*Árvores de Regressão* : Consistem na partição do espaço das covariáveis em regiões retangulares e no ajuste de um modelo simples (como uma constante) em cada uma delas, grande variedade de algoritmos.

O termo árvore de regressão é aplicado ao caso de variável resposta numérica e o termo árvore de classificação para o caso de variável resposta categórica. Em ambos os casos, as covariáveis podem ser categóricas e/ou numéricas.

Dentre os principais atrativos de árvores de classificação e regressão, destacam-se:

Baseiam-se em um conjunto mínimo de pressupostos; ,,,Servem como alternativa a diversos métodos estatísticos de classificação e regressão; ,,,

Permitem lidar com dados de estrutura complexa (elevada dimensão, dados ausentes, interações de diferentes ordens entre as covariáveis); ,,,

Produzem resultados simples e de fácil interpretação.

Seja y a variável resposta e x = (x1, x2, ..., xp) o vetor de covariáveis. Considere uma amostra de n observações de y e x.

O método CART inicia com a partição da amostra original em duas, segundo alguma regra do tipo

xk ≤ c | xk > c, ,,, para alguma covariável xk numérica e c algum valor amostrado de xk , ou

xk ∈ A | xk ∈/ A, ,,, para uma variável xk categórica e A uma particular categoria (ou um subconjunto de categorias) de xk .

Uma vez efetuada a partição, temos o espaço das covariáveis dividido em duas regiões, R1 e R2.

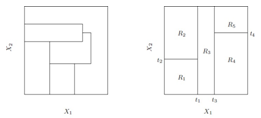
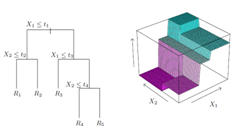
A variável responsável pela partição e o ponto de corte são escolhidos de forma que proporcionem o melhor ajuste possível para y .

Na sequência, o processo de partição é repetido em R1 e em R2, novamente buscando a variável e respectivo ponto de corte que proporcionem melhor ajuste.

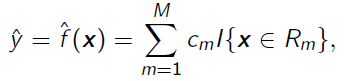
Neste passo, temos quatro regiões delimitadas no espaço das covariáveis: R3 e R4 (formadas a partir de R1); R5 e R6 (formadas a partir de R2).

O processo é repetido sucessivamente. No final, teremos M regiões delimitadas no espaço das covariáveis, que denotaremos por

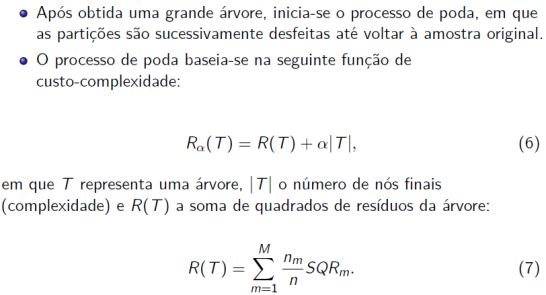
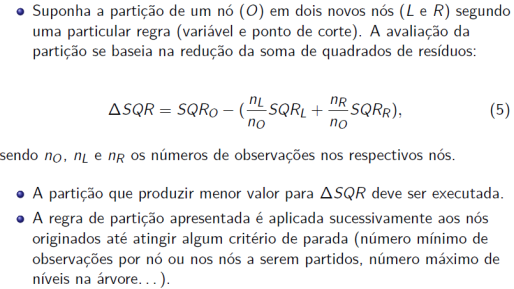
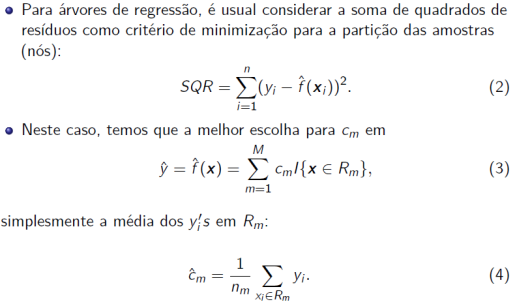
R1, R2, ..., RM . O resultado da aplicação do método CART pode ser representado por um diagrama contendo as partições e os grupos constituídos (nós), que denominamos árvore.

Podemos ainda expressar o resultado da aplicação do algoritmo CART por meio de um modelo de regressão na forma:

 sendo *cm* uma constante ajustada na região *Rm*, *i* = 1*,* 2*, ..., M.*

*Seleção das Partições – 6 e 7 processo de poda.*



O parâmetro α na função de custo-complexidade controla a complexidade do modelo.

Para diferentes valores de alfa tem-se diferentes árvores minimizando R alfa(T).

Tomando alfa = 0 tem-se como solução a maior árvore disponível (não podada), uma vez que não se penaliza sua complexidade.

Para alfa tendendo ao infinito tem-se penalização máxima para a complexidade e a solução é a não partição da amostra original.

Variando alfa a partir de zero tem-se uma sequência de árvores aninhadas, cada uma ótima para seu particular tamanho (número de nós finais).

É usual representar a função de custo-complexidade por meio de uma curva (versus alfa e ou |T |).

*Seleção do modelo:* Uma vez definida a sequência de árvores aninhadas, deve-se identificar, nessa sequência, a árvore ótima (correspondente à melhor escolha para α). Nesta etapa, é comum utilizar validação cruzada.

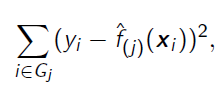
Seleção por *validação cruzada* é descrita na sequência:

Passo 1: Identificação de uma sequência de valores α1, α2, . . . , αk para α cada qual indicando uma das árvores na sequência aninhada como aquela que minimiza a função de custo-complexidade;

Passo 2: Dividir a base de dados em s grupos de tamanho (aproximado) s/n: G1, G2, ..., Gs ;

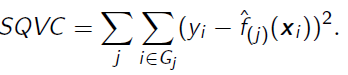
Passo 3: Ajustar o modelo à base completa (exceto pelas observações em Gj ) e determinar T1, T2, ..., Tk ;

Passo 4: Calcular a predição para cada observação i em Gj sob cada modelo Tj , j = 1, 2, ..., k;

Passo 5: Calcular a soma de quadrados dos erros de predição para o conjunto de observações em Gi : 

em que fˆ(j)(.) denota a predição sob o modelo ajustado sem as observações em Gj .

·

Passo 6: Os passos 3, 4 e 5 são repetidos para cada um dos demais grupos Gj . Ao término, para cada árvore T1, T2, ..., Tk tem-se a respectiva soma de quadrados de predição obtida por validação cruzada:Pode-se então selecionar a árvore que produz menor valor de SQVC ou a menor árvore tal que seu SVQC não seja muito maior daquela que produz SQVC mínimo.

Na prática, usa-se a regra do erro padrão, em que se seleciona a menor árvore tal que seu SQVC não exceda o SQVC mínimo por mais de um erro padrão de SQVC (estimado também na validação cruzada).

*Arvores de Classificação:*se aplicam quando a variável resposta é categórica (binária ou politômica);

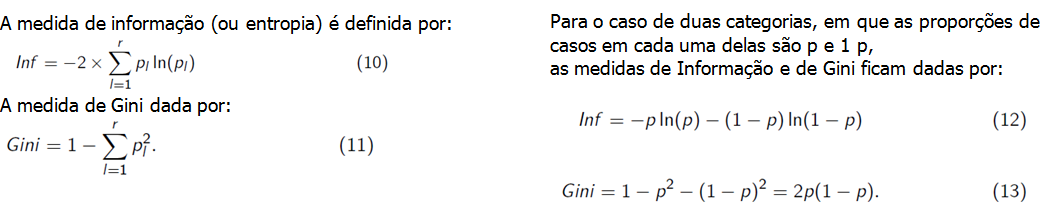
O algoritmo de árvores de classificação é semelhante ao de árvores de regressão, com algumas adaptações.

A diferença mais importante é a troca da soma de quadrados dos resíduos por alguma medida de heterogeneidade mais apropriada para dados categóricos.

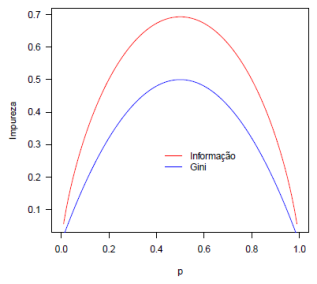
Dentre as alternativas, temos os critérios de Gini e da informação, conforme apresentados na sequência.

Vamos considerar um problema de classificação em que a resposta tenha r categorias, denotadas por 1, 2, ..., r .

Considere uma amostra (ou um nó) e p1, p2, ..., pr as proporções com que cada categoria é observada.

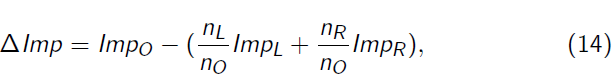


A Figura 1 apresenta o comportamento das medidas de informação e Gini para o caso de duas categorias.



Como pode ser observado na Figura 1, ambas as medidas são minimizadas quando os indivíduos da amostra pertencem a um mesmo grupo (pl ! 1, para algum l) e maximizadas quando as proporções são iguais nas diferentes categorias (p1 = p2 = ... = pr ).

Suponha a partição de um nó (O) em dois novos nós (L e R) segundo uma particular regra (variável e ponto de corte). A avaliação da partição se baseia na redução da medida de impureza:



em que Imp denota, genericamente, a medida de informação, de Gini ou qualquer outra medida de impureza.

Em árvores de classificação é comum classificar as observações em um nó m pela categoria mais frequente: 

em que pˆlm representa a proporção de indivíduos da categoria l em m, l = 1, 2, ..., r .

*Incorporando Perdas:*

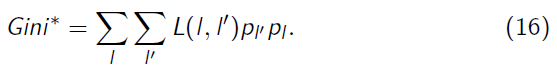
O ajuste da árvore de classificação segue os mesmos passos de uma árvore de regressão, com o ajuste de uma grande árvore, poda e seleção da árvore por validação cruzada.

Em problemas de classificação, pode ocorrer que o custo de classificação incorreta não seja o mesmo para todas as categorias da resposta.

Vamos admitir, novamente, um problema de classificação com r categorias (grupos).

Considere L (l , l j) o custo (perda) em classificar um indivíduo da categoria l na categoria l j. Obviamente, L (l , l ) = 0.

Uma maneira de incorporar os custos de má-classificação baseia-se na minimização do critério de Gini generalizado, definido por:



Para o caso de r = 2 grupos, o critério de Gini generalizado não se aplica, uma vez que o coeficiente associado a plj pl será o mesmo,

L(l , l j) + L(l j, l ): *Gini ∗ = L(1, 2)p1p2 + L(2, 1)p2p1 = [L(1, 2) + L(2, 1)]p1p2,* de forma que os custos simplesmente serão ignorados (tanto faz se L(1,2) > L(2,1) ou o contrário). Uma alternativa ao uso do crtitério de Gini generalizado, que funciona para r = 2, é incorporar pesos a priori.

*Notas:* O algoritmo tende a favorecer (proporcionar partições) covariáveis numéricas ou categóricas com grande número de categorias, uma vez

que essas oferecem maior número de partições possíveis;

Na presença de dados missing, o algoritmo usa os chamados surrogate splits (ou partições substitutas), buscando, dentre as demais covariáveis, a partição com maior nível de concordância em relação àquela para a qual não se dispõe dos dados.

Árvores de classificação e regressão são altamente instáveis.

Pequenas mudanças nos dados podem gerar ajustes consideravelmente diferentes.