

# CE092 - Extensões dos modelos de regressão

Cesar Augusto Taconeli

20 de setembro, 2017

# Árvores de classificação e regressão

# Métodos baseados em árvores

- Métodos baseados em árvores (*tree based methods*) consistem na partição do espaço das covariáveis em regiões retangulares e no ajuste de um modelo simples (como uma constante) em cada uma delas.
- Os métodos baseados em árvores compreendem uma grande variedade de algoritmos. Vamos nos concentrar, a princípio, no método CART (*Classification and regression trees*).
- 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.

# Árvores de classificação e regressão

- 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.

# Árvores de classificação e regressão

- Seja  $y$  a variável resposta e  $\mathbf{x} = (x_1, x_2, \dots, x_p)$  o vetor de covariáveis. Considere uma amostra de  $n$  observações de  $y$  e  $\mathbf{x}$ .
- O método CART inicia com a partição da amostra original em duas, segundo alguma regra do tipo

$$x_k \leq c \mid x_k > c,$$

para alguma covariável  $x_k$  numérica e  $c$  algum valor amostrado de  $x_k$ , ou

$$x_k \in A \mid x_k \notin A,$$

para uma variável  $x_k$  categórica e  $A$  uma particular categoria (ou um subconjunto de categorias) de  $x_k$ .

# Árvores de classificação e regressão

- Uma vez efetuada a partição, temos o espaço das covariáveis dividido em duas regiões,  $R_1$  e  $R_2$ .
- 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  $R_1$  e em  $R_2$ , novamente buscando a variável e respectivo ponto de corte que proporcionem melhor ajuste.

# Árvores de classificação e regressão

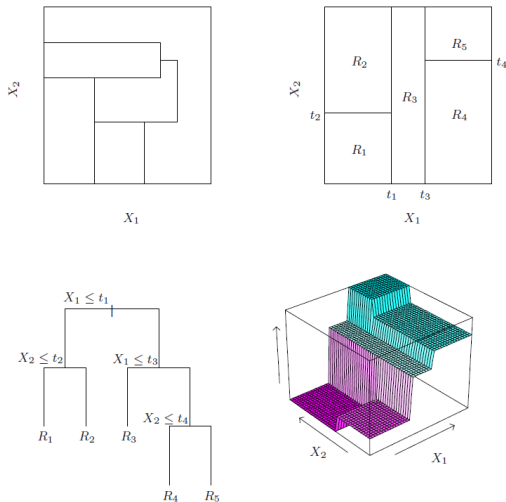
- Neste passo, temos quatro regiões delimitadas no espaço das covariáveis:  $R_3$  e  $R_4$  (formadas a partir de  $R_1$ );  $R_5$  e  $R_6$  (formadas a partir de  $R_2$ ).
- O processo é repetido sucessivamente. No final, teremos  $M$  regiões delimitadas no espaço das covariáveis, que denotaremos por  $R_1, R_2, \dots, R_M$ .

# Árvores de classificação e regressão

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



# Árvores de classificação e regressão



**Figura 1:** Ilustração - partições e árvores de regressão

# Árvores de classificação e regressão

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

$$\hat{y} = \hat{f}(\mathbf{x}) = \sum_{m=1}^M c_m I\{\mathbf{x} \in R_m\}, \quad (1)$$

sendo  $c_m$  uma constante ajustada na região  $R_m$ ,  $i = 1, 2, \dots, M$ .

# Árvores de regressão - Seleção das partições

- Para árvores de regressão, é usual considerar a soma de quadrados de resíduos como critério de minimização para a partição das amostras (nós):

$$SQR = \sum_{i=1}^n (y_i - \hat{f}(\mathbf{x}_i))^2. \quad (2)$$

- Neste caso, temos que a melhor escolha para  $c_m$  em

$$\hat{y} = \hat{f}(\mathbf{x}) = \sum_{m=1}^M c_m I\{\mathbf{x} \in R_m\}, \quad (3)$$

é simplesmente a média dos  $y_i$ 's em  $R_m$ :

$$\hat{c}_m = \frac{1}{n_m} \sum_{\mathbf{x}_i \in R_m} y_i. \quad (4)$$

# Árvores de regressão - Seleção das partições

- 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 soma de quadrados de resíduos:

$$\Delta SQR = SQR_O - \left( \frac{n_L}{n_O} SQR_L + \frac{n_R}{n_O} SQR_R \right), \quad (5)$$

sendo  $n_O$ ,  $n_L$  e  $n_R$  os números de observações nos respectivos nós.

- A partição que produzir menor valor para  $\Delta SQR$  deve ser executada.
- A regra de partição apresentada é aplicada sucessivamente aos nós originados até atingir algum critério de parada (número mínimo de observações por nó ou nos nós a serem partidos, número máximo de níveis na árvore. . . ).

# Árvores de regressão - O processo de poda

- Após obtida uma grande árvore, inicia-se o processo de poda, em que as partições são sucessivamente desfeitas até voltar à amostra original.
- O processo de poda baseia-se na seguinte função de custo-complexidade:

$$R_\alpha(T) = R(T) + \alpha|T|, \quad (6)$$

em que  $T$  representa uma árvore,  $|T|$  o número de nós finais (complexidade) e  $R(T)$  a soma de quadrados de resíduos da árvore:

$$R(T) = \sum_{m=1}^M \frac{n_m}{n} SQR_m. \quad (7)$$

# Árvores de regressão - O processo de poda

- O parâmetro  $\alpha$  na função de custo-complexidade controla a complexidade do modelo.
- Para diferentes valores de  $\alpha$  tem-se diferentes árvores minimizando  $R_\alpha(T)$ .
- Tomando  $\alpha = 0$  tem-se como solução a maior árvore disponível (não podada), uma vez que não se penaliza sua complexidade.

# Árvores de regressão - O processo de poda

- Para  $\alpha \rightarrow \infty$  tem-se penalização máxima para a complexidade e a solução é a não partição da amostra original.
- Variando  $\alpha$  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  $\alpha$  e ou  $|T|$ ).

# Árvores de regressão - Seleção do modelo

- Uma vez definida a sequência de árvores aninhadas, deve-se identificar, nessa sequência, a árvore ótima (correspondente à melhor escolha para  $\alpha$ ).
- Nesta etapa, é comum utilizar validação cruzada. Uma descrição informal da seleção por validação cruzada é descrita na sequência.



# Árvores de regressão - Seleção do modelo

- **Passo 1:** Identificação de uma sequência de valores  $\alpha_1, \alpha_2, \dots, \alpha_k$  para  $\alpha$  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)  $s/n$ :  $G_1, G_2, \dots, G_s$ ;
- **Passo 3:** Ajustar o modelo à base completa (exceto pelas observações em  $G_j$ ) e determinar  $T_1, T_2, \dots, T_k$ ;

# Árvores de regressão - Seleção do modelo

- **Passo 4:** Calcular a predição para cada observação  $i$  em  $G_j$  sob cada modelo  $T_j$ ,  $j = 1, 2, \dots, k$ ;
- **Passo 5:** Calcular a soma de quadrados dos erros de predição para o conjunto de observações em  $G_j$  :

$$\sum_{i \in G_j} (y_i - \hat{f}_{(j)}(\mathbf{x}_i))^2, \quad (8)$$

em que  $\hat{f}_{(j)}(\cdot)$  denota a predição sob o modelo ajustado sem as observações em  $G_j$ .

# Árvores de regressão - Seleção do modelo

- **Passo 6:** Os passos 3, 4 e 5 são repetidos para cada um dos demais grupos  $G_j$ . Ao término, para cada árvore  $T_1, T_2, \dots, T_k$  tem-se a respectiva soma de quadrados de predição obtida por validação cruzada:

$$SQVC = \sum_j \sum_{i \in G_j} (y_i - \hat{f}_{(j)}(\mathbf{x}_i))^2. \quad (9)$$

- 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).

# Árvores de classificação

- Como dito anteriormente, árvores 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.

# Árvores de classificação

- Vamos considerar um problema de classificação em que a resposta tenha  $r$  categorias, denotadas por  $1, 2, \dots, r$ .
- Considere uma amostra (ou um nó) e  $p_1, p_2, \dots, p_r$  as proporções com que cada categoria é observada.
- A medida de informação (ou entropia) é definida por:

$$Inf = -2 \times \sum_{l=1}^r p_l \ln(p_l) \quad (10)$$

- A medida de Gini é definida por:

$$Gini = 1 - \sum_{l=1}^r p_l^2. \quad (11)$$

# Árvores de classificação

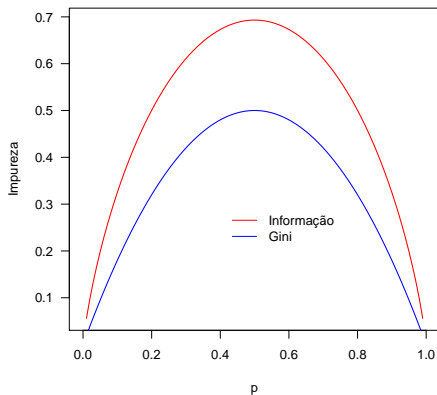
- Para o caso de duas categorias, em que as proporções de casos em cada uma delas são  $p$  e  $1 - p$ , as medidas de Informação e de Gini ficam dadas por:

$$\text{Inf} = -p \ln(p) - (1 - p) \ln(1 - p) \quad (12)$$

$$\text{Gini} = 1 - p^2 - (1 - p)^2 = 2p(1 - p). \quad (13)$$

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

# Árvores de classificação



**Figura 2:** Comparação dos critérios de informação e de Gini para  $r=2$  grupos

# Árvores de classificação

- Como pode ser observado na Figura 1, ambas as medidas são minimizadas quando os indivíduos da amostra pertencem a um mesmo grupo ( $p_l \rightarrow 1$ , para algum  $l$ ) e maximizadas quando as proporções são iguais nas diferentes categorias ( $p_1 = p_2 = \dots = p_r$ ).
- 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:

$$\Delta Imp = Imp_O - \left( \frac{n_L}{n_O} Imp_L + \frac{n_R}{n_O} Imp_R \right), \quad (14)$$

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



# Árvores de classificação

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

$$\hat{c}_m = \underset{l}{\operatorname{argmax}} \hat{p}_{lm}, \quad (15)$$

em que  $\hat{p}_{lm}$  representa a proporção de indivíduos da categoria  $l$  em  $m$ ,  $l = 1, 2, \dots, r$ .

- 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.

# Incorporando perdas

- 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  $L(l, l')$  o custo (perda) em classificar um indivíduo da categoria  $l$  na categoria  $l'$ . 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:

$$Gini^* = \sum_l \sum_{l'} L(l, l') p_{l'} p_l. \quad (16)$$

- Para o caso de  $r = 2$  grupos, o critério de Gini generalizado não se aplica, uma vez que o coeficiente associado a  $p_{l'} p_l$  será o mesmo,  $L(l, l') + L(l', l)$ :

$$Gini^* = L(1, 2)p_1 p_2 + L(2, 1)p_2 p_1 = [L(1, 2) + L(2, 1)]p_1 p_2, \quad (17)$$

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 critério de Gini generalizado, que funciona para  $r = 2$ , é incorporar pesos a priori (ver Therneau et al, 2017).

# Algumas notas sobre árvores de classificação e regressão

- 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.