.

ÁRVORES DE CLASSIFICAÇÃO E REGRESSÃO

Cesar Augusto Taconeli

Ilhéus-BA
Junho 2013

Sumário

CART - Classification And Regression Trees

⇒ BREIMAN, L.; FRIEDMAN, J. T., OLSHEN, R. A,. and STONE, C. J. (1984). Classification and regression trees.

⇒ Métodos de classificação e regressão baseados em partições binárias recursivas de uma amostra.

⇒ Estrutura gerada (amostra, sub-amostras e partiçõeses) representada por meio de um gráfico (Árvore).

CART - Classification And Regression Trees

$\Rightarrow$ Dados: uma (ou mais) variável resposta e um conjunto de variáveis preditoras.

$\Rightarrow$ Árvore de classificação - variável resposta numérica;

$\Rightarrow$ Árvore de regressão - variável resposta categórica.

Terminologia:

$\Rightarrow$ Nó: qualquer amostra ou sub-amostra representada numa Árvore;

$\Rightarrow$ Partição: regra responsável pela partição de um nó;

$\Rightarrow$ Nó inicial: a amostra original;

$\Rightarrow$ Nó pai: a amostra que é partida (em duas subamostras);

$\Rightarrow$ Nó filho: resultante da partição do nó pai;

$\Rightarrow$ Nó final: o nó que não dá origem a novos nós.

Ilustração

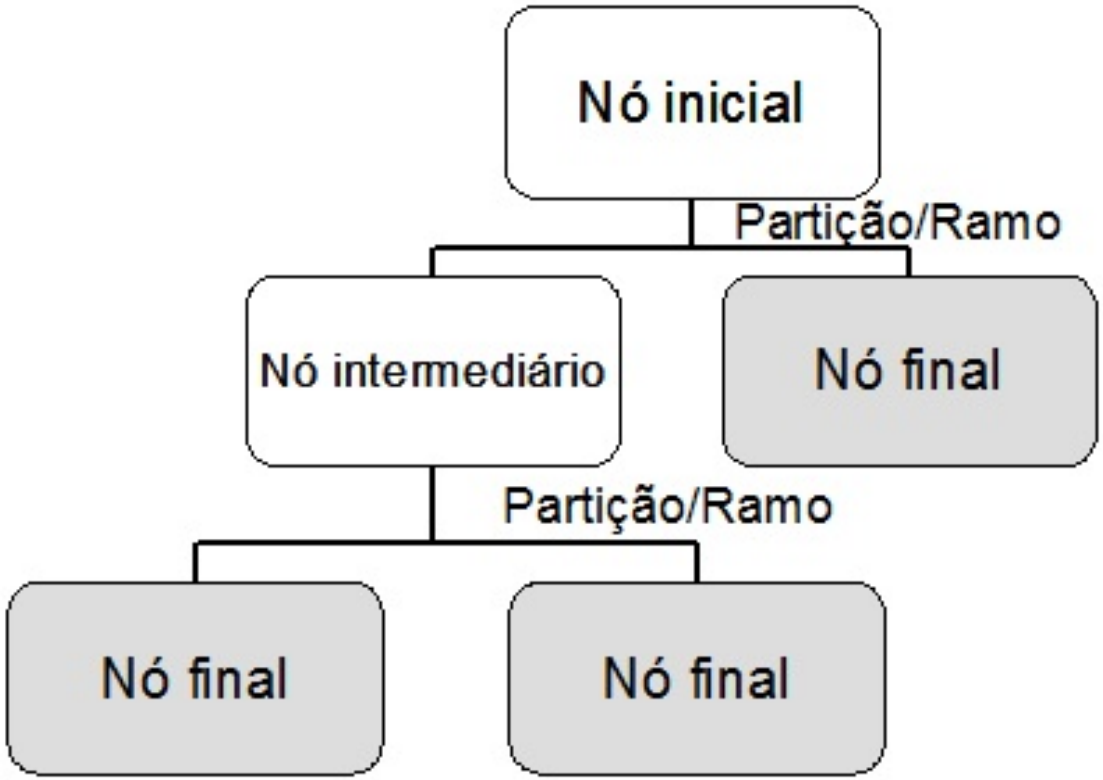

Figura 1 - Ilustração de uma árvore de classificação (ou regressão)

Aplicações:

$\Rightarrow$ Exploração de dados;

$\Rightarrow$ Modelagem preditiva.

Alternativa (ou complemento) a diversas técnicas estatísticas de classificação e regressão.

Aspectos positivos do CART

$\Rightarrow$ Rápida construção;

$\Rightarrow$ Resultados de fácil interpretação;

$\Rightarrow$ Permite lidar com variáveis em diferentes escalas (nominal, ordinal, numérica);

$\Rightarrow$ Detecção automática de interações entre as variáveis preditoras.

Aspectos positivos do CART

$\Rightarrow$ Dispõe de métodos para lidar com dados missing;

$\Rightarrow$ Invariante a transformações monótonas dos preditores;

$\Rightarrow$ Robusto quanto a presença de outliers na amostra.

Aspectos negativos do CART

$\Rightarrow$ Instabilidade frente a alterações nos dados amostrais;

$\Rightarrow$ Instabilidade causada por variáveis preditoras correlacionadas;

$\Rightarrow$ Viés de partição;

$\Rightarrow$ Ausência de significâncias estatísticas na construção da árvore.

Nota- Todos estes pontos negativos são contornáveis mediante modificações ou extensões do algoritmo original (veremos algumas alternativas).

Algoritmo:

1- Partição recursiva da amostra original e das sub-amostras geradas (ramificação);

2- Junção das partições executadas no passo 1, da base à origem da árvore (poda);

3- Seleção de uma árvore, dentre aquelas geradas no processo de poda;

4- Classificação dos nós finais e interpretação da árvore.

Dados

$$
\begin{array}{cccccc}
y_1 & x_{11} & x_{21} & \cdots & x_{m1} \\
y_2 & x_{12} & x_{22} & \cdots & x_{m2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
y_n & x_{1n} & x_{2n} & \cdots & x_{mn}
\end{array}
$$

Passo 1 - Ramificação

$\Rightarrow$ Partições executadas a partir de regras baseadas nos valores observados das variáveis preditoras.

$\Rightarrow$ Regras diferentes devem ser consideradas de acordo com a escala das variáveis preditoras.

Passo 1 - Ramificação

$\Rightarrow$ Variável preditora numérica:

Seja X_i uma variável preditora numérica, com valores amostrados $x_{i1}, x_{i2}, \ldots, x_{in}$.

$\Rightarrow$ Partições candidatas: $x_i \leq x_{ij}$, para todo* $\quad j = 1, 2, \ldots, n$.

* Restrições quanto ao tamanho do nó pai e dos nós filhos devem ser obervadas.

Passo 1 - Ramificação

Seja X_i uma variável preditora categórica, sendo $U = \{A.B, C, \ldots\}$ as categorias observadas na amostra.

$\Rightarrow$ Partições candidatas: $x_i \in S$, para todo* $S \subset U$.

Seja X_i uma variável preditora categórica ordenável, sendo $U = \{A_1.A_2, A_3, \ldots, A_k\}$ as categorias observadas na amostra.

$\Rightarrow$ Partições candidatas:
$x_i \in S$, sendo $S = \{A_1, A_2, A_3, \ldots, A_j\}$, para todo* $j = 1, 2, ..., k$.

Passo 1 - Ramificação

$\Rightarrow$ Executar a partição candidata que melhor 'explicar' a resposta.

$\Rightarrow$ Aquela responsável pela formação de nós mais homogêneos (menos impuros) internamente e heterogêneos entre si.

$\Rightarrow$ Deve-se definir uma medida adequada de impureza de acordo com a natureza do problema e a escala da variável resposta.

Passo 1 - Ramificação

$\Rightarrow$ Exemplos de medidas de impureza:

Seja Y uma variável categórica com categorias $A_1, A_2, \ldots, A_k$ presentes na amostra.

Suponha um nó t com $\hat{p}_1, \hat{p}_2, ..., \hat{p}_k$ as proporções de observações pertencentes a cada categoria.

$\Rightarrow$ Índice de Gini: $\phi(t) = \sum_{i=1}^{k} \hat{p}_i(1 - \hat{p}_i)$

$\Rightarrow$ Entropia (deviance): $\phi(t) = -\sum_{i=1}^{k} \hat{p}_i \log \hat{p}_i$

Passo 1 - Ramificação

Seja Y uma variável numérica e $y_1, y_2, \ldots, y_{n_t}$ os valores observados de Y em um nó t.

$\Rightarrow$ Índice anova: $\phi(t) = \sum_{i=1}^{n_t}(y_i - \bar{y}_t)^2$

Sendo $\bar{y}_t = \frac{\sum_{i=1}^{n_t} y_i}{n_t}$

Medidas de impureza apropriadas são definidas para dados de contagens, censurados, multivariados, longitudinais...

Passo 1 - Ramificação

$\Rightarrow$ Executar a partição que proporcionar maior redução na impureza do nó, ou seja, que maximizar:

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

onde t_L e t_R indicam os nós constituídos e n_L e n_R seus respectivos números de observações.

$\Rightarrow$ Executar, recursivamente, a partição dos nós gerados.

Passo 2 - Poda

$\Rightarrow$ Realizada a partir de uma 'grande' árvore, constituída no passo 1, desfazendo-se sucessivamente as partições executadas.

$\Rightarrow$ Baseada na minimização da chamada função de custo-complexidade.

$$R_\alpha(T) = R(T) + \alpha \mid \tilde{T} \mid$$

Passo 2 - Poda

onde:

$\Rightarrow$ T - árvore;

$\Rightarrow$ $\tilde{T}$ - conjunto de nós finais de T;

$\Rightarrow$ $| \tilde{T} |$ - número de nós finais de T;

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

$\Rightarrow$ α - parâmetro de complexidade

Aumentando α a partir de zero, obtém-se uma sequência aninhada de árvores que maximizam a função de custo complexidade, das quais uma será selecionada.

Passo 3 - Seleção

$\Rightarrow$ Avaliação da curva de custo-complexidade;

$\Rightarrow$ Custo de má-classificação estimado por validação cruzada;

$\Rightarrow$ Regra de um desvio padrão (1-SE Rule).

Passo 3 - Seleção

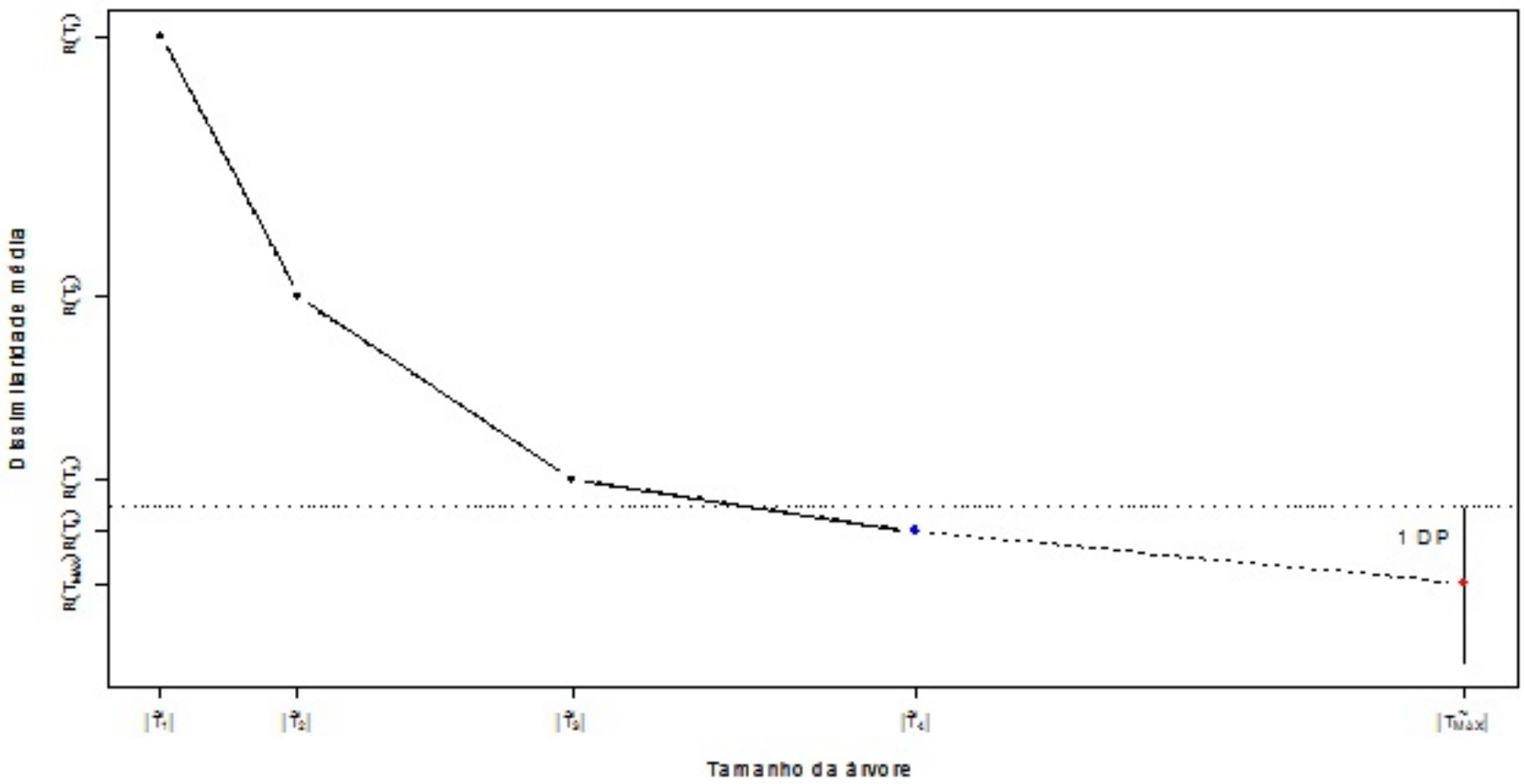

Figura 2 - Curva de custo-complexidade.

Passo 4 - Classificação e interpretação

$\Rightarrow$ Nós finais classificados de acordo com a distribuição dos resultados da variável resposta.

Algumas possibilidades:

$\Rightarrow$ Categoria mais frequente no nó (resposta categórica);

$\Rightarrow$ Média do nó (resposta numérica)...

Passo 4 - Classificação e interpretação

$\Rightarrow$ Interpretação dos resultados diretamente pela árvore.

$\Rightarrow$ Verficicação das variáveis que produzem as partições, suas interações e os nós finais resultantes.

Algumas extensões

⇒ Análise de dados ordinais

PICCARRETA, Raffaella. Classification trees for ordinal variables. Computational Statistics, v. 23, n. 3, p. 407-427, 2008.

⇒ Análise de dados de sobrevivência:

SEGAL, Mark Robert. Regression trees for censored data. Biometrics, p. 35-47, 1988.

LEBLANC, Michael; CROWLEY, John. Relative risk trees for censored survival data. Biometrics, p. 411-425, 1992.

Algumas extensões

$\Rightarrow$ Análise de dados longitudinais

SEGAL, Mark Robert. Tree-structured methods for longitudinal data. Journal of the American Statistical Association, v. 87, n. 418, p. 407-418, 1992.

SELA, Rebecca J.; SIMONOFF, Jeffrey S. RE-EM trees: a data mining approach for longitudinal and clustered data. Machine learning, v. 86, n. 2, p. 169-207, 2012.

Algumas extensões

⇒ Análise de dados multivariados

DE'ATH, Glenn. Multivariate regression trees: a new technique for modeling species-environment relationships. Ecology, v. 83, n. 4, p. 1105-1117, 2002.

⇒ Análise de dados de mistura

HOUSEMAN, E. Andres et al. Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. Bmc Bioinformatics, v. 9, n. 1, p. 365, 2008.

Bagging (Bootstrap aggregating)

BREIMAN, Leo. Bagging predictors. Machine learning, v. 24, n. 2, p. 123-140, 1996.

$\Rightarrow$ Árvores podem ser bastante instáveis frente a pequenas modificações na amostra.

$\Rightarrow$ Consequência: classificações pouco precisas.

$\Rightarrow$ Alternativa: Construir múltiplas árvores e agregar as classificações resultantes.

Bagging (Bootstrap aggregating)

Algoritmo

1- Seleção de uma amostra bootstrap a partir da amostra original;

2- Construção da árvore com base na amostra bootstrap selecionada;

3- Classificação dos elementos segundo a árvore construída;

Bagging (Bootstrap aggregating)

4- Repetição dos passos 1 a 3 um grande número (B) de vezes;

5- Obtenção de uma classificação agregada para cada elemento da amostra original (ou da amostra teste).

$\Rightarrow$ Classificação por voto (pela classificação mais frequente);

$\Rightarrow$ Classificação pela média.

Random forests

BREIMAN, Leo. Random forests. Machine learning, v. 45, n. 1, p. 5-32, 2001.

$\Rightarrow$ Extensão do algoritmo bagging;

$\Rightarrow$ Visa a obtenção de um grande número de árvores 'decorrelacionas';

$\Rightarrow$ Modificação - a cada partição, apenas um subconjunto de variáveis preditoras é considerado para a partição.

Random forests

Algoritmo:

1- Seleção de uma amostra bootstrap a partir da amostra original;

2- Construção da árvore com base na amostra bootstrap selecionada, respeitando o seguinte procedimento:

Random forests

Para cada partição:

i. Selecione aleatoriamente $p < m$ variáveis preditoras;

ii. Verifique a melhor partição proporcionada por estas p variáveis;

iii. Execute a melhor partição.

3- Classificação dos elementos segundo a árvore construída;

4- Repetição dos passos 1 a 3 um grande número (B) de vezes;

5- Obtenção de uma classificação agregada para cada elemento da amostra original (ou da amostra teste).

Random forests

Alguns 'sub-produtos' do random forest (e de outros algoritmos de agregação):

$\Rightarrow$ Medida de importância da variável

1- Calculada somando, para cada variável, a explicação proporcionada por suas partições no conjunto de árvores ou

2- Calculada pela diferença nas taxas de más-classificações para as árvores construídas com as variáveis preditoras na forma original vs aquelas obtidas permutando aleatoriamente os valores da j-ésima variável ($j = 1, 2, \ldots, m$).

$\Rightarrow$ Gráfico de proximidades

$\Rightarrow$ Medida de proximidade entre dois elementos: Número de árvores em que eles são alocados a um mesmo nó final.

$\Rightarrow$ Matriz de proximidades: Matriz $n \times n$ em que em cada entrada tem-se a medida de proximidade de um par de elementos.

$\Rightarrow$ Análise - Visualização do gráfico do escalonamento multidimensional da matriz de proximidades.

KRUSKAL, Joseph B. Nonmetric multidimensional scaling: a numerical method. Psychometrika, v. 29, n. 2, p. 115-129, 1964.

Cuidado! Gráficos de proximidade para random forests, em geral, têm resultados muito semelhantes.

Conditional trees

HOTHORN, Torsten; HORNIK, Kurt; ZEILEIS, Achim. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics, v. 15, n. 3, 2006.

Objetivos:

$\Rightarrow$ Eliminar possível viés de partição (variáveis preditoras numéricas tendem a 'inibir' partições baseadas em variáveis categóricas);

$\Rightarrow$ Incorporar medidas de significância à construção das árvores (roupagem mais estatística).

Conditional trees

$\Rightarrow$ Seleção da partição em duas etapas. Para cada nó:

1- Determinação da variável preditora com maior associação com a variável resposta;

2- Uma vez selecionada a variável, determinação do ponto de corte.

$\Rightarrow$ Para cada nó, a decisão por parti-lo ou não é baseada no teste da hipótese nula de que nenhuma variável preditora está associada à variável resposta (via permutação).

Árvores de Classificação e Regressão Multivariadas

DE'ATH, Glenn. Multivariate regression trees: a new technique for modeling species-environment relationships. Ecology, v. 83, n. 4, p. 1105-1117, 2002.

Extensão do algoritmo original de forma a contemplar múltiplas respostas.

$\Rightarrow$ Proposição de medidas de impureza e de má-classificaço adequadas.

$\Rightarrow$ Métodos gráficos para exploração dos resultados.

Árvores de Classificação e Regressão Multivariadas

Medidas de custo de má-classificação:

⇒ Coeficiente de entropia generalizado;

⇒ Índice Anova Multivariado;

⇒ Coeficientes baseados na matriz de distâncias (dissimilaridades).

⇒ Visualização - biplot.

GABRIEL, Karl Ruben. The biplot graphic display of matrices with application to principal component analysis. Biometrika, v. 58, n. 3, p. 453-467, 1971.

Referências

BREIMAN, L.; FRIEDMAN, J. T., OLSHEN, R. A,. and STONE, C. J. (1984). Classification and regression trees.

BREIMAN, Leo. Bagging predictors. Machine learning, v. 24, n. 2, p. 123-140, 1996.

BREIMAN, Leo. Random forests. Machine learning, v. 45, n. 1, p. 5-32, 2001.

DE'ATH, Glenn. Multivariate regression trees: a new technique for modeling species-environment relationships. Ecology, v. 83, n. 4, p. 1105-1117, 2002.

Referências

GABRIEL, Karl Ruben. The biplot graphic display of matrices with application to principal component analysis. Biometrika, v. 58, n. 3, p. 453-467, 1971.

HOTHORN, Torsten; HORNIK, Kurt; ZEILEIS, Achim. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics, v. 15, n. 3, 2006.

HOUSEMAN, E. Andres et al. Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. Bmc Bioinformatics, v. 9, n. 1, p. 365, 2008.

Referências

KRUSKAL, Joseph B. Nonmetric multidimensional scaling: a numerical method. Psychometrika, v. 29, n. 2, p. 115-129, 1964.

LEBLANC, Michael; CROWLEY, John. Relative risk trees for censored survival data. Biometrics, p. 411-425, 1992.

PICCARRETA, Raffaella. Classification trees for ordinal variables. Computational Statistics, v. 23, n. 3, p. 407-427, 2008.

Referências

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Viena, Áustria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

SEGAL, Mark Robert. Regression trees for censored data. Biometrics, p. 35-47, 1988.

SEGAL, Mark Robert. Tree-structured methods for longitudinal data. Journal of the American Statistical Association, v. 87, n. 418, p. 407-418, 1992.

SELA, Rebecca J.; SIMONOFF, Jeffrey S. RE-EM trees: a data mining approach for longitudinal and clustered data. Machine learning, v. 86, n. 2, p. 169-207, 2012.