Pontifícia Universidade Católica
do Rio de Janeiro

Rodrigo Pinto Moreira

Modelo de Regressão Logística com Transição Suave Estruturado por Árvore (STLR-Tree)

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio.

Orientador: Prof. Dr. Álvaro Veiga

Rio de Janeiro
Abril de 2008

Pontifícia Universidade Católica
do Rio de Janeiro

Rodrigo Pinto Moreira

Modelo de Regressão Logística com Transição Suave Estruturado por Árvore (STLR-Tree)

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica do Centro Técnico Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Álvaro de Lima Veiga Filho
Orientador
Departamento de Engenharia Elétrica – PUC-Rio

Prof. Marcelo Cunha Medeiros
Departamento de Economia – PUC-Rio

Prof. Joel Maurício Corrêa da Rosa
UFPR

Prof. José Eugenio Leal
Coordenador Setorial do Centro
Técnico Científico

Rio de Janeiro, 11 de abril de 2008

Rodrigo Pinto Moreira

Graduou-se em Ciência Estatísticas na Escola Nacional de Ciências Estatísticas - ENCE (Rio de Janeiro, Brasil). Durante o mestrado em Engenharia Elétrica trabalhou com técnicas de análise estatística multivariada, modelagem linear e não-linear e em projetos na área de seguros colaborando com seu orientador no desenvolvimento de modelos internos para seguradoras.

Ficha Catalográfica

Moreira, Rodrigo Pinto

Modelo de regressão logística com transição suave estruturado por árvore (STLR-Tree) / Rodrigo Pinto Moreira ; orientador: Álvaro de Lima Veiga Filho. – 2008.

81 f. : il. ; 30 cm

Dissertação (Mestrado em Engenharia Elétrica)– Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2008.

Inclui bibliografia

1. Engenharia elétrica – Teses. 2. Modelos não-lineares estruturados por árvore. 3. Classificação. 4. Regressão logística. Árvore de classificação e regressão (CART). I. Veiga Filho, Álvaro de Lima. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

CDD: 621.3

Agradecimentos

Ao meu orientador Álvaro Veiga Lima Filho, pelo apoio e incentivo a este trabalho.

À FAPERJ, CNPq, CAPES e à PUC-Rio, pelos auxílios financeiros concedidos, sem os quais este trabalho não poderia ter sido realizado.

Aos meus pais e minha irmã. Família que me atura e ajuda diariamente.

A todos os meus outros familiares, principalmente minhas avós e meu avô, que por muitos anos ainda me darão força.

Ao meu avô João (*in memoriam*), que certamente está me ajudando de um bom lugar.

À minha futura esposa, Suene, e à sua família. Ela foi a pessoa que mais me aturou no decorrer deste trabalho.

Aos meus queridos amigos da TDP, ENCE e todos os demais.

Aos meus companheiros da PUC-Rio, principalmente aos freqüentadores da sala L604, na favelinha.

Aos professores da ENCE, Kaizô Beltrão e Sandra Canton.

Aos professores Cristiano Fernandes, Marcelo Medeiros e Joel Corrêa da Rosa.

Ao mestrando do ICA, Gustavo Victor C. Ortega, pela ajuda com a aplicação de Redes Neurais e na obtenção dos dados para a mesma.

Ao pessoal da secretaria e do suporte do departamento de Engenharia Elétrica.

Enfim, a todos aqueles que contribuíram de forma direta ou indireta na realização deste feito.

Resumo

Moreira, Rodrigo Pinto; Veiga, Álvaro. **Modelo de Regressão Logística com Transição Suave Estruturado por Árvore (STLR-Tree)**. Rio de Janeiro, 2008. 82p. Dissertação de Mestrado — Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Este trabalho tem como objetivo principal adaptar o modelo STR-Tree, o qual é a combinação de um modelo *Smooth Transition Regression* com *Classification and Regression Tree (CART)*, a fim de utilizá-lo em Classificação. Para isto algumas alterações foram realizadas em sua forma estrutural e na estimação. Devido ao fato de estarmos fazendo classificação de variáveis dependentes binárias, se faz necessária a utilização das técnicas empregadas em Regressão Logística, dessa forma a estimação dos parâmetros da parte linear passa a ser feita por Máxima Verossimilhança. Assim o modelo, que é paramétrico não-linear e estruturado por árvore de decisão, onde cada nó terminal representa um regime os quais têm seus parâmetros estimados da mesma forma que em uma Regressão Logística, é denominado *Smooth Transition Logistic Regression Tree (STLR-Tree)*. A inclusão dos regimes, determinada pela divisão dos nós da árvore, é feita baseada em testes do tipo Multiplicadores de Lagrange, que em sua forma para o caso Gaussiano utiliza a Soma dos Quadrados dos Resíduos em suas estatísticas de teste, aqui é substituída pela Função Desvio (*Deviance*), que é equivalente para o caso dos modelos não Gaussianos, cuja distribuição da variável dependente pertença à família exponencial. Na aplicação a dados reais selecionou-se dois conjuntos das variáveis explicativas de cada uma das duas bases utilizadas, que resultaram nas melhores taxas de acerto, verificadas através de Tabelas de Classificação (Matrizes de Confusão). Esses conjuntos de variáveis foram usados com outros métodos de classificação existentes, são eles: *Generalized Additive Models (GAM), Regressão Logística, Redes Neurais, Análise Discriminante, k-Nearest Neighbor (k-NN)* e *Classification and Regression Trees (CART)*.

Palavras–chave

Modelos não-lineares estruturados por árvore. Classificação. Regressão Logística. Árvores de Classificação e Regressão (CART).

Abstract

Moreira, Rodrigo Pinto; Veiga, Álvaro. **Smooth Transition Logistic Regression Model Tree**. Rio de Janeiro, 2008. 82p. MsC Thesis — Departament of Electric Engineering, Pontifícia Universidade Católica do Rio de Janeiro.

The main goal of this work is to adapt the STR-Tree model, which is the combination of a *Smooth Transition with Regression* model with *Classification and Regression Tree (CART)*, in order to use it in Classification. Some changes were made in its structural form and in the estimation. Due to the fact we are doing binary dependent variables classification, is necessary to use the techniques employed in Logistic Regression, so the estimation of the linear part will be made by Maximum Likelihood. Thus the model, which is nonlinear parametric and structured by a decision tree, where each terminal node represents a regime that have their parameters estimated in the same way as in a Logistic Regression, is called *Smooth Transition Logistic Regression Tree (STLR-Tree)*. The inclusion of the regimes, determined by the splitting of the tree´s nodes, is based on Lagrange Multipliers tests, which for the Gaussian cases uses the Residual Sum-of-squares in their test statistic, here are replaced by the *Deviance function*, which is equivalent to the case of non-Gaussian models, that has the distribution of the dependent variable in the exponential family. After applying the model in two datasets chosen from the bibliography comparing with other methods of classification such as: *Generalized Additive Models (GAM), Logistic Regression, Neural Networks, Discriminant Analyses, k-Nearest Neighbor (k-NN)* and *Classification and Regression Trees (CART)*. It can be seen, verifying in the Classification Tables (Confusion Matrices) that STLR-Tree showed the second best result for the overall rate of correct classification in three of the four applications shown, being in all of them, behind only from GAM.

Keywords

Tree structured nonlinear models. Classification. Logistic Regression. Classifications and Regression Trees (CART).

Sumário

Lista de figuras

Lista de tabelas

Jamais considere seus estudos como uma obrigação, mas como uma oportunidade invejável para aprender a conhecer a influência libertadora da beleza do reino do espírito, para seu próprio prazer pessoal e para proveito da comunidade à qual seu futuro trabalho pertencer.

Albert Einstein, *Notas Autobiográficas Albert Einstein.*