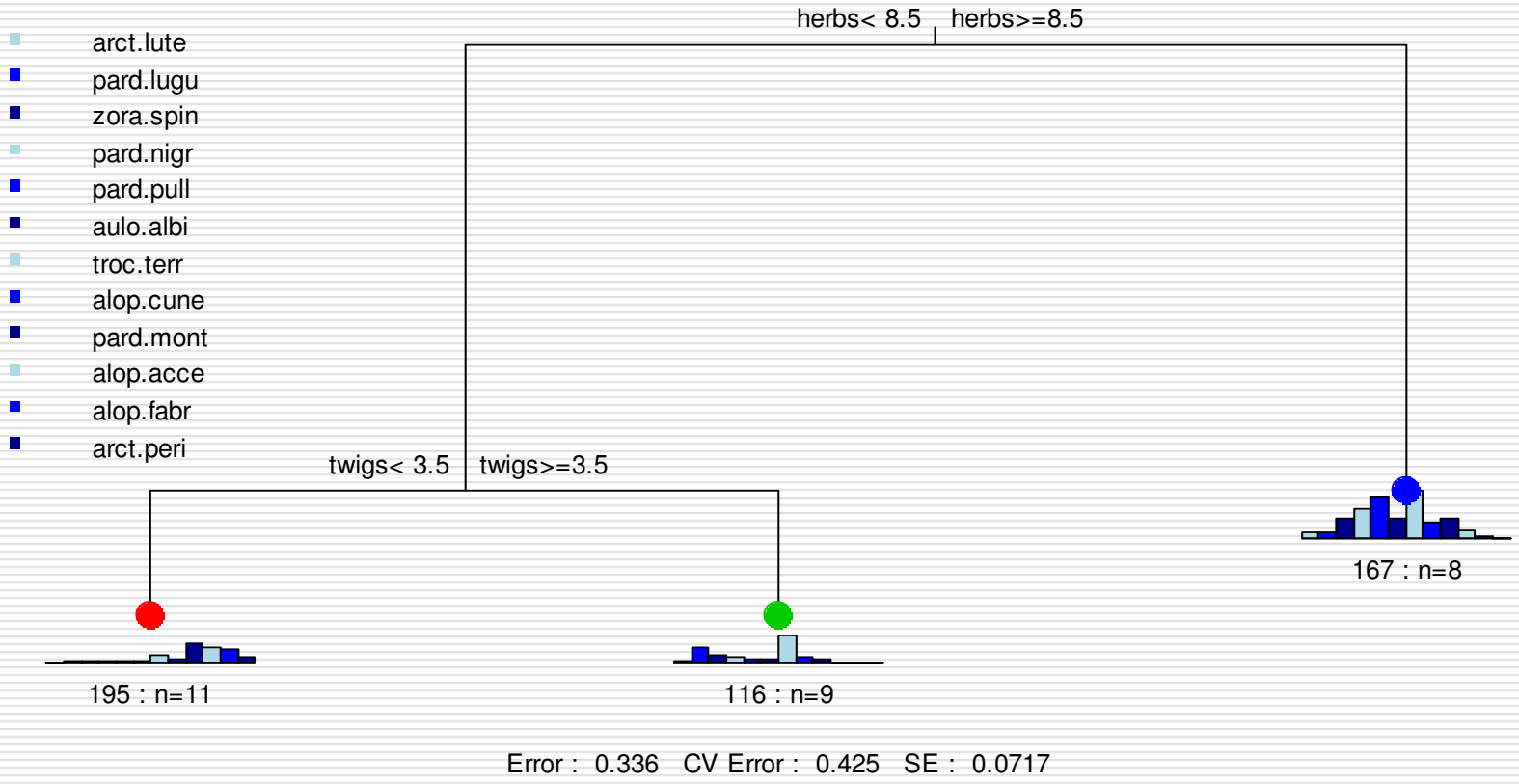

Figura 11 – Árvore de regressão multivariada

2. Árvores de Regressão multivariadas

- Biplots (Gabriel, 1971)
  - Gráfico bi-dimensional representando uma matriz de dados, com um ponto para cada uma das *n* observações e um vetor para cada uma das *p* variáveis
  - A disposição dos pontos e vetores nos diferentes quadrantes do gráfico representa as correlações entre as variáveis e as observações.

2. Árvores de Regressão multivariadas

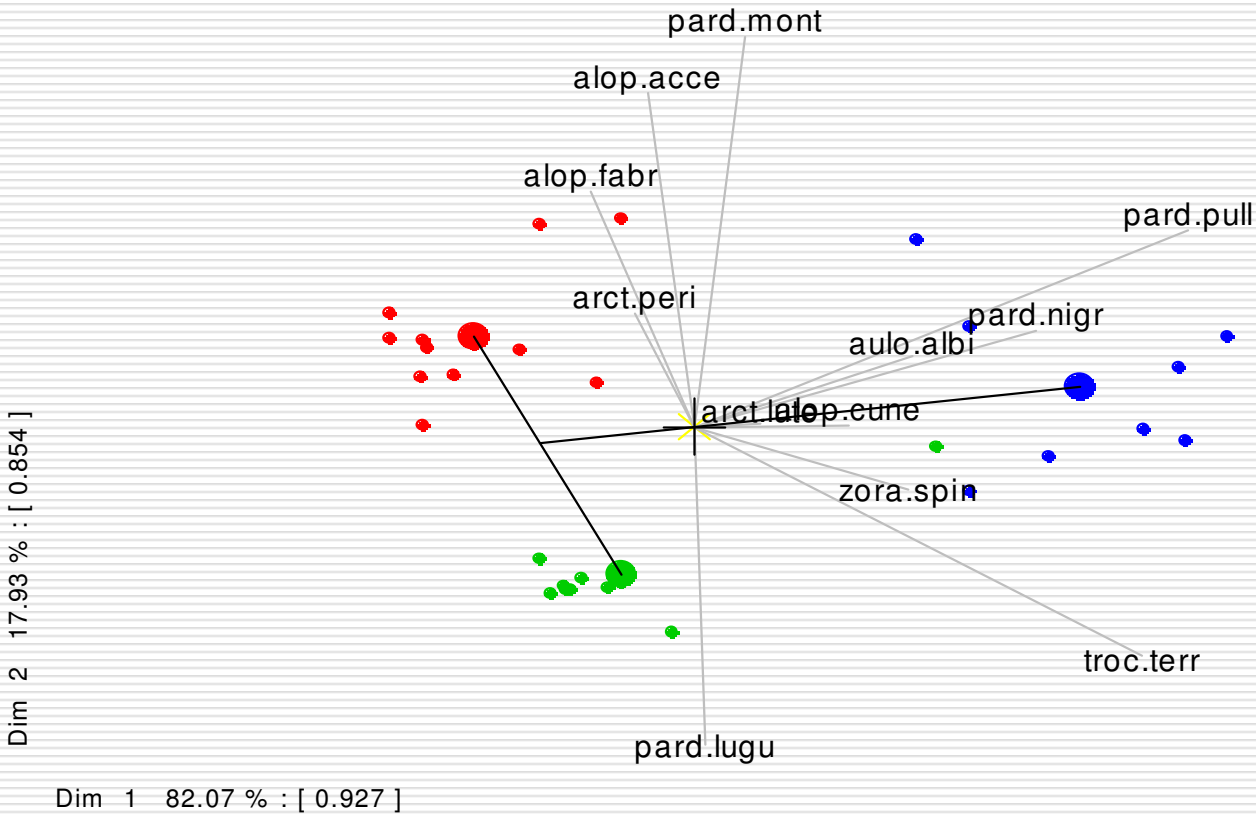

Figura 12 – Tree biplot

3. Conclusão

- CART
  - Alternativa não paramétrica a diversos procedimentos estatísticos;
  - Flexibilidade e simplicidade da técnica;
  - Extensão multivariada: análise conjunta de duas ou mais variáveis respostas;
  - Identificação de fatores ambientais associados à abundância de espécies de aranhas caçadoras.

4. Referências

- BREIMAN, L., J.H. FRIEDMAN, R.A. OLSHEN, AND C.G. STONE. (1984), **Classification and regression trees**. Wadsworth International Group, California, 358p, 1984.

- DE'ATH, G. Multivariate Regression Trees: A New Technique for Modeling Species-Environment Relationships. **Ecology**, 83, 4, 1105–1117, 2002.

- GABRIEL, K. R. The biplot graphical display of matrices with application to principal component analysis. **Biometrika,** 58, 453–467,1971.

- VAN DE ART, P.J., N. SMEECK ENSERINCK. Correlations between distributions of hunting spiders (Lycosidae, Ctenidae) and environmental characteristics in a dune área. **Netherlands Journal of Zoology**, 25,1-45, 1975.