Discente: Deivison Venicio Souza

Universidade Federal do Paraná - Departamento de Estatística
Especialização em Data Sciece e Big Data
Prof. Cesar Augusto Taconeli
Avaliação - 19/10/2018

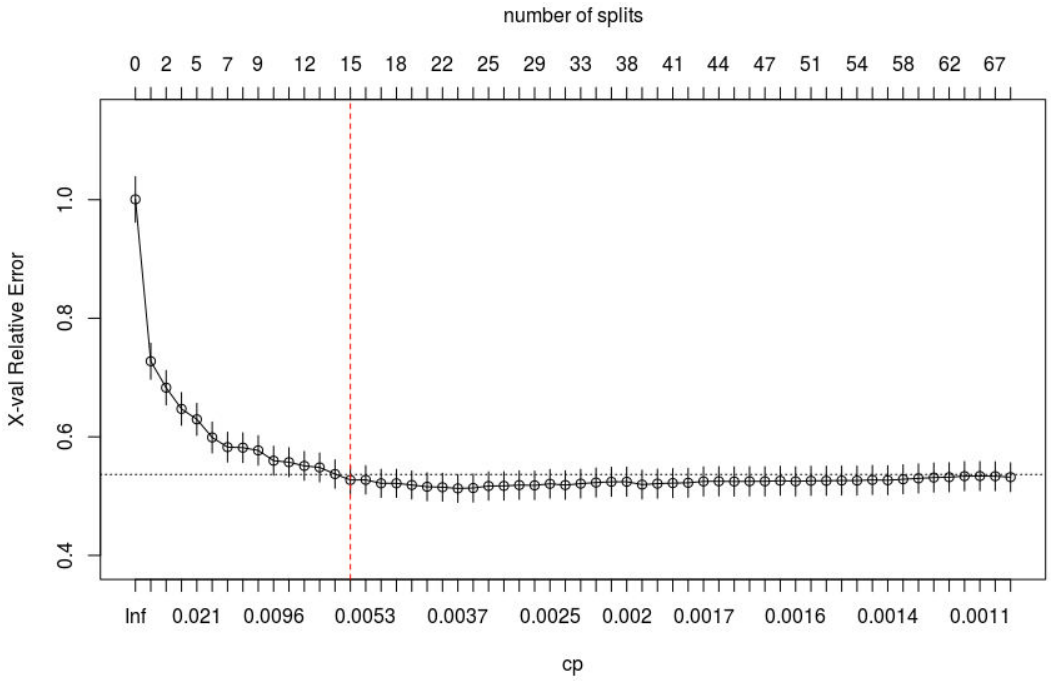
Vamos considerar a aplicação de uma árvore de regressão à base de dados abalone, do pacote **AppliedPredictiveModeling**. Os dados referem-se a 4177 espécimes de abalone, tipo de molusco encontrado ao longo das águas costeiras de todos os continentes. A variável resposta é a idade do molusco, aferida pelo número de anéis internos, que é um procedimento demorado e pouco adequado. O objetivo é ajustar um modelo que permita estimar a idade a partir de outras medidas, que são obtidas com maior facilidade. Para maiores detalhes a respeito da base, consultar a documentação e o link fornecido. Para a análise, as primeiras 3000 linhas deverão ser usadas para ajuste, e as demais para validação.

1. Qual o tamanho da árvore (número de nós finais) selecionada por validação cruzada? Quantas são as partições? Nota: Fixe a semente com set.seed(1). Estabeleça cp = 0.001 para o processo de poda.

Usando cp = 0.001 no argumento control da função rpart o tamanho da árvore (nós finais) foi construída uma árvore cujos valores de tamanho e número de divisões foram:

- Número de partições = **68 divisões.**
- Tamanho da árvore = **69 nós terminais.**

A regra 1-SE foi utilizada para podar a árvore a partir dos dados de validação cruzada. Para identificar o valor de cp indicado para podar segundo a regra 1-SE usou-se da função plotcp, cuja visualização gráfica gerada está abaixo. A linha pontilhada indica o limiar de decisão da regra 1-SE. A ideia é de que árvores menores (menos complexas) que tenham erros de validação (`xerror`) menor (ou dentro) desse limiar terão desempenho semelhante a árvore com "menor estatística desempenho" na validação cruzada (menor `xerror`), com a vantagem de ser mais parcimoniosa (mais simples). Portanto, para este caso a regra 1-SE sugere que **uma árvore com 15 divisões faz efetivamente o mesmo trabalho do que a árvore com 68 divisões** (que possui o menor xerror =



0.53199). Por fim, a árvore com 15 divisões pode ser considerado o modelo mais parcimonioso, cujo erro não é mais do que 1-SE (erro padrão) acima do erro do melhor modelo (árvore com 68 divisões).

- Número de partições após regra 1-SE = **15 divisões**
- Tamanho da árvore após regra 1-SE = **16 nós finais.**

2. Quantas covariáveis aparecem no ajuste da árvore?

Antes da poda (modelo original) as variáveis usadas para construir a árvore foram:

Diameter, Height, LongestShell, ShellWeight, ShuckedWeight , Type, VisceraWeight e WholeWeight.

Após a poda da árvore usando a regra 1-SE apenas duas variáveis permaneceram no modelo final:

ShellWeight e ShuckedWeight

3. Qual a idade estimada para moluscos com:

Para este previsão considerei o modelo de árvore podado pela regra 1-SE. Então, os resultados foram:

a) ShellWeight=0.18 e ShuckedWeight=0.25.

R: Predição de Rings = **8.506944**

b) ShellWeight=0.31 e ShuckedWeight=0.45.

R: Predição de Rings = **10.739872**

4. Qual o resíduo para cada um dos dados? Considere, para o primeiro, Rings=8 e para o segundo Rings=10.

Os resíduos são dados pela diferença entre os valores observado (reais) e estimados. Assim, para as predições da questão 3 têm-se:

- Resíduo para predição de Rings (quando ShellWeight=0.18 e ShuckedWeight=0.25) = **0.5069444**

- Resíduo para predição de Rings (quando ShellWeight=0.31 e ShuckedWeight=0.45) = **0.7398721**

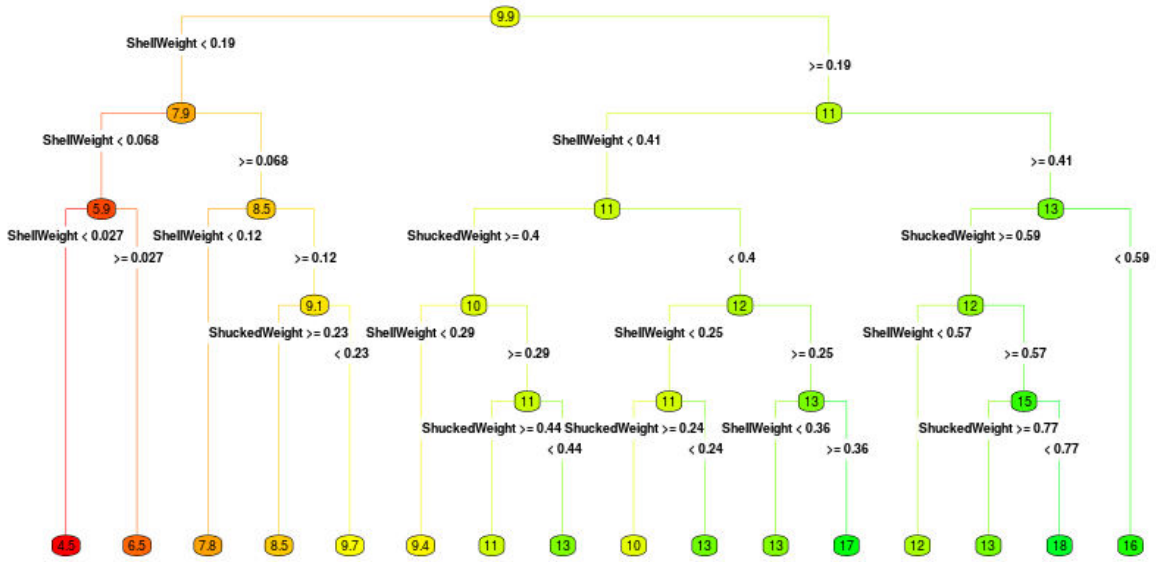
5. Usando os dados de validação, calcule e apresente o valor da soma de quadrados de resíduos.

O valor da soma de quadrados de resíduos no conjunto de validação foi de = **5768.793**

O valor de MSE no conjunto de validação foi de = **4.901268**

O valor de RMSE no conjunto de validação foi de = **2.213881**

A árvore final podada está a abaixo:



Obs: O código utilizado para solução deste trabalho está em anexo ao e-mail.
