



Jessica Quintanilha Kubrusly

**Regressão Construtiva por Regiões Definidas
Implicitamente**

Tese de Doutorado

Tese apresentada ao Programa de Pós-graduação em Matemática do Departamento de Matemática da PUC-Rio como requisito parcial para obtenção do título de Doutor em Matemática

Orientador: Prof. Hélio Côrtes Vieira Lopes

Rio de Janeiro
junho de 2009



Jessica Quintanilha Kubrusly

Regressão Construtiva por Regiões Definidas Implicitamente

Tese apresentada ao Programa de Pós-graduação em Matemática do Departamento de Matemática do Centro Técnico Científico da PUC-Rio como requisito parcial para obtenção do título de Doutor em Matemática. Aprovada pela comissão examinadora abaixo assinada.

Prof. Hélio Côrtes Vieira Lopes

Orientador

Departamento de Matemática — PUC-Rio

Prof. Beatriz Vaz de Melo Mendes

Departamento de Métodos Estatísticos do Instituto de
Matemática — UFRJ

Prof. Dirce Uesu Pesco

Departamento de Geometria — UFF

Prof. Luiz Henrique de Figueiredo

IMPA

Prof. Cristiano Fernandes

Departamento de Engenharia Elétrica — PUC-Rio

Prof. Sinesio Pesco

Departamento de Matemática — PUC-Rio

Prof. Thomas Lewiner

Departamento de Matemática — PUC-Rio

Prof. José Eugênio Leal

Coordenador Setorial do Centro Técnico Científico — PUC-Rio

Rio de Janeiro, 22 de junho de 2009

Todos os direitos reservados. Proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador.

Jessica Quintanilha Kubrusly

No ano de 2003 graduou-se em Engenharia Elétrica pela Pontifícia Universidade Católica do Rio de Janeiro, com ênfase em Sistemas de Apoio à Decisão. Em 2005 concluiu seu mestrado em Matemática Aplicada, também pela Pontifícia Universidade Católica do Rio de Janeiro, onde desenvolveu junto com seus orientadores uma ferramenta estatística para o cálculo de reservas.

Ficha Catalográfica

Kubrusly, Jessica

Regressão Construtiva por Regiões Definidas Implicitamente / Jessica Quintanilha Kubrusly; orientador: Hélio Côrtes Vieira Lopes. — Rio de Janeiro : PUC–Rio, Departamento de Matemática, 2009.

v., 89 f: il. ; 29,7 cm

1. Tese (doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Matemática.

Inclui referências bibliográficas.

1. Matemática – Tese. 2. Árvores de Regressão. 3. Estatística. 4. Modelos não Lineares. 5. Modelos não Paramétricos. 6. IBNR. I. Lopes, Hélio Côrtes Vieira. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Matemática. III. Título.

CDD: 510

Agradecimentos

Agradeço à Funenseg, pela bolsa concedida, à PUC-Rio, pelo suporte em todos os aspectos e ao professor Luiz Roberto Cunha, pela confiança e incentivo durante esse período. Sem dúvida foram ajudas fundamentais para que esse trabalho fosse concretizado.

Gostaria muito de agradecer ao professor Hélio pela sua dedicação, não só nesses últimos quatro anos de doutorado, mas também por todos os oito anos de convivência. Sua simpatia e profissionalismo contribuíram muito para que esta tese acontecesse e para que a relação entre aluno e orientador fosse a melhor possível.

Não poderia deixar de agradecer a todos da PUC-Rio, funcionários, alunos e professores, que fizeram parte desse agradável ambiente acadêmico e por isso são totais responsáveis pela enorme saudade que sinto desde já.

Por último, quero agradecer aos familiares e amigos, que nesses últimos meses foram muito compreensíveis com a minha constante ansiedade e falta de tempo por causa da tese.

Resumo

Kubrusly, Jessica; Lopes, Hélio Côrtes Vieira. **Regressão Construtiva por Regiões Definidas Implicitamente**. Rio de Janeiro, 2009. 89p. Tese de Doutorado — Departamento de Matemática, Pontifícia Universidade Católica do Rio de Janeiro.

Os métodos de regressão baseados em árvores são modelos não lineares e não paramétricos, estudados desde a década de 80, quando houve a criação do algoritmo CART. Até hoje há muita pesquisa nessa área e cada vez mais novos métodos são apresentados com o objetivo de aperfeiçoar os modelos já existentes. Esse trabalho propõe um novo método chamado de Regressão Construtiva em Regiões Implícitas (RCRI). Sua principal diferença, com relação aos demais métodos baseados em árvores, está na forma como o domínio é particionado. Até o momento essa partição era formada por retângulos com arestas paralelas aos eixos, porém o algoritmo RCRI permitiu que as partições fossem formadas por regiões mais flexíveis definidas implicitamente. Além disso, o trabalho também propõe uma extensão intervalar para o modelo. Duas diferentes aplicações desse novo método também são sugeridas. A primeira em atuária, que busca melhorar a estimativa da reserva IBNR fornecida pelo já usual modelo *Chain Ladder*. A segunda em geologia, que utiliza as informações existentes nos poços para realizar inferências sobre dados faltantes.

Palavras-chave

Árvores de Regressão. Estatística. Modelos não Lineares. Modelos não Paramétricos. IBNR.

Abstract

Kubrusly, Jessica; Lopes, Hélio Côrtes Vieira (Adviser). **Constructive Regression on Implicitly Defined Regions**. Rio de Janeiro, 2009. 89p. D.Sc. Thesis — Departamento de Matemática, Pontifícia Universidade Católica do Rio de Janeiro.

Tree-based methods are playing an important role in non-linear and non-parametric regression. They have been studied since the 80's, when the CART algorithm was proposed. Nowadays there is a lot of research in this area and new methods are being created, aiming at improving existing models. This work proposes a new tree-based method called Constructive Regression on Implicit Regions. Its main difference, with respect to other tree-based methods, is how the domain is partitioned. The proposed algorithm allows partitions formed by flexible regions whose borders are implicitly defined. Moreover, the work also proposes an interval extension to the model. Two different applications of this new method are also proposed. The first one is in actuary, which looks for improvements in the estimation of IBNR reserves, already provided by the usual Chain Ladder model. The second one is in geology, which uses the well data to perform inferences about the missing data in the well itself.

Keywords

Regression Trees. Statistics. Non Linear Models. Non Parametric Models. IBNR.

Sumário

1	Introdução	11
2	Método de Regressão Construtiva em Regiões Implícitas	17
2.1	Partição do Domínio e Caracterização da Árvore	17
2.2	Construção do Polinômio p	18
2.3	Critérios de Parada	22
3	Suavidade em Árvores de Regressão	24
3.1	Funções de Transição	24
3.2	Aplicação em Árvores	28
3.3	Regressão Construtiva nas Regiões Implícitas	30
4	Algoritmo e Comentários sobre o Método	35
4.1	Algoritmo: Regressão Construtiva por Regiões Implícitas	35
4.2	Alguns Comentários sobre o Método	36
5	Testes e Resultados	38
5.1	1ª Teste	39
5.2	2º Teste	45
5.3	Testes com Dados Reais	51
6	Aplicações	53
6.1	Cálculo de Reservas de Sinistros Ocorridos e Ainda não Avisados	53
6.2	Geologia de Petróleo	61
7	Extensão Intervalar	65
7.1	O Que Muda na Estrutura?	66
7.2	Algoritmo Intervalar	69
7.3	Aplicação	71
8	Conclusão	74
	Referências Bibliográficas	76
A	Estimadores por Mínimos Quadrados para Regressões Lineares	81
A.1	Condição para Existência	82
A.2	Estimador não Tendencioso	82
A.3	Condição para Consistência	82
B	SVR - <i>Support Vector Regression</i>	84
B.1	Problema Primal	84
B.2	Problema Dual	85
B.3	Núcleo	85
B.4	Solução	85

C	Aritmética Intervalar	87
C.1	Intervalos	87
C.2	Operações Elementares	88
C.3	Funções Intervalares	88

Lista de figuras

1.1	Métodos Baseados em Árvores	12
2.1	Exemplo da árvore para o método RCRI.	18
2.2	Exemplo de partição do domínio para uma dimensão.	19
3.1	Valores da função de transição por regiões.	24
3.2	Diagrama das regiões definidas pelos conjuntos A e B .	25
3.3	Gráficos dos exemplos citados de função degrau.	27
3.4	Criação de cada nó interno da árvore.	28
3.5	Exemplo da aplicação de suavidade em árvores de regressão.	29
3.6	Ilustração do incremento no número de folhas.	30
3.7	Exemplo das regressões local e global para uma dimensão.	33
4.1	Ilustração para a avaliação do método RCRI em um ponto qualquer.	37
5.1	Regra de geração dos dados para o caso A do 1º teste.	40
5.2	Resultado do caso A do 1º teste.	40
5.3	Subdivisão do domínio para o caso A do 1º teste.	41
5.4	Regra de geração dos dados para o caso B do 1º teste.	43
5.5	Resultado do caso B do 1º teste.	44
5.6	Subdivisão do domínio para o caso B do 1º teste.	44
5.7	Arquitetura das árvores para o caso A do 2º teste.	46
5.8	Curvas de nível zero dos polinômios do caso A do 2º teste.	47
5.9	Subdivisão do domínio para o caso A do 2º teste.	47
5.10	Arquitetura das árvores para o caso B do 2º teste.	49
5.11	Curvas de nível zero dos polinômios do caso B do 2º teste.	50
5.12	Subdivisão do domínio para o caso B do 2º teste.	50
6.1	Triângulo de desenvolvimento para calcular o IBNR.	54
6.2	Criação dos dados de entrada para a aplicação em IBNR.	55
6.3	Dados utilizados na estimativa do IBNR.	57
6.4	Resultado para dados TRV.	60
6.5	Relação entre ρ_{hob} e os diferentes solos.	62
6.6	Exemplo de uma amostra de ρ_{hob} .	62
7.1	Exemplo intervalar em uma variável.	69

Lista de tabelas

5.1	Valores nas folhas para o caso A do 1º teste.	41
5.2	Estatísticas do erro quadrático para o caso A do 1º teste.	42
5.3	Valores nas folhas para o caso B do 1º teste.	45
5.4	Estatísticas do erro quadrático para o caso B do 1º teste.	45
5.5	Valores nas folhas para o caso A do 2º teste.	48
5.6	Estatísticas para os erro quadrático entre as saídas das árvores 1 e 2 do caso A do 2º teste.	48
5.7	Valores nas folhas para o caso B do 2º teste.	49
5.8	Estatísticas para os erro quadrático entre as saídas das árvores 1 e 2 do caso B do 2º teste.	51
5.9	Resultado para os dados <i>Abalone</i> , <i>Housing</i> e <i>Computer Hardware</i> .	52
6.1	Valores das saídas para os dados ABC e RAA.	58
6.2	Valores das saídas para os dados TRV.	59
6.3	Estatísticas para os erros quadráticos do IBNR.	60
6.4	Estatísticas, dos quatro poços, para os erros médios quadráticos nas estimativas de rhob.	64
7.1	Estatísticas, dos quatro poços, para a média dos tamanhos dos intervalos gerados.	72
7.2	Redução no tamanho do intervalo fornecido, por cada método e em cada um dos quatro poços.	72
7.3	Estatísticas, dos quatro poços, para as porcentagens de acerto.	73