1
Introdução

Nas mais variadas áreas, como por exemplo, na engenharia, economia e atuária, existem decisões que envolvem incertezas. Sempre que há incerteza, há a necessidade da utilização de ferramentas estatísticas, que procuram compreender o cenário e determinar o melhor rumo a ser seguido. Por causa disso, cada vez mais, torna-se necessário o estudo e o aprimoramento dessas técnicas.

Uma dessas ferramentas, que vem sendo estudada e aprimorada por anos, são os modelos de aprendizagem supervisionada. Estes procuram, a partir de uma amostra de tamanho N, com dados de entrada $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \in \mathbb{R}^d$ e dados de saída $Y = \{y_1, y_2, \ldots, y_N\} \in \mathbb{R}$, encontrar uma função

$$\begin{array}{rccc} f: & \mathbb{R}^d & \longrightarrow & \mathbb{R} \\ & \mathbf{x} & \longmapsto & y = f(\mathbf{x}) \end{array}$$

que descreva a relação entre $\mathbf{x}_i$ e y_i. Em geral $\mathbf{x}_i$ é chamado de variável explicativa e y_i de variável de predição.

São muitas as formas como esse problema pode ser resolvido. Provavelmente a alternativa mais simples é o modelo de regressão linear, definido por:

$$y_i = \alpha + \beta \mathbf{x}_i + \varepsilon_i,$$

onde α e $\beta = (\beta_1, \ldots, \beta_d)$ são constantes a serem determinadas e ε_i representa a componente aleatória no mecanismo gerador de y_i.

Por um lado existem as vantagens de um modelo simples como esse, seus coeficientes podem ser facilmente encontrados através de mínimos quadrados ordinários, por exemplo. Em compensação, sua estrutura baseia-se em hipóteses muito fortes, nesse caso, a suposta relação linear entre os elementos de X e Y. Algumas vezes essa é uma suposição bem razoável, mas na maioria das vezes não.

PUC-Rio - Certificação Digital Nº 0521445/CA

Muitas das alternativas não lineares vem de métodos não paramétricos, que de forma geral são modelados por:

$$y_i = f(\mathbf{x}_i) + \varepsilon_i$$

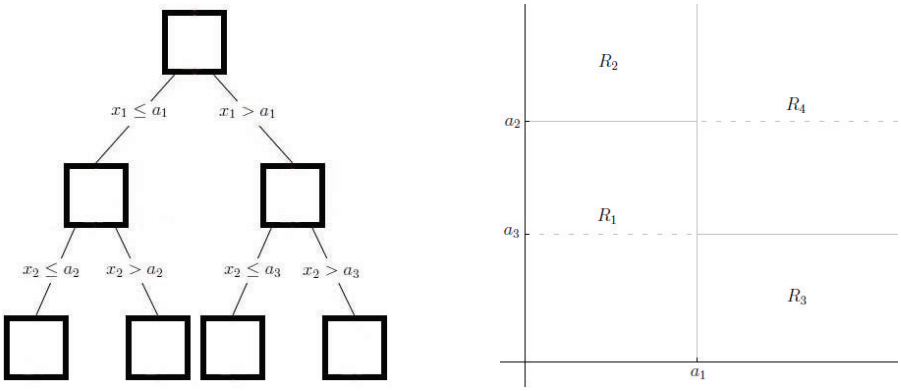
onde f é uma função que descreve a relação entre $\mathbf{x}_i$ e y_i e $\{\varepsilon_i\}_{i=1}^{N}$ são variáveis aleatórias independentes e identicamente distribuídas que representam a componente aleatória da estimativa. Os métodos de regressão baseados em árvores são exemplos de modelos não lineares e não paramétricos e o foco desse trabalho. No texto a seguir serão dados mais detalhes sobre eles. Outros exemplos de modelos não paramétricos podem ser encontrados no capítulo 5 do livro de Wasserman (48).

Métodos Baseados em Árvores

Uma regressão baseada em árvores é um modelo cuja função de predição f é descrita por:

$$f(\mathbf{x}) = \sum_{k=1}^{l} f_k(\mathbf{x}) I_k(\mathbf{x} \in R_k) \tag{1-1}$$

onde $\{R_k\}_{k=1}^{l}$ são retângulos que formam uma partição do domínio de f, f_k são funções de regressão definidas em R_k e I_k é a função indicadora, que assume valor 1 se $\mathbf{x} \in R_k$ e zero caso contrário. Tais retângulos são definidos pela estrutura de árvore binária: cada divisão de um nó interno particiona o domínio por retas paralelas aos eixos e no final do processo cada folha i define um retângulo R_i, como mostra a figura 1.1 a seguir.



1.1(a): Árvore 1.1(b): Retângulos

Figura 1.1: Métodos Baseados em Árvores

A partição por retas paralelas aos eixos possibilita interpretações sobre o modelo gerado, o que é um ponto muito positivo. Por outro lado, o modelo tem dificuldade em reconhecer padrões mais elaborados no domínio.

PUC-Rio - Certificação Digital Nº 0521445/CA

Entre os métodos baseados em árvores, o que diferencia um modelo do outro é: a forma como é escolhida a coordenada j, na qual será feita a divisão em cada nó; o valor a em que essa coordenada será dividida; os critérios que determinam se um nó é um nó interno ou terminal, ou seja, uma folha; e o formato das funções f_k.

O algoritmo CART (*Classification and Regression Tree*), proposto por Breiman, Friedman, Olshen e Stone em 1984 (5), é com certeza um marco na bibliografia de árvores de classificação (variável de predição inteira) e regressão (variável de predição contínua). Apesar de antigo, continua sendo uma importante referência e comparação para qualquer método nessa área.

O algoritmo de regressão do modelo CART procede da seguinte maneira. A escolha da coordenada j e do valor de $a \in \mathbb{R}$ definem uma divisão nos dados de entrada: $X = X_L \cup X_R$ e $Y = Y_L \cup Y_R$, onde

$$\begin{array}{rclcrcl} X_L & = & \{\mathbf{x}_i \in X \mid x_{ij} \leq a\} & & X_R & = & \{\mathbf{x}_i \in X \mid x_{ij} > a\} \\ Y_L & = & \{y_i \in Y \mid \mathbf{x}_i \in X_L\} & & Y_R & = & \{y_i \in Y \mid \mathbf{x}_i \in X_R\}. \end{array}$$

e x_{ij} é a j-ésima coordenada do vetor $\mathbf{x}_i$. O valor de a escolhido para a coordenada j, a_j, é aquele que maximiza o decrescimento no erro médio quadrático, ou seja, aquele que maximiza $\Delta\mathcal{R} = \mathcal{R} - \mathcal{R}_L - \mathcal{R}_R$, com

$$\begin{array}{ccc} \mathcal{R} = \sum_{y_i \in Y}(y_i - \bar{y})^2 & \mathcal{R}_L = \sum_{y_i \in Y_L}(y_i - \bar{y}_L)^2 & \mathcal{R}_R = \sum_{y_i \in Y_R}(y_i - \bar{y}_R)^2 \\ \bar{y} = \frac{1}{\#Y}\sum_{y_i \in Y} y_i & \bar{y}_L = \frac{1}{\#Y_L}\sum_{y_i \in Y_L} y_i & \bar{y}_R = \frac{1}{\#Y_R}\sum_{y_i \in Y_R} y_i. \end{array}$$

Depois de computados todos os valores de a_j e seu respectivos decrescimentos $\Delta\mathcal{R}_j$, a coordenada escolhida para ser feita a divisão é aquela com maior $\Delta\mathcal{R}_j$. Essa sequência de passos se repete até que a árvore esteja completa. Por fim, suas funções de regressão em R_k são constantes e definidas pela média amostral entre os pontos que pertencem ao retângulo R_k:

$$\mathbf{f}_k(\mathbf{x}) = c_k = \frac{\sum_{\{i \mid \mathbf{x}_i \in R_k\}} y_i}{\sum_{\{i \mid \mathbf{x}_i \in R_k\}} 1}.$$

Outro método bastante conhecido é o algoritmo MARS (*Multivariate Adaptive Regression Splines*). Ele foi desenvolvido por Friedman em 1991 (17) e pode ser visto como uma modificação do modelo CART. Sua modelagem é preparada para trabalhar com grandes quantidades de dados de entrada e é um método de regressão mais flexível, que generaliza a saída constante por partes do modelo CART para funções contínuas e diferenciáveis, geradas através de *splines*. Mais detalhes sobre esse modelo podem ser encontrados no livro de Hastie, Tibshirani e Friedman (20), e no livro de Hastie e Tibshirani (19).

PUC-Rio - Certificação Digital Nº 0521445/CA

Em 1994 Chaudhuri apresentou no artigo (9) o modelo SUPPORT (*Smoothed and Unsmoothed Piecewise-Polynomial Regression Trees*). A principal diferença nesse método é que as funções f_k são polinomiais e a saída final é suave. Seu algoritmo recursivo é composto por três etapas. Na primeira delas o domínio é dividido em retângulos, de forma semelhante ao modelo CART, até que os dados em cada retângulo estejam adequadamente ajustados por um polinômio de grau pré-estabelecido. Na segunda etapa os coeficientes desse polinômio são escolhidos de forma a melhor ajustar, não só os dados em cada retângulo, como também em uma vizinhança dele. Para terminar, o estimador final é obtido através de uma média ponderada dos polinômios já definidos, onde o conceito de suavidade é aplicado para garantir que a saída do estimador não seja descontínua, como no caso do modelo CART.

Bem mais recente, em 2008, da Rosa, Veiga e Medeiros apresentaram no artigo (11) um novo algoritmo em árvores de regressão, denominado STR-Tree (*Smooth Transition Regression Tree*). Como no modelo CART, nele as funções f_k são constantes, $\mathbf{f}_k(\mathbf{x}) = c_k$. Seu grande diferencial é a utilização da função logística, apresentada abaixo, para realizar a suavidade na estimativa.

$$G(\mathbf{x}; j, a, \lambda) = \frac{1}{1 + e^{-\lambda(x_j - a)}}$$

Na estrutura do STR-Tree cada nó interno guarda uma função logística, onde x_j é a coordenada de divisão desse nó, a o valor em que essa divisão é feita e λ um parâmetro de suavidade. Observe que $G(\mathbf{x}; j, a, \lambda) \in [0, 1]$. Sua estimativa final não será mais dada pela equação (1-1) e sim por

$$f(\mathbf{x}) = \sum_{k=1}^{l} c_k B_k(\mathbf{x}),$$

onde B_k é produto de funções logísticas e l o número de folhas na árvore. Por exemplo, se a árvore tiver quatro folhas e três nós internos, como na figura 1.1(a), as funções B_k serão:

$$\begin{aligned}
B_1(\mathbf{x}) &= G(\mathbf{x}; 1, a_1, \lambda) G(\mathbf{x}; 2, a_1, \lambda) \\
B_2(\mathbf{x}) &= G(\mathbf{x}; 1, a_1, \lambda)(1 - G(\mathbf{x}; 2, a_2, \lambda)) \\
B_3(\mathbf{x}) &= (1 - G(\mathbf{x}; 1, a_1, \lambda)) G(\mathbf{x}; 2, a_3, \lambda) \\
B_4(\mathbf{x}) &= (1 - G(\mathbf{x}; 1, a_1, \lambda))(1 - G(\mathbf{x}; 2, a_3, \lambda))
\end{aligned}$$

A idéia intuitiva é que $B_k(\mathbf{x})$ indica um peso associado à regressão na folha k quando o dado de entrada for $\mathbf{x}$. Esse peso é inversamente relacionado à distancia entre $\mathbf{x}$ e a fronteira de R_k O interessante é que para um mesmo

PUC-Rio - Certificação Digital Nº 0521445/CA

dado de entrada $\mathbf{x}$ a soma dos pesos para todas as folhas é 1, isto é, a árvore gera pela sua própria construção uma partição da unidade.

Contribuição

A proposta desse trabalho é um novo modelo baseado em árvores de regressão, chamado de método de Regressão Construtiva em Regiões Implícitas (RCRI). Também se trata de um modelo com árvores binárias e funções f_k constantes. Além disso, assim como as regressões nos modelos SUPPORT e STR-Tree, sua saída é uma função diferenciável.

A primeira diferença entre esse novo modelo e os demais é a forma como a suavidade é feita. Apesar de bem parecida com a suavidade no modelo STR-Tree, a função logística é substituída por uma função degrau $\mathbf{s}$, polinomial, tal que o conjunto $\{t \in \mathbb{R} \mid \mathbf{s}(t) \neq 0 \text{ e } \mathbf{s}(t) \neq 1\}$ é limitado. Esta pequena mudança já garante mais eficiência computacional para o algoritmo, como será mostrado mais adiante.

A função degrau escolhida para esse trabalho foi sugerida por Lage, Bordignon, Petronetto, Veiga, Tavares, Lewiner e Lopes em (30), onde, com o domínio restrito ao plano, as retas que definem a partição não são mais paralelas aos eixos, e sim na direção de maior distribuição dos dados de entrada, obtida pela Análise de Componentes Principais. Esta idéia foi originada pelo artigo (4) de Bordignon, Castro, Lewiner, Lopes e Tavares, que trata do uso de árvores binárias para compressão de pontos esparsos.

A segunda e grande diferença é que no RCRI cada nó interno não particiona mais o domínio por retas paralelas aos eixos. A partição é feita por conjunto de nível zero de polinômios definidos em $\mathbb{R}^d$ e por isso podem assumir formatos bem mais elaborados, como por exemplo, no plano esses conjuntos podem ser retas não necessariamente paralelas aos eixos, elipses, hipérboles ou até curvas de graus maiores. Dessa forma, a partição do domínio, que antes era feita por retângulos, agora é feita por regiões definidas implicitamente.

Esse novo modelo foi aplicado para dados de atuária e de geologia. Tais conjuntos de dados foram analisados pela primeira vez através de algoritmos de árvores de regressão. Para terminar essa seção de contribuições tem que ser destacado que, ainda nesse trabalho, o método RCRI foi expandido para sua forma intervalar, onde ele passa a trabalhar não só com dados reais como também com dados intervalares na variável de predição.

Organização

O capítulo 2 a seguir apresenta a construção recursiva do método RCRI. Lá estão a caracterização de sua árvore, a definição das regiões implícitas que

PUC-Rio - Certificação Digital Nº 0521445/CA

particionam o domínio, os critérios de parada e outros detalhes necessários para definir o algoritmo.

Já no capítulo 3 o objetivo é apresentar como é feita a suavidade nesse novo modelo. Para isso, primeiro são apresentados os conceitos de função de transição e com eles a idéia de suavidade é introduzida naturalmente. Ainda nesse capítulo são definidas duas diferentes formas de estimação dos parâmetros das regressões nas regiões que particionam o domínio.

Nesse momento toda a estrutura do método RCRI já é conhecida e seu algoritmo já está completo. O próximo passo é a realização de testes, que são apresentados no capítulo 5, a fim de avaliar seu desempenho em dados gerados e em dados reais. Esse capítulo termina com uma comparação entre o novo método e os métodos CART (5), MARS (17) e STR-Tree (11).

As aplicações em atuária e geologia estão apresentadas no capítulo 6. Lá cada aplicação é desenvolvida, seu problema é contextualizado, sua modelagem é descrita com detalhes, a escolha dos parâmetros justificadas, os resultados apresentados e devidamente comentados.

No capítulo 7 é feita a extensão do método RCRI para dados intervalares. Todo esse capítulo se desenvolve com o objetivo de mostrar as modificações sofridas pelo algoritmo original e o desempenho dessa nova versão, ilustrado com exemplos e uma aplicação em geologia.

Por fim, no capítulo 8, são apresentadas as conclusões finais. Os apêndices A, B e C apresentam pequenos resumo sobre mínimos quadrados para regressões lineares, SVR e aritmética intervalar, respectivamente, e podem vir a auxiliar na compreensão do texto.

PUC-Rio - Certificação Digital Nº 0521445/CA