5
Testes e Resultados

Nesse capítulo o método RCRI será testado de três formas diferentes. A primeira delas, apresentada na seção 5.1, verifica como o método se comporta quando as diferentes regiões que subdividem o domínio são limitadas por curvas polinomiais de grau maior que g. O segundo teste, seção 5.2, verifica se o método é capaz de reproduzir uma árvore caso os dados de entrada tenham sido, via simulação, gerados por ela. Para terminar, a seção 5.3 verifica seu desempenho para dados reais.

Os valores utilizados para os parâmetros do método RCRI foram:

g : Para todos os nós internos, o grau do polinômio **p** será 2.

d_{lim} : Esse parâmetro influencia diretamente no número de subdivisões do domínio. Quanto maior ele for, mais subdivisões serão possíveis. É preciso escolher d_{lim} de forma que o domínio tenha subdivisões suficientes para representar bem o conjunto de entrada, mas não sejam tantas as subdivisões a ponto de restarem poucos dados da amostra em cada uma das regiões. Então, a escolha desse parâmetro depende das características do conjunto de dados.

n_{lim} : Como já foi comentado na seção 2.3, sobre os critérios de parada, o valor de n_{lim} depende do tamanho N da amostra. Ele será escolhido de forma que os dados sejam igualmente distribuídos entre as folhas, caso a árvore seja completa. Para isso, $n_{lim} = \frac{N}{2^{d_{lim}}}$.

α : Para o teste de hipótese do 3º critério de parada, foi escolhido o valor de $\alpha = 0.05$.

p_λ : O parâmetro λ é determinado a partir do valor de p_λ da seguinte maneira: λ é tal que $100p_\lambda\%$ dos dados estão dentro da fronteira. Dessa forma, com um valor pré-estabelecido de p_λ, é possível atribuir para cada nó um valor diferente de λ, o que é uma característica interessante, uma vez que quanto mais profundo o nó menor é a sua região. Será usado $p_\lambda = 0.5$.

5.1
1ª Teste

O primeiro teste tem como objetivo verificar o comportamento do método quando os dados fornecidos vêm de uma regra que a árvore não pode realizar. Esse é o caso em que o comportamento dos dados de entrada depende de regiões no domínio definidas implicitamente por curvas de grau maior que g.

Para realizar esse teste serão feitas as seguintes etapas:

Etapa 1.1) São gerados, por simulação, a amostra inicial de tamanho $N = 1000$, composta pelos conjuntos $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{1000}\} \subset \mathbb{R}^2$ e $Y = \{y_1, y_2, \ldots, y_{1000}\} \subset \mathbb{R}$. A simulação é feita de forma que $\mathbf{x}_i$ seja uniformemente distribuído no quadrado $[-1, 1] \times [-1, 1]$ e y_i seja função de $\mathbf{x}_i$, de acordo com uma regra pré-estabelecida.

Etapa 1.2) A árvore é criada a partir dessa amostra inicial.

Etapa 1.3) É feita a análise do resultado obtido.

Etapa 1.4) Outra simulação gera mais 1000, agora para testes fora da amostra. Os conjuntos X e Y são gerados como na etapa 1.1. Para cada entrada $\mathbf{x}_i$ a árvore criada determina uma estimativa para y_i. Dessa forma é calculado o erro médio quadrático do método.

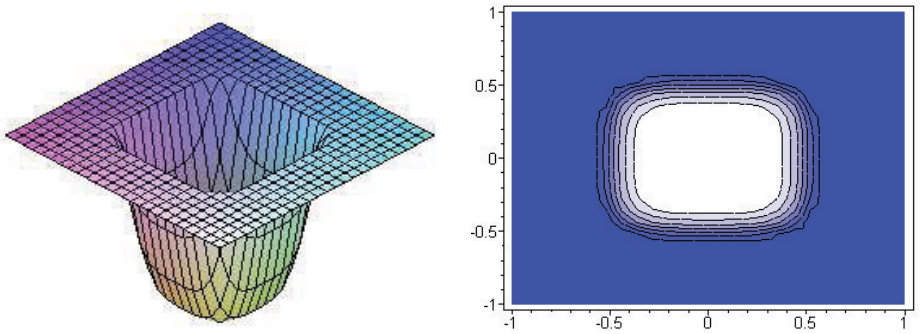
Essa sequência de etapas foi feita duas vezes, casos A e B. Em ambos os casos os dados de entrada têm dimensão 2, já que dessa forma é possível visualizar as saídas e fazer análises em cima dos gráficos gerados. Os detalhes de cada uma desses casos estão descritos a seguir.

1º Teste: Caso A

Etapa 1.1)

Os dados foram gerados de forma que $\mathbf{x}_i = (x_{i1}, x_{i2})$ seja uniformemente distribuído no quadrado $[-1, 1] \times [-1, 1]$ e $y_i = \mathbf{F}_A(\mathbf{x}_i) = \mathbf{F}_A(x_{i1}, x_{i2})$, para $1 \leq i \leq 1000$. A regra $\mathbf{F}_A$ é descrita a seguir e ilustrada nas figuras 5.1(a) e 5.1(b).

$$\mathbf{F}_A(x_1, x_2) = \begin{cases} 0,1; & \text{se } 10\,(x_1^4 + x_2^4) < 0,1 \\ 1,0; & \text{se } 10\,(x_1^4 + x_2^4) > 1,0 \\ 10\,(x_1^4 + x_2^4)\,; & \text{caso contrário} \end{cases}$$



5.1(a): Gráfico de $\mathbf{F}_A$ 5.1(b): Curvas de níveis de $\mathbf{F}_A$

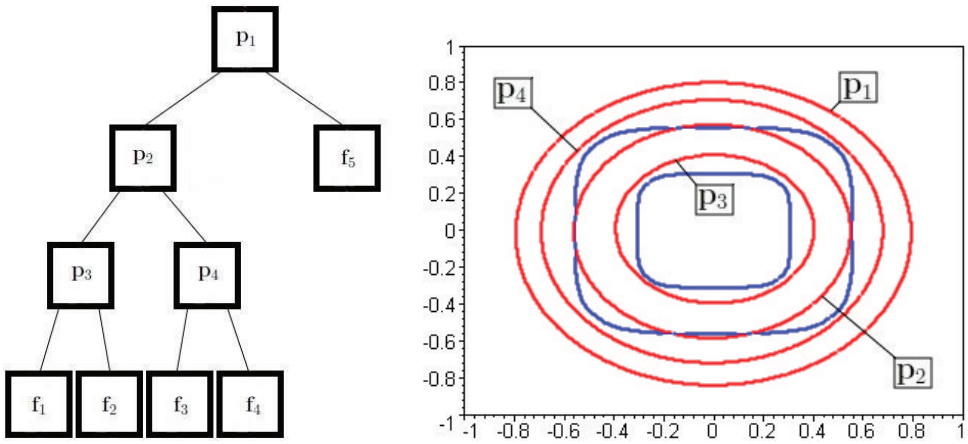
Figura 5.1: Regra de geração dos dados para o caso A do 1º teste.

Etapa 1.2)

A árvore criada possui as seguintes características:

- Número de nós internos = 4;
- Número de folhas = 5;
- Número de elementos por folha: $\{135, 136, 126, 136, 467\}$;
- Profundidade de cada folha: $\{3, 3, 3, 3, 1\}$;
- Os critérios de parada por folha: $\{1^o, 1^o, 1^o, 2^o, 2^o\}$.

A figura 5.2(a) mostra a arquitetura da árvore. Os polinômios de cada nó interno são ilustrados na figura 5.2(b), onde as curvas em vermelho, indicadas por $\mathbf{p}_i$, representam a curva de nível zero do polinômio i, e as curvas em azul representam as fronteiras das regiões definidas por $\mathbf{F}_A$, como na figura 5.1(b).



5.2(a): Árvore 5.2(b): Curvas

Figura 5.2: Resultado do caso A do 1º teste.

As regiões definidas por cada folha, que formam a partição do domínio, estão apresentadas na figura 5.3. Já os valores das regressões em cada uma

dessas regiões estão na tabela 5.1. Observe que esses valores dependem do método de estimação utilizado, local ou global.

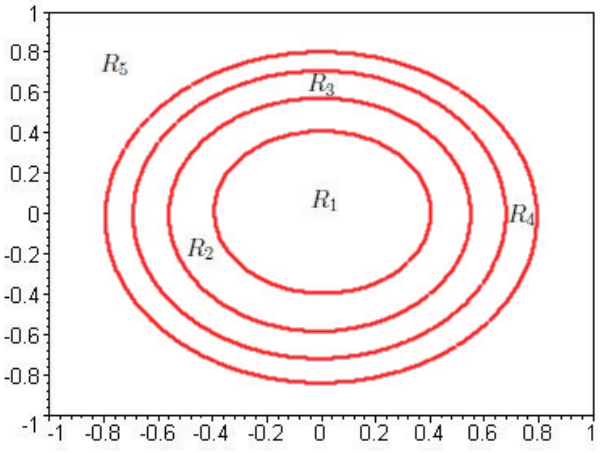


Figura 5.3: Subdivisão do domínio para o caso A do 1º teste.

Estimativa Local	Estimativa Global
$c_1 = 0,12061653772458952$	$c_1 = 0,08920717121357472$
$c_2 = 0,4926414240983919$	$c_2 = 0,37512486166175124$
$c_3 = 0,8790218565300529$	$c_3 = 1,0393158375779796$
$c_4 = 0,9999126639398305$	$c_4 = 0,9936878091284133$
$c_5 = 0,9998994590442155$	$c_5 = 1,0005008938912598$

Tabela 5.1: Valores nas folhas para o caso A do 1º teste.

Etapa 1.3)

Como as curvas em azul da figura 5.2(b) são de grau 4 e os polinômios dos nós internos são de grau 2, as curvas vermelhas jamais conseguiriam coincidir com essas curvas azuis. Por isso podemos dizer que o resultado é satisfatório, uma vez que a árvore reconheceu os padrões do domínio e se aproximou razoavelmente das curvas que definem esses padrões.

Acompanhando a construção da árvore é possível observar que primeiro é feito uma separação grosseira dos dados, definida pela curva $p_1(x_1, x_2) = 0$. Em seguida são feitas outras separações mais refinadas, no interior da região. Esse comportamento é muito satisfatório, indicando que quanto mais profundo o nó na árvore mais fino será o seu ajuste.

Os resultados das regressões nas folhas também foram bons. A região R_1 que faz o papel do conjunto $\{(x_1, x_2) \in \mathbb{R}^2 \mid \mathbf{F}_A(x_1, x_2) = 0,1\}$ ficou com estimativa próxima de $0,1$. Já as regiões R_5 e R_4 que fazem o papel do conjunto $\{(x_1, x_2) \in \mathbb{R}^2 \mid \mathbf{F}_A(x_1, x_2) = 1,0\}$ ficaram com estimativas praticamente iguais

à 1.0. As demais regiões, que representam a transição entre os valores 0.1 e 1.0, também tiveram estimativas esperadas.

Etapa 1.4)

Nesse momento 1000 novos dados, para testes fora da amostra, foram gerados de forma semelhante à anterior: $\mathbf{x}_i = (x_{i1}, x_{i2})$ uniformemente distribuído no quadrado $[-1, 1] \times [-1, 1]$ e $y_i = \mathbf{F}_A(\mathbf{x}_i)$. Para cada $\mathbf{x}_i$ gerado foi computada a estimativa fornecida pelo modelo RCRI, que será chamada de $\widehat{\mathbf{F}}(\mathbf{x}_i)$. Em seguida foram determinados os erros quadráticos,

$$\text{Erro}_i^2 = \left(\widehat{\mathbf{F}}(\mathbf{x}_i) - \mathbf{F}_A(\mathbf{x}_i)\right)^2 ,$$

tanto para a estimativa local quanto para a estimativa global. A tabela 5.2 a seguir apresenta a média, a mediana, o MAD e os valores mínimo e máximo para a amostra desses erros quadráticos calculados, onde MAD é uma medida de variação definida da seguinte maneira:

$$\text{MAD}\left\{x_i\right\}_{i=1}^n = \text{Mediana}\left\{\left|x_i - \text{Mediana}\left\{x_j\right\}_{j=1}^n\right|\right\}_{i=1}^n . \qquad (5\text{-}1)$$

RCRI	média	mediana	MAD	mínimo	máximo
Local	$5,57 \times 10^{-3}$	$1,01 \times 10^{-8}$	$1,25 \times 10^{-9}$	$8,23 \times 10^{-9}$	$9,32 \times 10^{-2}$
Global	$3,10 \times 10^{-3}$	$5,08 \times 10^{-6}$	$5,02 \times 10^{-6}$	$1,03 \times 10^{-11}$	$9,67 \times 10^{-2}$

Tabela 5.2: Estatísticas do erro quadrático para o caso A do 1º teste.

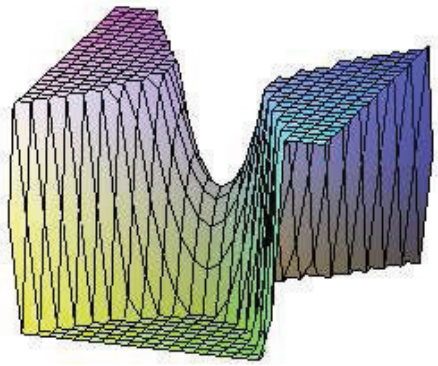
Considerando que para a função $\mathbf{F}_A$ existem duas regiões predominantes no domínio, uma com valor 0,1 e a outra com valor 1,0 , e que o erro médio quadrático, para o caso local por exemplo, é $5,57 \times 10^{-3}$, a estimativa fornecida pelo método RCRI é bastante satisfatória. Nesse caso, em média, os valores que deveriam dar 0,1 vão estar entre $0,0253$ e $0,1746$ e os valores que deveriam dar 1,0 vão estar entre $0,9253$ e $1,0746$.

1ª Teste: Caso B

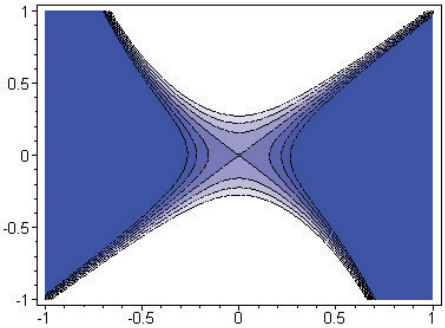
Etapa 1.1)

Como no caso anterior, os dados de entrada foram gerados de forma que $\mathbf{x}_i = (x_{i1}, x_{i2})$ seja uniformemente distribuído no quadrado $[-1, 1] \times [-1, 1]$, para $1 \leq i \leq 1000$. Nesse caso a geração de y_i obedece a regra $\mathbf{F}_B$, $y_i = \mathbf{F}_B(\mathbf{x}_i) = \mathbf{F}_B(x_{i1}, x_{i2})$, descrita a seguir e ilustrada nas figuras 5.4(a) e 5.4(b).

$$\mathbf{F}_B(x_1, x_2) = \begin{cases} -1,0; & \text{se } 10\,(x_1^4 - x_1^3 x_2 + x_1^2 - x_2^2) < -1,0 \\ 1,0; & \text{se } 10\,(x_1^4 - x_1^3 x_2 + x_1^2 - x_2^2) > 1,0 \\ 10\,(x_1^4 - x_1^3 x_2 + x_1^2 - x_2^2)\,; & \text{caso contrário} \end{cases}$$



5.4(a): Gráfico de $\mathbf{F}_B$



5.4(b): Curvas de níveis de $\mathbf{F}_B$

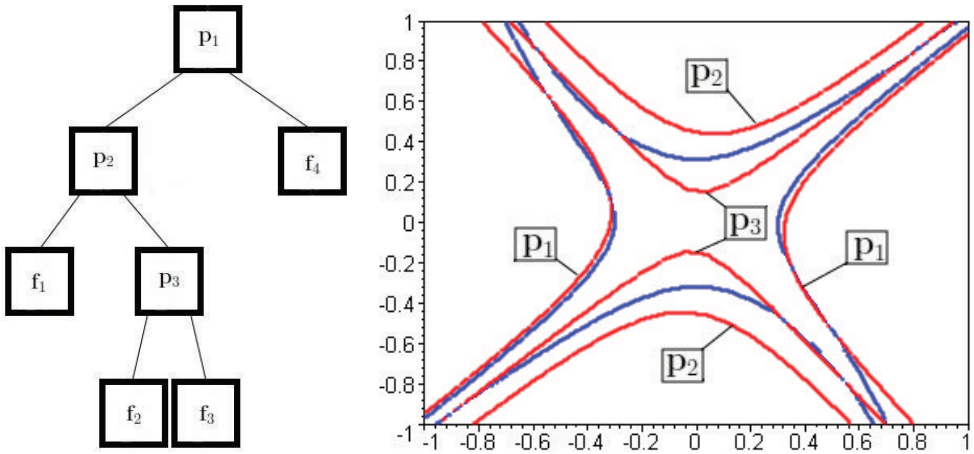
Figura 5.4: Regra de geração dos dados para o caso B do 1º teste.

Etapa 1.2)

A árvore criada possui as seguintes características:

- Número de nós internos = 3;
- Número de folhas = 4;
- Número de elementos por folha: $\{230, 145, 164, 461\}$;
- Profundidade de cada folha: $\{2, 3, 3, 1\}$;
- Os critérios de parada por folha: $\{2^o, 1^o, 1^o, 3^o\}$.

A figura 5.5(a) mostra a arquitetura da árvore. Os polinômios de cada nó interno estão ilustrados na figura 5.5(b), onde as curvas em vermelho, indicadas por $\mathbf{p}_i$, representam a curva de nível zero do polinômio i, e as curvas em azul representam as fronteiras das regiões definidas por $\mathbf{F}_B$, como na figura 5.4(b).



5.5(a): Árvore 5.5(b): Curvas

Figura 5.5: Resultado do caso B do 1º teste.

As regiões definidas por cada folha estão apresentadas na figura 5.6 e os valores das regressões em cada uma delas estão na tela 5.3.

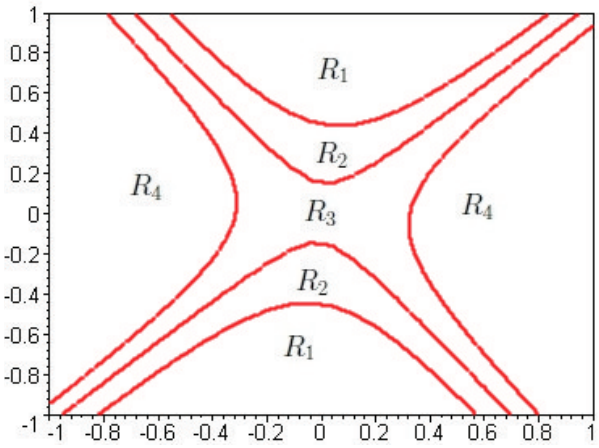


Figura 5.6: Subdivisão do domínio para o caso B do 1º teste.

Etapa 1.3)

Observando o gráfico da figura 5.5(b), a curva $\mathbf{p}_1(x_1, x_2) = 0$ representa muito bem o conjunto $\{(x_1, x_2) \in \mathbb{R}^2 \mid \mathbf{F}_B(x_1, x_2) = 1, 0\}$. Além disso, as curvas $\mathbf{p}_2(x_1, x_2) = 0$ e $\mathbf{p}_3(x_1, x_2) = 0$ conseguem reconhecer bem os pontos (x_1, x_2) tais que $\mathbf{F}_B(x_1, x_2) = -1, 0$.

Novamente pode ser observado que primeiro é feito uma separação grosseira e conforme a árvore fica mais profunda, mais fino é o ajuste. Como no caso anterior, os resultados das regressões nas folhas também foram bons.

Estimativa Local	Estimativa Global
$c_1 = -0,998844245167958$	$c_1 = -0,9895313831161557$
$c_2 = -0,8642838325577452$	$c_2 = -1,0827358162558174$
$c_3 = 0,33959199102208903$	$c_3 = 0,17383848266153643$
$c_4 = 0,9396077707797265$	$c_4 = 1,0573804421958635$

Tabela 5.3: Valores nas folhas para o caso B do 1º teste.

Etapa 1.4)

Nesse momento 1000 novos dados para testes fora da amostra foram gerados de forma semelhante à anterior: $\mathbf{x} = (x_1, x_2)$ uniformemente distribuído no quadrado $[-1, 1] \times [-1, 1]$ e $y = \mathbf{F}_B(x_1, x_2)$. Como no caso A, foram computados os erros quadráticos para cada $\mathbf{x}_i$ gerado, tanto para a estimativa local quanto para a estimativa global. A tabela 5.4 a seguir apresenta a média, a mediana, o MAD e os valores mínimo e máximo para a amostra desses erros quadrados calculados.

RCRI	média	mediana	MAD	mínimo	máximo
Local	$4,69 \times 10^{-2}$	$3,65 \times 10^{-3}$	$3,65 \times 10^{-3}$	$1,34 \times 10^{-6}$	$1,63$
Global	$3,66 \times 10^{-2}$	$3,29 \times 10^{-3}$	$3,18 \times 10^{-3}$	$5,01 \times 10^{-8}$	$1,30$

Tabela 5.4: Estatísticas do erro quadrático para o caso B do 1º teste.

Considerando que para a função $\mathbf{F}_B$ existem duas regiões predominantes no domínio, uma com valor $-1,0$ e a outra com valor $1,0$, e que o erro médio quadrático, para o caso local por exemplo, é $4,69 \times 10^{-2}$, a estimativa fornecida pelo método RCRI é bastante satisfatória. Nesse caso, em média, os valores que deveriam dar $-1,0$ vão estar entre $-1,2165$ e $-0,7834$ e os valores que deveriam dar 1,0 vão estar entre $0,7834$ e $1,2165$.

5.2
2º Teste

Esse segundo teste busca verificar se o método é capaz de reproduzir uma árvore já existente quando ele é construído a partir de dados gerados, via simulação, pela própria árvore. A seguir, as etapas para a execução desse teste.

Etapa 2.1) São gerados, por simulação, a amostra inicial de tamanho $N = 1000$, composta pelos conjuntos $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{1000}\} \subset \mathbb{R}^2$ e $Y = \{y_1, y_2, \ldots, y_{1000}\} \subset \mathbb{R}$. A simulação é feita de forma que $\mathbf{x}_i$ seja uniformemente distribuído no quadrado $[-1, 1] \times [-1, 1]$ e y_i seja a saída de uma árvore já existente, chamada de "árvore 1", para o dado de entrada $\mathbf{x}_i$.

Etapa 2.2) Uma segunda árvore, chamada de "árvore 2", é criada a partir dessa amostra inicial de dados.

Etapa 2.3) É feita a análise do resultado obtido.

Etapa 2.4) Outra simulação gera mais 1000 dados, agora para testes fora da amostra. O conjunto X, de dados de entrada, é gerado como na etapa 2.1: $\mathbf{x}_i$ uniformemente distribuído no quadrado $[-1,1] \times [-1,1]$. Nessa etapa são gerados dois conjuntos de dados de saída. O primeiro com as estimativas fornecidas pela árvore 1 e o segundo com as estimatívas fornecidas pela árvore 2, ambos supondo que os dados de entrada são os elementos de X. Dessa forma é calculado a diferença quadrática entre as duas estimativas e feita as devidas comparações.

Como nesse teste o importante é comparar a estrutura das duas árvores, tanto faz o método para estimar a regressão nas folhas, local ou global. Por isso, apenas a forma local será usada.

Da mesma forma que no teste 1, a sequência de etapas acima foi feita duas vezes, casos A e B, com dados de entrada de dimensão 2.

2º Teste: Caso A

Etapa 2.1)

O formato da árvore 1 está exposta na figura 5.7. Suas curvas de nível zero dos polinômios nos nós internos estão apresentados na figura 5.8(a) e sua partição do domínio pode ser vista na figura 5.9(a). Já os valores das regressões em cada região aparecem na primeira coluna da tabela 5.5.

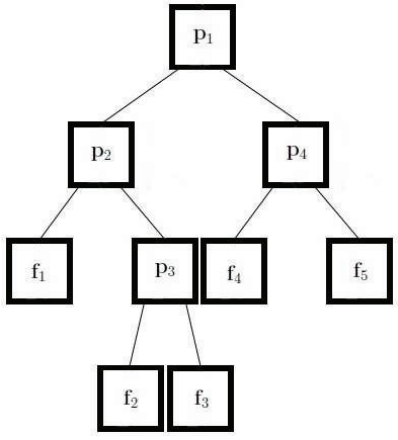
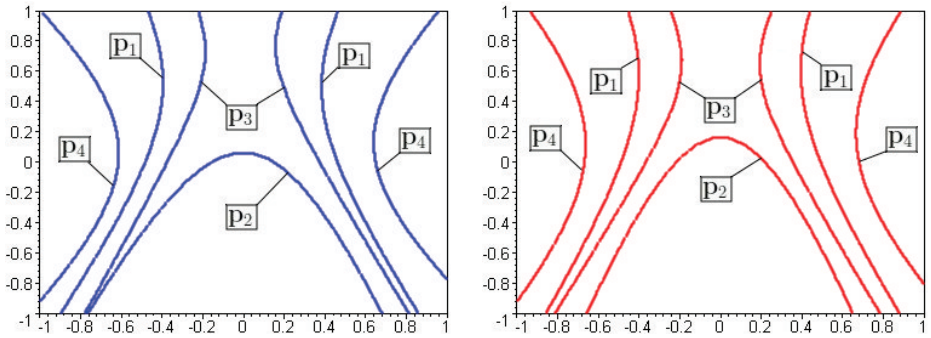


Figura 5.7: Arquitetura das árvores para o caso A do 2º teste.

Foram gerados 1000 dados $\mathbf{x}_i \in \mathbb{R}^2$, uniformemente distribuídos no quadrado $[-1,1] \times [-1,1]$. Para cada um deles foi computado a estimativa fornecida pela árvore 1, chamadas de $\widehat{\mathbf{F}}_1(\mathbf{x}_i) = y_i$. Dessa forma foi criada a amostra de dados $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{1000}\}$ e $Y = \{y_1, y_2, \ldots, y_{1000}\}$.

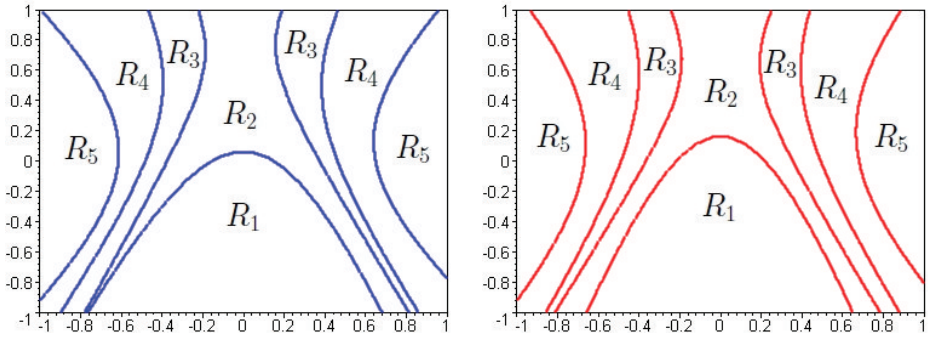
Etapa 2.2)

A partir da amostra de dados formada pelos conjuntos X e Y, construída na etapa anterior, foi criada a árvore 2. Seu formato é o mesmo que da árvore 1, isto é, o formato apresentado na figura 5.7. As curvas de nível zero de seus polinômios estão apresentados na figura 5.8(b) e as regiões definidas por cada folha na figura 5.9(b). Os valores das regressões em cada região aparecem na segunda coluna da tabela 5.5.



5.8(a): Árvore 1 5.8(b): Árvore 2

Figura 5.8: Curvas de nível zero dos polinômios do caso A do 2º teste.



5.9(a): Árvore 1 5.9(b): Árvore 2

Figura 5.9: Subdivisão do domínio para o caso A do 2º teste.

Árvore 1	Árvore 2
$c_1 = 0,02608871874592359$	$c_1 = 0,0889158297577668$
$c_2 = 0,5040017690666077$	$c_2 = 0,49701306011303975$
$c_3 = 1,0365876047995826$	$c_3 = 1,162378619041381$
$c_4 = 1,5359452607389446$	$c_4 = 1,4890861118529133$
$c_5 = 1,9990462601784886$	$c_5 = 1,9423833830510138$

Tabela 5.5: Valores nas folhas para o caso A do 2º teste.

Etapa 2.3)

A semelhança entre as duas árvores é evidente. Elas não só tem a mesma arquitetura como também definem partições do domínio visualmente muito parecidas. Além disso, os valores das regressões em cada região são bem parecidos. Com isso concluí-se que o método foi capaz de reproduzir a árvore 1.

Etapa 2.4)

Nesse momento 1000 novos dados de entrada, para testes fora da amostra, foram gerados de forma semelhante à anterior: $\mathbf{x}_i$ uniformemente distribuído no quadrado $[-1,1]\times[-1,1]$. Para cada $\mathbf{x}_i$ gerado foi computada a estimativa fornecida pela árvore 1, chamadas de $\widehat{\mathbf{F}}_1(\mathbf{x}_i)$, e pela árvore 2, chamadas de $\widehat{\mathbf{F}}_2(\mathbf{x}_i)$. Em seguida foram determinados os erros quadráticos,

$$\text{Erro}_i^2 = \left(\widehat{\mathbf{F}}_1(\mathbf{x}_i) - \widehat{\mathbf{F}}_2(\mathbf{x}_i)\right)^2 .$$

A tabela 5.6 a seguir apresenta a média, a mediana, o MAD e os valores mínimo e máximo, para a amostra dos erros quadrados calculados.

média	mediana	MAD	mínimo	máximo
$1,22\times10^{-3}$	$2,83\times10^{-4}$	$5,56\times10^{-4}$	$5,42\times10^{-14}$	$2,15\times10^{-2}$

Tabela 5.6: Estatísticas para os erro quadrático entre as saídas das árvores 1 e 2 do caso A do 2º teste.

Os baixos valores na tabela 5.6 acima mostram, mais uma vez, o desempenho satisfatório do método RCRI nesse teste, ratificando assim a conclusão de que o método foi capaz de reproduzir bem a árvore 1.

2º Teste: Caso B

Etapa 2.1)

O formato da árvore 1 está exposta na figura 5.10. Suas curvas de nível zero dos polinômios nos nós internos estão apresentados na figura 5.11(a) e sua partição do domínio pode ser vista na figura 5.12(a). Já os valores das regressões em cada região aparecem na primeira coluna da tabela 5.7.

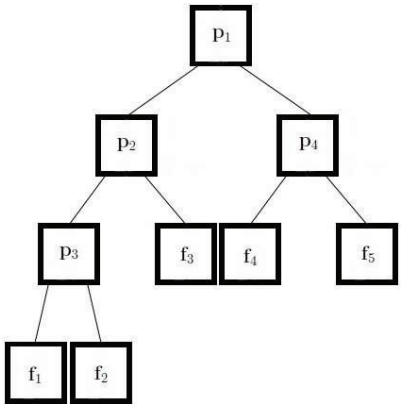


Figura 5.10: Arquitetura das árvores para o caso B da 2º teste.

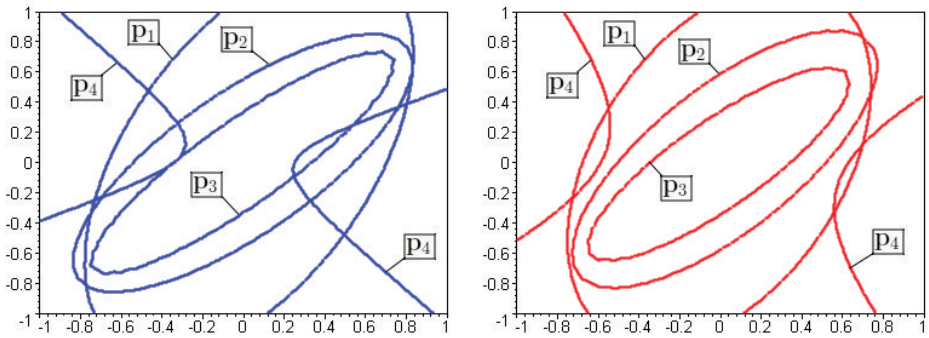
Foram gerados 1000 dados $\mathbf{x}_i \in \mathbb{R}^2$, uniformemente distribuído no quadrado $[-1,1] \times [-1,1]$. Para cada um deles foi computado a estimativa fornecida pela árvore 1, chamadas de $\widehat{\mathbf{F}}_1(\mathbf{x}_i) = y_i$. Dessa forma foi criada a amostra de dados $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{1000}\}$ e $Y = \{y_1, y_2, \ldots, y_{1000}\}$.

Etapa 2.2)

A partir da amostra de dados formada pelos conjuntos X e Y, construída na etapa anterior, foi criada a árvore 2. Seu formato é o mesmo que da árvore 1, isto é, o formato apresentado na figura 5.10. As curvas de nível zero de seus polinômios estão apresentados na figura 5.11(b) e as regiões definidas por cada folha na figura 5.12(b). Os valores das regressões em cada região aparecem na segunda coluna da tabela 5.7.

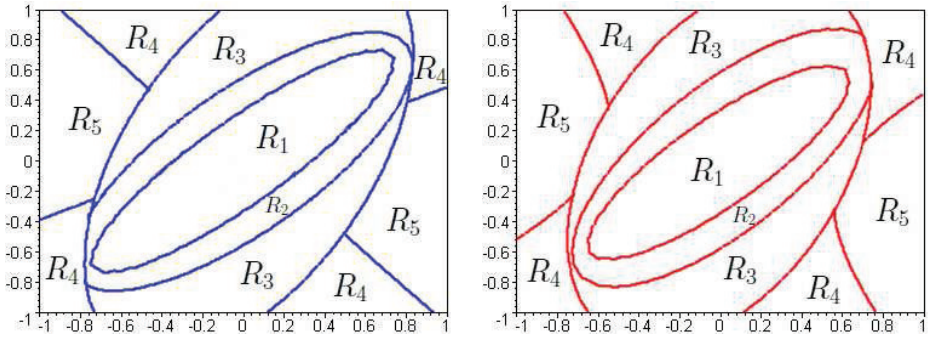
Árvore 1	Árvore 2
$c_1 = -0,3318529586419327$	$c_1 = -0,3046679941847812$
$c_2 = -0,028627739424873528$	$c_2 = 0,06089862704721668$
$c_3 = 0,6665116254993362$	$c_3 = 0,626162432991329$
$c_4 = 0,8334923161675619$	$c_4 = 0,7672438096727838$
$c_5 = 1,296923571198397$	$c_5 = 1,2215662097768125$

Tabela 5.7: Valores nas folhas para o caso B do 2º teste.



5.11(a): Árvore 1 5.11(b): Árvore 2

Figura 5.11: Curvas de nível zero dos polinômios do caso B do 2º teste.



5.12(a): Árvore 1 5.12(b): Árvore 2

Figura 5.12: Subdivisão do domínio para o caso B do 2º teste.

Etapa 2.3)

Novamente, a semelhança entre as duas árvores é evidente. Elas não só tem a mesma arquitetura como também definem partições do domínio visualmente muito parecidas.

Além disso, os valores das regressões nas regiões que definem a partição do domínio são bem parecidos. Com isso concluí-se que o método foi capaz de reproduzir, também para o caso B, a árvore 1.

Etapa 2.4)

Nesse momento 1000 novos dados de entrada, para testes fora da amostra, foram gerados de forma semelhante à anterior: $\mathbf{x}_i$ uniformemente distribuído no quadrado $[-1, 1] \times [-1, 1]$. Para cada $\mathbf{x}_i$ gerado foi computada a estimativa fornecida pela árvore 1, $\widehat{\mathbf{F}}_1(\mathbf{x}_i)$, pela árvore 2, $\widehat{\mathbf{F}}_2(\mathbf{x}_i)$, e em seguida determinados os erros quadráticos, como no caso A. As estatísticas para esses erros quadráticos estão na tabela 5.8 a seguir:

média	mediana	MAD	mínimo	máximo
$1,11 \times 10^{-2}$	$4,39 \times 10^{-3}$	$3,47 \times 10^{-3}$	$3,12 \times 10^{-11}$	$1,84 \times 10^{-1}$

Tabela 5.8: Estatísticas para os erro quadrático entre as saídas das árvores 1 e 2 do caso B do 2º teste.

Os baixos valores na tabela 5.8 acima mostram, também para esse caso, o desempenho satisfatório do método RCRI nesse teste.

5.3 Testes com Dados Reais

Nessa seção o objetivo é, além de verificar o desempenho do método RCRI em dados reais, compará-lo com outros modelos. Os dados rodados estão listados a seguir. O número de atributo é o número de coordenadas do vetor $\mathbf{x}$, ou seja, o número de coordenadas dos vetores $\mathbf{x}_i$.

1. *Abalone*: Conjunto de dados com 4177 linhas. Cada linha possui um valor a ser estimado e mais 8 atributos.

2. *Housing*: Conjunto de dados com 506 linhas. Cada linha possui um valor a ser estimado e mais 13 atributos.

3. *Computer Hardware*: Conjunto de dados com 209 linhas. Cada linha possui um valor a ser estimado e mais 6 atributos.

Esses conjuntos de dados também foram rodados para os modelos CART (5), MARS (17) e STR-Tree (11). Tais resultados encontram-se no trabalho de da Rosa, Veiga e Medeiros (11). Além disso, esses dados podem ser encontrados na *UC Irvine Machine Learning Repository*, no endereço `http://archive.ics.uci.edu/ml/datasets.html`.

Resultados

Como deseja-se comparar os resultados, a forma de rodar os dados foi semelhante a usada no trabalho de da Rosa, Veiga e Medeiros (11). Para garantir um resultado justo foi realizada a sequência de passos listadas abaixo. Os resultados estão expostos na tabela 5.9.

- A amostra foi dividida, de forma aleatória, em 10 partes iguais.
- O método foi rodado 10 vezes. Para cada uma das vezes os dados de uma das partes foram usados como dados fora da amostra. Os demais, das outras 9 partes, foram usadas como dados dentro da amostra.

– Em cada uma das vezes que o método foi rodado, o erro médio quadrático da estimativa foi computado. Assim foi formado um conjunto de 10 erros médios quadráticos, uma para cada rodada.

– A comparação entre os modelos será baseada nas seguintes estatísticas para esse conjunto de erros: mínimo, máximo, mediana e MAD. Onde MAD é uma medida de variação definida na equação (5-1).

	RCRI (local)	RCRI (global)	STR-Tree (LM)	SRT-Tree (CV)	CART	MARS
Abalone:						
Mediana	**4.38**	4.53	5.23	6.26	5.93	4.50
MAD	0.49	0.52	0.55	0.63	0.45	0.40
Mínimo	3.66	3.72	3.85	4.21	4.54	3.62
Máximo	5.73	5.87	7.79	8.38	8.20	5.84
Housing:						
Mediana	19.64	17.48	13.91	12.06	20.49	**11.71**
MAD	3.48	3.39	3.34	2.96	6.99	2.91
Mínimo	10.11	9.54	5.64	6.49	8.35	6.30
Máximo	30.55	31.89	41.03	43.32	51.78	39.05
Computer Hardware: ($\times 10^3$)						
Mediana	2.77	3.63	2.56	3.05	5.56	**2.34**
MAD	1.62	1.84	1.33	1.94	4.10	1.27
Mínimo	0.45	0.48	0.55	0.28	0.47	0.51
Máximo	14.3	12.3	18.1	26.7	43.1	14.3

Tabela 5.9: Resultado para os dados *Abalone*, *Housing* e *Computer Hardware*.

De forma geral o resultado foi muito bom. No conjunto de dados *Abalone* o método RCRI teve o melhor desempenho, para os dados de *Computer Hardware* foi um dos melhores e, apenas para os dados *Housing*, não teve uma saída tão boa. Isso mostra a competitividade do RCRI.