Pontifícia Universidade Católica
do Rio de Janeiro

Marina Sequeiros Dias

Regressão Construtiva em Variedades Implícitas

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
agosto de 2012

Pontifícia Universidade Católica
do Rio de Janeiro

Marina Sequeiros Dias

Regressão Construtiva em Variedades Implícitas

Tese apresentada como requisito parcial para obtenção do título de Doutor pelo Programa de Pós–Graduação em Matemática do Departamento de Matemática do Centro Técnico Científico da PUC–Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Hélio Côrtes Vieira Lopes
Orientador
Departamento de Informática — PUC–Rio

Prof. Marcos Craizer
Departamento de Matemática – PUC-Rio

Prof. Marcelo Gattass
Departamento de Informática – PUC-Rio

Prof. Abelardo Borges Barreto Jr.
PETROBRÁS

Prof. Marcos de Oliveira Lage Ferreira
Instituto de Computação – UFF

Prof. Jessica Quintanilha Kubrusly
Instituto de Matemática – UFF

Prof. José Eugênio Leal
Coordenador Setorial do Centro
Técnico Científico — PUC–Rio

Rio de Janeiro, 30 de agosto de 2012

Marina Sequeiros Dias

Graduou-se em Licenciatura em Matemática na UERJ (Universidade do Estado do Rio de Janeiro), em 2004. Concluiu o mestrado em Matemática Aplicada pela PUC-Rio (Pontifícia Universidade Católica do Rio de Janeiro), em 2007.

Ficha Catalográfica

Dias, Marina Sequeiros

Regressão Construtiva em Variedades Implícitas / Marina Sequeiros Dias; orientador: Hélio Côrtes Vieira Lopes. — Rio de Janeiro : PUC–Rio, Departamento de Matemática, 2012.

v., 156 f: il. ; 29,7 cm

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

Inclui referências bibliográficas.

1. Matemática – Tese.
2. Redução da Dimensionalidade. 3. Aprendizado de variedades. 4. Votação por tensores. 5. Regressão. 6. Partição da Unidade. 7. Aproximação de função. I. Lopes, Hélio Côrtes Vieira. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Matemática. III. Título.

CDD: 510

A minha família

Agradecimentos

A Deus, em primeiro lugar.

Ao professor Hélio Lopes por me incentivar a fazer o doutorado e pela orientação.

Aos meus pais, Benjamim e Maria das Dores, por toda colaboração e por sempre incentivarem meus estudos.

Ao meu marido Eduardo por todo companheirismo e apoio.

Ao meu irmão Adriano pelo incentivo e apoio.

À Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) pelo acolhimento.

A FUNENSEG pela bolsa de Doutorado.

Aos professores do Departamento de Matemática da PUC-Rio pelos ensinamentos.

A Universidade Federal Fluminense, pelo auxílio qualificação Stricto Sensu concedido como forma de incentivo a qualificação dos servidores docentes do quadro efetivo e permanente da UFF.

Aos meus amigos, por estarem sempre participando de todos os momentos de minha vida, em especial, aos amigos que fiz na PUC e aos amigos da Universidade Federal Fluminense do Pólo Universitário de Volta Redonda, que me incentivaram na conclusão dos estudos.

Às secretárias Creuza Nascimento e Katia Beatriz Aguiar pelo carinho e aos Auxiliares Administrativos do Departamento de Matemática da PUC-Rio pela colaboração de sempre.

Resumo

Dias, Marina Sequeiros; Lopes, Hélio Côrtes Vieira. **Regressão Construtiva em Variedades Implícitas**. Rio de Janeiro, 2012. 156p. Tese de Doutorado — Departamento de Matemática, Pontifícia Universidade Católica do Rio de Janeiro.

Métodos de aprendizagem de variedades assumem que um conjunto de dados de alta dimensão possuem uma representação de baixa dimensionalidade. Tais métodos podem ser empregados para simplificar os dados e obter um melhor entendimento da estrutura da qual os dados fazem parte. Nesta tese, utiliza-se o método de aprendizagem de variedades chamado votação por tensores para obter informação da dimensionalidade intrínseca dos dados, bem como estimativas confiáveis da orientação dos vetores normais e tangentes em cada ponto da variedade. Em seguida, propõe-se um método construtivo para aproximar a variedade implícita e realizar uma regressão. O método é chamado de Regressão Construtiva em Variedades Implícitas (RCVI). Com os resultados obtidos no método de votação por tensores, busca-se uma aproximação da variedade através de uma partição do domínio, controlada pelo erro, baseada em malhas 2^n-ádicas (n denota o número de características dos dados de entrada) e em árvore binária com funções de transição suave. A construção consiste em dividir os dados em vários subconjuntos, de maneira a aproximar cada subconjunto de dados com funções implícitas simples. Nesse trabalho empregamos funções polinomiais multivariadas. A forma global pode ser obtida combinando essas estruturas simples. A cada dado de entrada está associada uma saída e a partir de uma boa aproximação da variedade, utilizando esses dados de entrada, busca-se obter uma boa estimativa da saída. Dessa forma, os critérios de parada da subdivisão do domínio incluem uma precisão, definida pelo usuário, na aproximação da variedade, bem como um critério envolvendo a dispersão das saídas em cada subdomínio. Para avaliar o desempenho do método proposto, realiza-se uma regressão com dados reais, compara-se com métodos de aprendizagem supervisionada e efetua-se ainda uma aplicação na área de dados de poços de petróleo.

Palavras–chave

Redução da Dimensionalidade ; Aprendizado de variedades ; Votação por tensores ; Regressão ; Partição da Unidade ; Aproximação de função.

Abstract

Dias, Marina Sequeiros; Lopes, Hélio Côrtes Vieira (advisor). **Constructive Regression on Implicit Manifolds**. Rio de Janeiro, 2012. 156p. D.Sc. Thesis — Departamento de Matemática, Pontifícia Universidade Católica do Rio de Janeiro.

Manifold Learning Methods assume that a high-dimensional data set has a low-dimensional representation. These methods can be employed in order to simplify data, and to obtain a better understanding of the structure of which the data belong. In this thesis, a tensor voting approach is employed as a technique of manifold learning, to obtain information about the intrinsic dimensionality of the data and reliable estimates of the orientation of normal and tangent vectors at each data point in the manifold. Next, a constructive method is proposed to approximate an implicit manifold and perform a regression. The method is called Constructive Regression on Implicit Manifold (RCVI). With the obtained results, search is made in order to obtain a manifold approximation, which consists in a domain partition, error-controlled, based on 2^n-trees (n means the number of features of the input data set) and binary partition trees with smooth transition functions. The construction implies in partition the data set into several subsets in order to approximate each subset with a simple implicit function. In this work, it is used multivariate polynomial functions. The global shape can be obtained by combining these simple structures. Each input data set is associated with an output data, then, from a good manifold approximation using those input data set, it is hoped that occurs a good estimate of the output data. Therefore, the stop criteria of the domain subdivision include a precision, defined by the user, on the manifold approximation, as well as a criterion that involves the output dispersion on each subdomain. To evaluate the performance of the proposed method, a regression on real data is computed, and compared with some supervised learning algorithms and also an application on well data is performed.

Keywords

Dimensionality Reduction ; Manifold learning ; Tensor Voting ; Regression ; Unity Partition ; Function Approximation .

Sumário

1 Introdução **13**
1.1 Contribuições 15
1.2 Estrutura do trabalho 16

2 Variedades **17**

3 Redução da dimensionalidade **22**
3.1 Dimensionalidade intrínseca dos dados 23
3.2 Métodos de redução da dimensionalidade 24
3.3 Aproximação de Funções 32

4 Votação por tensores **38**
4.1 Representação dos dados 40
4.2 O processo de votação 43
4.3 Experimentos 54

5 Representação de variedades implícitas e Partição da Unidade **69**
5.1 Aproximação da variedade implícita 73
5.2 Partição da Unidade 80

6 O método de Regressão Construtiva em Variedades Implícitas **95**
6.1 Estimativa Local 98
6.2 Estimativa Global 99

7 Resultados **100**
7.1 Dados sintéticos 100
7.2 Dados reais 126
7.3 Aplicação em Geologia 128

8 Conclusão e trabalhos futuros **139**

Referências Bibliográficas **142**

A Conceitos de topologia **149**

Lista de figuras

2.1 Variedade de dimensão 1 em $\mathbb{R}^2$ 18
2.2 Variedade de dimensão 2 em $\mathbb{R}^3$ 18

4.1 Princípio Gestalt da proximidade 38
4.2 Princípio Gestalt da boa continuação 38
4.3 Tensor Orientado ou palito. 41
4.4 Tensor não orientado ou bola. 41
4.5 Tensor genérico. 41
4.6 Decomposição de um tensor tridimensional em seus componentes: bola, placa e palito. (69) 42
4.7 Votação Palito. **A** deposita um voto palito em **B**, **C** e **D**. O voto é uma função da posição relativa do votante palito, do receptor e da orientação do votante. 44
4.8 Votação Bola 47
4.9 Votação por tensores genéricos. Neste esquema, o votante é um tensor com duas normais ($\mathbf{v}_{n1}$ e $\mathbf{v}_{n2}$) em um espaço tridimensional. 48
4.10 Hélice 55
4.11 *Swiss Roll* 58
4.12 Superfície implícita no $\mathbb{R}^4$ definida pela equação complexa $\mathbf{w} = \mathbf{z}^2$ 62

5.1 T_xM 71
5.2 Árvore com suas funções de transição em cada região do domínio. 87

7.1 Poço 1 132
7.2 Poço 2 133
7.3 Poço 3 135
7.4 Poço 4 137

A.1 Deformação de uma rosquinha em um copo de café. (Lee,2011(34)) 149
A.2 Uma carta local. (Lee,2003(33)) 152
A.3 Mudança de coordenadas 153

Lista de tabelas

4.1 Curva no $\mathbb{R}^3$. Taxa da estimativa correta da dimensionalidade(DE) e erro (em graus) da orientação tangente, em função do parâmetro σ e do número de vizinhos (viz). 56

4.2 Curva no $\mathbb{R}^3$. Taxa da estimativa correta da dimensionalidade(DE) e erro (em graus) da orientação tangente, em função do parâmetro σ e do número de vizinhos (viz). 56

4.3 Curva no $\mathbb{R}^3$. Taxa da estimativa correta da dimensionalidade(DE) e erro (em graus) da orientação tangente, em função do parâmetro σ e do número de vizinhos (viz). 57

4.4 Curva no $\mathbb{R}^3$. Taxa da estimativa correta da dimensionalidade (DE) e erro (em graus) da orientação tangente, em função do parâmetro σ e do número de vizinhos (viz). 57

4.5 *Swiss Roll*. Taxa da estimativa correta da dimensionalidade (DE) e erro (em graus) da orientação normal, em função do parâmetro σ e do número de vizinhos (viz). 58

4.6 *Swiss Roll*. Taxa da estimativa correta da dimensionalidade (DE) e erro (em graus) da orientação normal, em função do parâmetro σ e do número de vizinhos (viz). 59

4.7 *Swiss Roll*. Taxa da estimativa correta da dimensionalidade (DE) e erro (em graus) da orientação normal, em função do parâmetro σ e do número de vizinhos (viz). 59

4.8 *Swiss Roll*. Taxa da estimativa correta da dimensionalidade (DE) e erro (em graus) da orientação normal, em função do parâmetro σ e do número de vizinhos (viz). 59

4.9 Curva no $\mathbb{R}^4$. Taxa da estimativa correta da dimensionalidade(DE) e erro (em graus) da orientação tangente, em função do parâmetro σ e do número de vizinhos (viz). 60

4.10 Curva no $\mathbb{R}^4$. Taxa da estimativa correta da dimensionalidade (DE) e erro (em graus) da orientação tangente, em função do parâmetro σ e do número de vizinhos (viz). 61

4.11 Curva no $\mathbb{R}^4$. Taxa da estimativa correta da dimensionalidade(DE) e erro (em graus) da orientação tangente, em função do parâmetro σ e do número de vizinhos (viz). 61

4.12 Curva no $\mathbb{R}^4$. Taxa da estimativa correta da dimensionalidade (DE) e erro (em graus) da orientação tangente, em função do parâmetro σ e do número de vizinhos (viz). 61

4.13 Superfície no $\mathbb{R}^4$. Taxa da estimativa correta da dimensionalidade(DE) e erro (em graus) da orientação normal, em função do parâmetro σ e do número de vizinhos (viz). 64

4.14 Superfície no $\mathbb{R}^4$. Taxa da estimativa correta da dimensionalidade(DE) e erro (em graus) da orientação normal, em função do parâmetro σ e do número de vizinhos (viz). 64

4.15 Superfície no $\mathbb{R}^4$. Taxa da estimativa correta da dimensionalidade(DE) e erro (em graus) da orientação normal, em função do parâmetro σ e do número de vizinhos (viz). 65
4.16 Superfície no $\mathbb{R}^4$. Taxa da estimativa correta da dimensionalidade(DE) e erro (em graus) da orientação normal, em função do parâmetro σ e do número de vizinhos (viz). 65
4.17 Volume no $\mathbb{R}^4$. Taxa da estimativa correta da dimensionalidade(DE) e erro (em graus) da orientação normal, em função do parâmetro σ e do número de vizinhos (viz). 67
4.18 Volume no $\mathbb{R}^4$. Taxa da estimativa correta da dimensionalidade (DE) e erro (em graus) da orientação normal, em função do parâmetro σ e do número de vizinhos (viz). 67
4.19 Volume no $\mathbb{R}^4$. Taxa da estimativa correta da dimensionalidade(DE) e erro (em graus) da orientação normal, em função do parâmetro σ e do número de vizinhos (viz). 68
4.20 Volume no $\mathbb{R}^4$. Taxa da estimativa correta da dimensionalidade(DE) e erro (em graus) da orientação normal, em função do parâmetro σ e do número de vizinhos (viz). 68

7.1 Curva no $\mathbb{R}^3$. Taxa da estimativa correta da dimensionalidade (DE). As estatísticas dos ângulos referem-se aos erros de orientação obtidos. 102
7.2 Subdivisão espacial da curva no $\mathbb{R}^3$ por Malha 2^n-ádica. 103
7.3 Aproximação da curva no $\mathbb{R}^3$ por Malha 2^n-ádica. 103
7.4 Dados da subdivisão espacial da curva no $\mathbb{R}^3$ pela Árvore BSP. 104
7.5 Aproximação da curva no $\mathbb{R}^3$ pela Árvore BSP. 106
7.6 Curva no $\mathbb{R}^3$. Regressão fora da amostra com Malha 2^n-ádica e Árvore BSP. 107
7.7 Curva no $\mathbb{R}^3$. Regressão fora da amostra. 107
7.8 Curva no $\mathbb{R}^3$. Regressão fora da amostra 2. 107
7.9 *Swiss Roll*. Taxa da estimativa correta da dimensionalidade (DE). As estatísticas dos ângulos referem-se aos erros de orientação obtidos. 108
7.10 Subdivisão espacial da superfície no $\mathbb{R}^3$ por Malha 2^n-ádica. 108
7.11 Aproximação da superfície no $\mathbb{R}^3$ por Malha 2^n-ádica. 109
7.12 Dados da subdivisão espacial da superfície no $\mathbb{R}^3$ pela Árvore BSP. 109
7.13 Aproximação da superfície no $\mathbb{R}^3$ pela Árvore BSP. 110
7.14 Superfície no $\mathbb{R}^3$. Regressão fora da amostra. 111
7.15 Superfície no $\mathbb{R}^3$. Regressão fora da amostra 2. 111
7.16 Curva no $\mathbb{R}^4$. Taxa da estimativa correta da dimensionalidade (DE). As estatísticas dos ângulos referem-se aos erros de orientação obtidos. 112
7.17 Subdivisão espacial da curva no $\mathbb{R}^4$ por Malha 2^n-ádica. 112
7.18 Aproximação da curva no $\mathbb{R}^4$ por Malha 2^n-ádica. 113
7.19 Dados da subdivisão espacial da curva no $\mathbb{R}^4$ pela Árvore BSP. 114
7.20 Aproximação da curva no $\mathbb{R}^4$ pela Árvore BSP. 115
7.21 Curva no $\mathbb{R}^4$. Regressão fora da amostra. 116
7.22 Curva no $\mathbb{R}^4$. Regressão fora da amostra 2. 117
7.23 Superfície no $\mathbb{R}^4$. Taxa da estimativa correta da dimensionalidade (DE). As estatísticas dos ângulos referem-se aos erros de orientação obtidos. 117

7.24 Subdivisão espacial da superfície no $\mathbb{R}^4$ por Malha 2^n-ádica. 117
7.25 Aproximação da superfície no $\mathbb{R}^4$ por Malha 2^n-ádica. 118
7.26 Dados da subdivisão espacial da superfície no $\mathbb{R}^4$ pela Árvore BSP. 119
7.27 Aproximação da superfície no $\mathbb{R}^4$ pela Árvore BSP. 120
7.28 Superfície no $\mathbb{R}^4$. Regressão fora da amostra. 121
7.29 Superfície no $\mathbb{R}^4$. Regressão fora da amostra 2. 121
7.30 Volume no $\mathbb{R}^4$. Taxa da estimativa correta da dimensionalidade (DE). As estatísticas dos ângulos referem-se aos erros de orientação obtidos. 121
7.31 Subdivisão espacial do volume no $\mathbb{R}^4$ por Malha 2^n-ádica. 122
7.32 Aproximação do volume no $\mathbb{R}^4$ por Malha 2^n-ádica. 123
7.33 Dados da subdivisão espacial do volume no $\mathbb{R}^4$ pela Árvore BSP. 123
7.34 Aproximação do volume no $\mathbb{R}^4$ pela Árvore BSP. 124
7.35 Volume no $\mathbb{R}^4$. Regressão fora da amostra. 125
7.36 Volume no $\mathbb{R}^4$. Regressão fora da amostra 2. 126
7.37 Regressão com dados reais 128
7.38 Resumo dos dados do poço 1. 132
7.39 Dados do poço 1. Variável prevista: DT. 132
7.40 Dados do poço 1. Variável prevista: RHOB. 133
7.41 Resumo dos dados do poço 2. 134
7.42 Dados do poço 2. Variável prevista: DT. 134
7.43 Dados do poço 2. Variável prevista: RHOB. 134
7.44 Resumo dos dados do poço 3. 135
7.45 Dados do poço 3. Variável prevista: DT. 136
7.46 Dados do poço 3. Variável prevista: RHOB. 136
7.47 Resumo dos dados do poço 4. 137
7.48 Dados do poço 4. Variável prevista: DT. 137
7.49 Dados do poço 4. Variável prevista: RHOB. 138