Smooth Transition Logistic Regression-Tree (STLR-Tree)

Rodrigo Pinto Moreira

Orientador: Álvaro Veiga
Departamento de Engenharia Elétrica
Pontifícia Universidade Católica do Rio de Janeiro

Dissertação de Mestrado

11 de abril de 2008

Sumário

Sumário

Motivação e Objetivo

- Na busca por uma melhor explicação do comportamento complexo dos modelos de regressão e das séries temporais, a utilização da modelagem não-linear vem crescendo ao longo dos anos amparada pelo advento da informática, a modernização dos pacotes estatísticos e pelo simples fato de fornecerem um maior conhecimento sobre o fenômeno em estudo do que os modelos lineares.
- Adaptar o modelo STR-Tree, o qual é a combinação de um modelo Smooth Transition Regression (STR) com Classification and Regression Tree (CART), a fim de utilizá-lo em Classificação, fazendo a estimação de seus parâmetros lineares como em uma Regressão Logística.

Motivação e Objetivo

- Na busca por uma melhor explicação do comportamento complexo dos modelos de regressão e das séries temporais, a utilização da modelagem não-linear vem crescendo ao longo dos anos amparada pelo advento da informática, a modernização dos pacotes estatísticos e pelo simples fato de fornecerem um maior conhecimento sobre o fenômeno em estudo do que os modelos lineares.
- Adaptar o modelo STR-Tree, o qual é a combinação de um modelo Smooth Transition Regression (STR) com Classification and Regression Tree (CART), a fim de utilizá-lo em Classificação, fazendo a estimação de seus parâmetros lineares como em uma Regressão Logística.

Sumário

Classification and Regression Trees (CART)

- Fácil entendimento;
- Particionam de forma recursiva o espaço das covariáveis, $\mathbb{X}$;
- Estrutura é representada e ajustada em um gráfico que cresce de um nó inicial (ou nó raíz),em direção aos nós terminais (ou folhas) passando pelos nós intermediários (ou nós geradores, criadores);
- Cada nó gerador na posição j dá origem a dois novos nós nas posições $2j+1$ e $2j+2$.

Classification and Regression Trees (CART)

- Fácil entendimento;
- Particionam de forma recursiva o espaço das covariáveis, $\mathbb{X}$;
- Estrutura é representada e ajustada em um gráfico que cresce de um nó inicial (ou nó raíz),em direção aos nós terminais (ou folhas) passando pelos nós intermediários (ou nós geradores, criadores);
- Cada nó gerador na posição j dá origem a dois novos nós nas posições $2j+1$ e $2j+2$.

Classification and Regression Trees (CART)

- Fácil entendimento;
- Particionam de forma recursiva o espaço das covariáveis, $\mathbb{X}$;
- Estrutura é representada e ajustada em um gráfico que cresce de um nó inicial (ou nó raíz),em direção aos nós terminais (ou folhas) passando pelos nós intermediários (ou nós geradores, criadores);
- Cada nó gerador na posição j dá origem a dois novos nós nas posições $2j+1$ e $2j+2$.

Classification and Regression Trees (CART)

- Fácil entendimento;
- Particionam de forma recursiva o espaço das covariáveis, $\mathbb{X}$;
- Estrutura é representada e ajustada em um gráfico que cresce de um nó inicial (ou nó raíz),em direção aos nós terminais (ou folhas) passando pelos nós intermediários (ou nós geradores, criadores);
- Cada nó gerador na posição j dá origem a dois novos nós nas posições $2j+1$ e $2j+2$.

Formulação Matemática

Seja $\mathbf{x}_i = (x_{1i}, x_{2i}, ..., x_{qi})' \in \mathbb{X} \subseteq \mathbb{R}^q$ o vetor que contém q variáveis explicativas (covariáveis ou preditores) para uma resposta univariada contínua, $y_i \in \mathbb{R}, \quad i = 1, ..., n$. Suponha que a relação entre y_i e $\mathbf{x}_i$ segue o modelo de regressão

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

Um modelo de árvore de regressão com K folhas é um modelo de particionamento recursivo do espaço das covariáveis, $\mathbb{X}$, que aproxima $f(\cdot)$ por uma função geral não-linear, $H(\mathbf{x}_i; \psi)$ de $\mathbf{x}_i$ e definida pelo vetor de parâmetros $\psi \in \mathbb{R}^r$ onde r é o número total de parâmetros do modelo.

Formulação Matemática

$H(\mathbf{x}_i;\boldsymbol{\psi})$ é uma função constante por partes definida por K subregiões $k_j(\boldsymbol{\theta}_j), \quad i = 1, ..., K$ de $\mathbb{K} \subset \mathbb{R}^q$. A determinação dessas subregiões é feita pelo vetor de parâmetros não-lineares $\boldsymbol{\theta}_j, \quad j = 1, ..., K$ onde

$$f(\mathbf{x}_i) \approx H(\mathbf{x}_i;\boldsymbol{\psi}) = \sum_{j=1}^{K} \beta_j I_j(\mathbf{x}_i;\boldsymbol{\theta}_j) \tag{1}$$

em que

$$I_j(\mathbf{x}_i;\boldsymbol{\theta}_j) = \begin{cases} 1 & , se\ \mathbf{x}_i \in k_j(\boldsymbol{\theta}_j) \\ 0 & , se\ \mathbf{x}_i \notin k_j(\boldsymbol{\theta}_j) \end{cases};$$

e o vetor de parâmetros é $\boldsymbol{\psi} = (\beta_1, ..., \beta_K, \boldsymbol{\theta}'_1, ..., \boldsymbol{\theta}'_K)'$.

Exemplo Numérico

Cada nó gerador tem uma variável de transição $x_{s_j i} \in \mathbf{x}_i$ associada, onde $s_j \in \mathbb{S} = \{1, 2, ..., m\}$.
$\mathbb{J}$ índices dos nós geradores e $\mathbb{T}$ índices dos nós terminais.

- Exemplo ($d = 1$) e $K = 2$ nós terminais:

$$y_i = \beta_1 I_0(\mathbf{x}_i; s_0, c_0) + \beta_2[1 - I_0(\mathbf{x}_i; s_0, c_0)] + \epsilon_i$$

onde

$$I_0(\mathbf{x}_i; s_0, c_0) = \begin{cases} 1 & , se\ \mathbf{x}_{s_0 i} \leq c_0 \\ 0 & , se\ \mathbf{x}_{s_0 i} > c_0 \end{cases}$$

e $s_0 \in \mathbb{S} = 1, 2, ..., m$.

Exemplo Numérico

Esse exemplo numérico ilustra como é feita a divisão no espaço das covariáveis, $\mathbb{X} \subseteq \mathbb{R}^2$ a partir da estrutura do modelo.

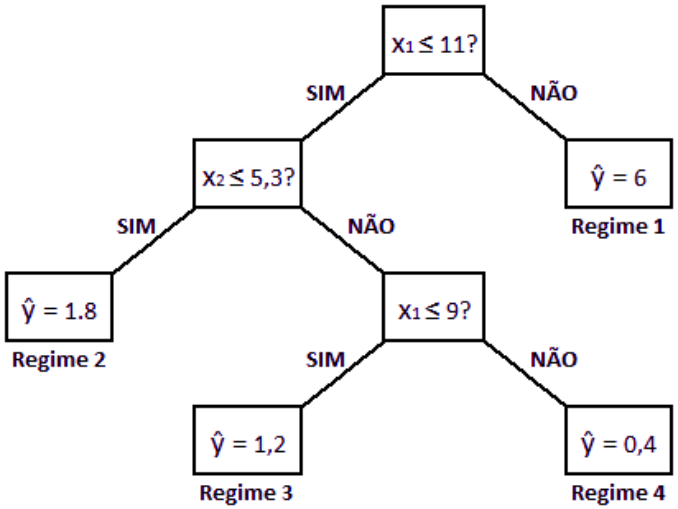

Figura: Estrutura do modelo

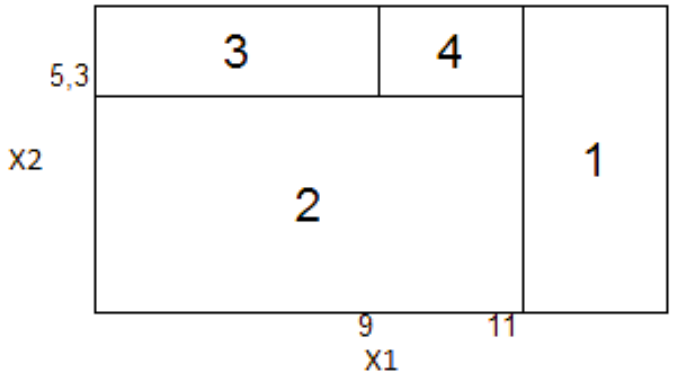

Figura: Divisão do espaço das covariáveis

Exemplo Numérico

Tabela com as sentenças lógicas e o correspondente valor da variável dependente estimada, $\hat{y}$

Divisões de $\mathbb{X}$	$\hat{y}$
se $x_1 \geq 11$	6
se $x_1 < 11$ e $x_2 < 5,3$	1,8
se $x_1 < 9$ e $x_2 \geq 5,3$	1,2
se $9 < x_1 < 11$ e $x_2 \geq 5,3$	0,4

Tabela: Divisão do espaço das covariáveis - Sentenças lógicas

Algoritmo de Crescimento

Basicamente busca-se, a partir do nó raíz, x_{s_0} e c_0 que minimizam a soma dos erros quadráticos:

$$SQ^{Arv1} = \sum_{i=1}^{n} \{y_i - \beta_1 I_0(\mathbf{x}_i; s_0, c_0) - \beta_2[1 - I_0(\mathbf{x}_i; s_0, c_0)]\}^2$$

A estimação dos parâmetros β_1 e β_2 é dada por

$$\hat{\beta}_1^{MQ} = \frac{\sum_{i=1}^{n} y_i I_0(\mathbf{x}_i; s_0, c_0)}{\sum_{i=1}^{n} I_0(\mathbf{x}_i; s_0, c_0)}$$

$$\hat{\beta}_2^{MQ} = \frac{\sum_{i=1}^{n} y_i[1 - I_0(\mathbf{x}_i; s_0, c_0)]}{\sum_{i=1}^{n}[1 - I_0(\mathbf{x}_i; s_0, c_0)]}$$

Sumário

MLG

A Regressão Logística é um caso particular de Modelos Lineares Generalizados (MLG), os quais permitem o trabalho de modelagem adimitindo que a variável dependente (Y_i) siga uma distribuição que pertença à *família exponencial*, além da Normal (Gaussiana). Podendo ser uma Binomial, Poisson, Gamma, dentre outras.

$$y_i = \sum_{j=1}^{p} x_{ji}\beta_j + \epsilon_i, \quad i = 1, ..., n$$

$$y_i = \mathbb{E}(Y_i|\mathbf{x}_i) + \epsilon_i$$

Componentes de um MLG

O modelo pode ser dividido em três partes:

- Componente aleatória: componente da variável aleatória Y_i, $i = 1, ...n$, admitindo que a mesma tenha distribuição pertencente à família exponencial;
- Preditor linear: representado por η e denominado por

$$\eta_i = \sum_{j=1}^{p} x_{ji}\beta_j, \quad i = 1, ..., n$$

- Função de ligação: função monotônica diferenciável que liga o preditor linear a parte aleatória, ou seja, $g(\mu_i) = \eta_i$, $i = 1, ...n$

Ligações Canônicas

Temos disponíveis outras funções de ligação clássicas como, por exemplo, para o caso de uma distribuição Binomial, em que $\mu \in (0, 1)$

1. Logito: $g(\mu_i) = \log\left(\frac{\mu}{1-\mu}\right)$
2. Probito: $g(\mu_i) = \Phi^{-1}(\mu)$
   onde $\Phi(\cdot)$ é uma função de distribuição acumulada Normal padrão
3. Complemento Log-log: $g(\mu_i) = \log[-\log(1-\mu)]$

Regressão Logística

Diferentemente do caso linear Gaussiano aqui $y_i = \pi(\mathbf{x}_i) + \epsilon_i$, onde ϵ_i assume apenas dois valores dependendo daquele assumido por y_i.

- Se $y_i = 1$ então $\epsilon_i = 1 - \pi(\mathbf{x}_i)$ com probabilidade $\pi(\mathbf{x}_i)$;
- Caso $y_i = 0$, $\epsilon_i = -\pi(\mathbf{x}_i)$ com probabilidade $1 - \pi(\mathbf{x}_i)$.

$$\mathbb{E}[Y_i|\mathbf{x}_i] = \mu_i = \sum_{j=1}^{p} x_{ji}\beta_j, \quad i = 1, ..., n$$

Após a estimação dos β´s podemos encontrar os valores de $\hat{\mu_1}, ..., \hat{\mu_n}$ escrevendo assim o modelo estimado da seguinte maneira

$$\hat{\mu_i} = \sum_{j=1}^{p} x_{ji}\hat{\beta}_j, \quad i = 1, ..., n$$

Regressão Logística

Teremos que definir a probabilidade de interesse, ou *probabilidade de sucesso*, $\mathbb{P}(Y_i = 1) = \pi_i$ e a *probabilidade de fracasso* $\mathbb{P}(Y_i = 0) = 1 - \pi_i$. Para investigar a relação entre a probabilidade de sucesso π_i e o vetor de covariáveis escrevemos o modelo

$$\mathbb{E}(Y_i|\mathbf{x}_i) = \pi_i = \sum_{j=1}^{p} x_{ji}\beta_j, \quad i = 1, ..., n$$

Entretanto tal igualdade não pode ser aceita dado que $-\infty < n < \infty$ e $0 < \pi < 1$.

$$g(\pi_i) = \eta_i$$

$$g(\pi_i) = \sum_{j=1}^{p} x_{ji}\beta_j, \quad i = 1, ..., n$$

Regressão Logística

Optamos pela logito (ou função logística)

$$\log\left[\frac{\mathbb{P}(Y_i=1|\mathbf{x}_i)}{\mathbb{P}(Y_i=0|\mathbf{x}_i)}\right]=\log\left(\frac{\pi_i}{1-\pi_i}\right)=\sum_{j=1}^{p}x_{ji}\beta_j,\quad i=1,...,n$$

Isolando a probabilidade resposta π_i

$$\frac{\pi_i}{1-\pi_i}=e^{\sum_{j=1}^{p}x_{ji}\beta_j},\quad i=1,...,n$$

Após algumas contas

$$\pi_i=\frac{e^{\eta_i}}{1+e^{\eta_i}},\quad i=1,...,n$$

Especificação do modelo

Se olharmos apenas para o caso em que π é um escalar temos a função de *Máxima Verossimilhança* para $y_1, ..., y_n$ seguindo uma distribuição *Bernoulli* (π) dada por

$$L(\beta) = \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i} = \pi^{\sum_{i=1}^{n} y_i}(1-\pi)^{n-\sum_{i=1}^{n}}, \quad 0 \leq \pi \leq 1.$$

A expressão do log da função acima, também chamada *log-verossimilhança* é

$$\log L(\beta) = l(\beta) = \sum_{i=1}^{n} y_i \log(\pi) + \left(n - \sum_{i=1}^{n} y_i\right) \log(1-\pi), 0 \leq \pi \leq 1.$$

Especificação do modelo

Para uma Regressão Logística π depende de outras covariáveis, $x_1, ..., x_n$, assim subtituindo o escalar π pela função $\pi(\mathbf{x}_i)$ temos

$$L(\beta) = \prod_{i=1}^{n} \pi(\mathbf{x}_i)^{y_i}[1 - \pi(\mathbf{x}_i)]^{1-y_i} = \pi(\mathbf{x}_i)^{\sum_{i=1}^{n} y_i}[1 - \pi(\mathbf{x}_i)]^{n-\sum_{i=1}^{n}}$$

$$\begin{aligned} l(\beta) = & \sum_{i=1}^{n} y_i \log[\pi(\mathbf{x}_i)] + \left(n - \sum_{i=1}^{n} y_i\right) \log[1 - \pi(\mathbf{x}_i)] \\ = & \sum_{i=1}^{n} y_i \log[\pi(\mathbf{x}_i)] + n \log[1 - \pi(\mathbf{x}_i)] - \sum_{i=1}^{n} y_i \log[1 - \pi(\mathbf{x}_i)] \\ = & \sum_{i=1}^{n} y_i \log\left[\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)}\right] + n \log[1 - \pi(\mathbf{x}_i)]. \end{aligned}$$

Especificação do modelo

sabemos que

$$\log\left[\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right] = \beta'\mathbf{x}_i$$

assim

$$\pi(\mathbf{x}_i) = \frac{e^{\beta'\mathbf{x}_i}}{1+e^{\beta'\mathbf{x}_i}}$$

$$1-\pi(\mathbf{x}_i) = \frac{1}{1+e^{\beta'\mathbf{x}_i}}$$

então

$$l(\beta) = \sum_{i=1}^{n}\left[y_i\beta'\mathbf{x}_i - \log(1+e^{\beta'\mathbf{x}_i})\right]$$

Especificação do modelo

Função Desvio (*Deviance*)

$$D = -2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{\pi}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - \hat{\pi}_i}{m_i - y_i}\right)\right],$$

onde $0 \leq y_i \leq m_i$ e, no caso, $m_i = 1, \forall t$

$$D = -2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{\pi}_i}\right) + (1 - y_i)\log\left(\frac{1 - \hat{\pi}_i}{1 - y_i}\right)\right]$$

$$\begin{aligned} D = - \quad & 2\sum_{i=1}^{n}[y_i \log(y_i) + (1 - y_i)\log(1 - y_i) - \\ - \quad & y_i \log\left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i}\right) - \log(1 - \hat{\pi}_i)\Bigg] \end{aligned}$$

como y assume apenas os valores 0 e 1 temos que

$$y_i \log(y_i) = (1 - y_i)\log(1 - y_i) = 0$$

Especificação do modelo

A seleção dos regressores pode ser feita utilizando-se Stepwise, ou através dos critérios de informação, *AIC* (*Akaike Information Criterion*) e *BIC* (*Bayesian Information Criterion*).

$$AIC = -2\frac{l(\boldsymbol{\beta})}{n} + 2\frac{p}{n}$$

$$BIC = -2\frac{l(\boldsymbol{\beta})}{n} + p\frac{\log(n)}{n}$$

onde p é o número de parâmetros, n a quantidade de observações e $l(\boldsymbol{\beta})$ é o log da função de verossimilhança

Estimação

- Máxima Verossimilhança

Derivando a primeira vez $l(\beta)$ em relação ao parâmetro β (*Função Escore*) e igualando a zero teremos

$$\begin{aligned}\frac{\partial l(\beta)}{\partial \beta} &= \sum_{i=1}^{n}\left[y_i \mathbf{x}_i - \frac{e^{\beta' \mathbf{x}_i}\mathbf{x}_i}{(1+e^{\beta' \mathbf{x}_i})}\right] \\ &= \sum_{i=1}^{n}\mathbf{x}_i\left[y_i - \pi(\mathbf{x}_i)\right]\end{aligned}$$

Na forma matricial

$$\frac{\partial l(\beta)}{\partial \beta} = \mathbf{X}'(\mathbf{y}-\boldsymbol{\pi}).$$

Estimação

O algoritmo também requer a segunda derivada (ou *Hessiano*)

$$\begin{aligned}\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta'} &= -\sum_{i=1}^{n}\left[\frac{(1+e^{\beta'\mathbf{x}_i})e^{\beta'\mathbf{x}_i}\mathbf{x}_i\mathbf{x}_i' - e^{\beta'\mathbf{x}_i}\mathbf{x}_i' e^{\beta'\mathbf{x}_i}\mathbf{x}_i'}{(1+e^{\beta'\mathbf{x_i}})^2}\right] \\ &= -\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\pi(\mathbf{x}_i)[1-\pi(\mathbf{x}_i)].\end{aligned}$$

Matricialmente

$$\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta'} = -\mathbf{X}'\mathbf{W}\mathbf{X}.$$

Estimação

A *Informação de Fisher* para β é conhecida pela expressão

$$I(\beta) = -E\left[\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta'}\right].$$

Agora, com esses elementos, falta apenas determinar um valor inicial para β, chamaremos de $\beta^{(m)}$, e assim dar início ao algoritmo de Newton-Raphson, que expande-se a Função Escore, $U(\beta)$ em torno do valor inicial de forma que

$$U(\beta) \cong U(\beta^{(m)}) + \frac{\partial}{\partial\beta'}U(\beta^{(m)})(\beta^{(m+1)} - \beta^{(m)}), \quad m = 0, 1, ...$$

Estimação

Iterativamente obtem-se

$$\beta^{(m+1)} = \beta^{(m)} + \left[-\frac{\partial}{\partial \beta'} U(\beta^{(m)}) \right]^{-1} U(\beta^{(m)}), \quad m = 0, 1, ...$$

A matriz $-\frac{\partial}{\partial \beta'} U(\beta^{(m)})$ deve ser positiva definida, e como não se pode garantir tal hipótese, substitui-se a mesma pelo seu valor esperado

$$E \left[-\frac{\partial}{\partial \beta'} U(\beta^{(m)}) \right]^{-1} = I^{-1}(\beta^{(m)})$$

e assim, continuando o processo iterativo

$$\beta^{(m+1)} = \beta^{(m)} + I^{-1}(\beta^{(m)}) U(\beta^{(m)}), \quad m = 0, 1, ...$$

Estimação

Trabalhando mais uma vez com a forma matricial em que $\mathbf{y}$ é o vetor $(n \times 1)$ de valores y_i, $\mathbf{X}$ a matriz $(n \times (p+1))$ de valores x_i, $\boldsymbol{\pi}$ o vetor $(n \times 1)$ das probabilidades ajustadas com o *i-ésimo* elemento igual a $\pi(x_i; \beta^{(m)})$ e $\mathbf{W}$ a matriz diagonal $(n \times n)$ de pesos com o *i-ésimo* elemento igual a $\pi(x_i; \beta^{(m)})(1 - \pi(x_i; \beta^{(m)}))$ tem-se

$$\begin{aligned} \beta^{(m+1)} &= \beta^{(m)} + (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \boldsymbol{\pi}) \\ &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}[\mathbf{X}\beta^{(m)} + \mathbf{W}^{-1}(\mathbf{y} - \boldsymbol{\pi})] \\ &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z} \end{aligned}$$

onde $\mathbf{z} = \mathbf{X}\beta^{(m)} + \mathbf{W}^{-1}(\mathbf{y} - \boldsymbol{\pi})$

Avaliação do ajuste

- Tabela de Classificação: Para esta análise deve-se estipular um ponto de corte, c^*, geralmente usa-se o valor 0,5. Este será comparado aos valores estimados do modelo de regressão logística, $\hat{\pi}(\mathbf{x}_i)$, e desta forma obtém-se os valores de $\hat{y}_i$ da seguinte maneira:

$$\text{se } \hat{\pi}(\mathbf{x}_i) > c^* \text{ então } \hat{y}_i = 1$$
$$\text{se } \hat{\pi}(\mathbf{x}_i) \leq c^* \text{ então } \hat{y}_i = 0.$$

Avaliação do ajuste

Abaixo encontra-se uma tabela de classificação e desta tiramos algumas medidas relevantes para avaliar o ajuste:

	Predito ($\hat{y}$)		
Observado (y)	0	1	
0	A	B	A + B
1	C	D	C + D
	A + C	B + D	A + B + C + D

Figura: Tabela de Classificação

- Taxa de acerto total: $\left(\frac{A+D}{A+B+C+D}\right) \times 100\%$
- Taxa de acertos para 0: $\left(\frac{A}{A+B}\right) \times 100\%$
- Taxa de acertos para 1: $\left(\frac{D}{C+D}\right) \times 100\%$

Sumário

GAM

A classe dos GAM´s tem como fundamento a substituição da forma linear $\sum \beta_j \mathbf{x}_j$ pela soma de funções suavizadas das variáveis explicativas, $\sum f_j(\mathbf{x}_j)$.

- Regressão Logística Aditiva

Com a substituição dos termos lineares da regressão logística pelas funções suavizadas, a expressão do modelo toma a forma

$$\log\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right] = f_0 + f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) + ... + f_p(\mathbf{x}_p)$$

Um suavizador usual e que será utilizado nas aplicações feitas nesse trabalho é o *Cubic Smoother Splines*, que faz a busca pela $f(x)$ que minimize a *Soma dos Quadrados dos Resíduos Penalizada*, denotada por

$$PRSS(f,\lambda) = \sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \int_a^b f''(t)^2 dt, \quad a \leq x_1 \leq ... \leq b$$

onde λ é o parâmetro de suavização que deve ser escolhido.

GAM

Um suavizador usual e que será utilizado nas aplicações feitas nesse trabalho é o *Cubic Smoother Splines*, que faz a busca pela $f(x)$ que minimize a *Soma dos Quadrados dos Resíduos Penalizada*, denotada por

$$PRSS(f, \lambda) = \sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \int_a^b f''(t)^2 dt, \quad a \leq x_1 \leq ... \leq b$$

onde λ é o parâmetro de suavização que deve ser escolhido.

GAM

A estimação é feita por *Máxima Verossimilhança Penalizada*. Tal método segue o mesmo princípio apresentado anteriormente onde deve-se maximizar a log-verossimilhança, guardadas as devidas alterações com a inclusão do termo penalizador como segue

$$\begin{aligned} l(f;\lambda) &= \sum_{i=1}^{n}[y_i \log \pi(x_i) + (1-y_i)\log(1-\pi(x_i))] - \frac{1}{2}\lambda \int [f''(t)]^2 dt \\ &= \sum_{i=1}^{n}[y_i f(x_i) - \log(1+e^{f(x_i)})] - \frac{1}{2}\lambda \int [f''(t)]^2 dt \end{aligned}$$

onde $\pi(x) = \mathbb{P}(Y=1|X=x)$.

k-NN

Dado um ponto x_0 no espaço n-dimensional, que deva ser classificado, encontra-se k pontos pertencentes a amostra de treinamento $(x_{(i)}, i = 1, ..., k)$ mais próximos em distância (geralmente distância Euclidiana) do novo ponto. Assim, este tem sua classificação feita de acordo com a maioria das classificações existentes de seus k vizinhos com a finalidade de formar $\hat{Y}$ que é definido como

$$\hat{Y}(x_{(i)}) = \frac{1}{k} \sum_{x_0 \in N_k(x_{(i)})} y_0,$$

onde $N_k(x_{(i)})$ é a vizinhança de $x_{(i)}$ definida pelos k pontos amostrais próximos de x_0

k-NN

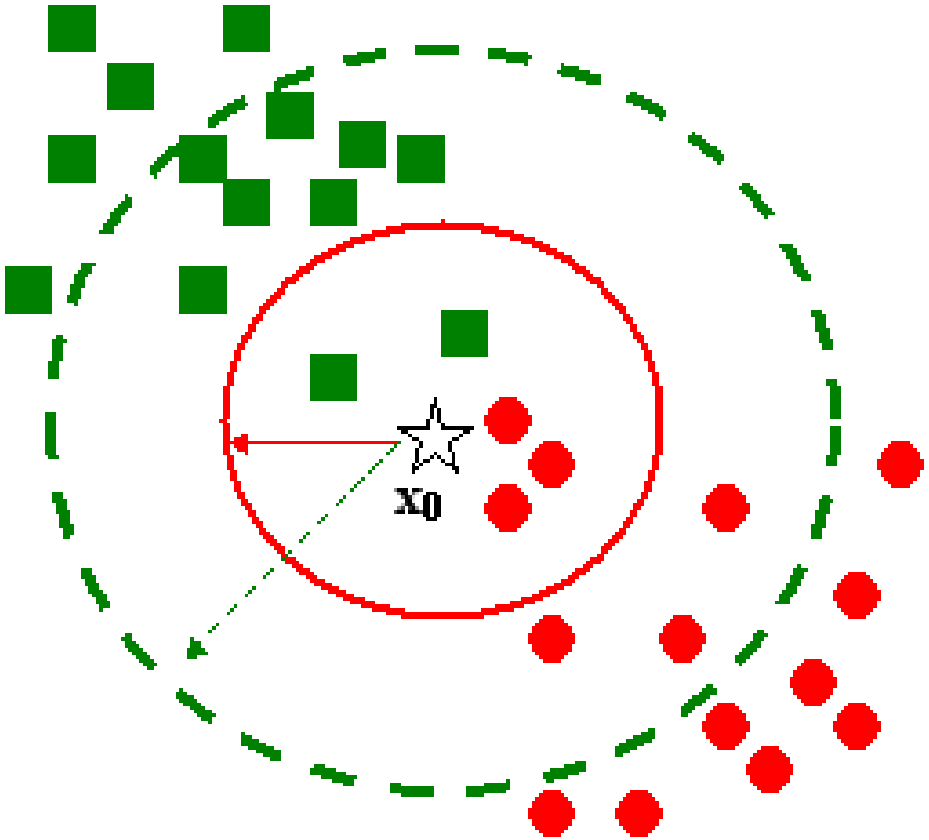

Figura: Exemplo: k-Nearest Neighbor

Análise Discriminante

A Análise Discriminante é uma metodologia que permite classificar duas ou mais populações e com esta separação prévia poder alocar um novo objeto a uma das classes existentes. Para tal é calculada uma função, que é a combinação linear das covariáveis, denominada *função discriminante*. Os principais pressupostos desta função são: a variável dependente deve seguir uma distribuição Normal multivariada e as matrizes de covariância (Σ) sejam iguais.

Sumário

Formulação Matemática

Seja $\mathbf{z}_i \subseteq \mathbf{x}_i$ tal que $\mathbf{x}_i \in \mathbb{X} \subseteq \mathbb{R}^q$ e $\mathbf{z}_i \in \mathbb{R}^p$ onde $p \leq q$. Considere $\tilde{\mathbf{z}}_i = (1, \mathbf{z}_i)'$. Um modelo paramétrico $\mathcal{M}$ definido pela função $H_{\mathbb{J}\mathbb{T}}(\mathbf{x}_i; \psi) : \mathbb{R}^{q+1} \rightarrow \mathbb{R}$, indexado pelo vetor de parâmetros $\psi \in \mathbf{\Psi}$, um subconjunto compacto do espaço Euclidiano, é chamado modelo Smooth Transition Regression Tree (STR-Tree) se

$$y_i = H_{\mathbb{J}\mathbb{T}}(\mathbf{x}_i; \psi) = \sum_{k \in \mathbb{T}} \beta_k' \tilde{\mathbf{z}}_i B_{\mathbb{J}_k}(\mathbf{x}_i; \boldsymbol{\theta}_k) + \epsilon_i$$

onde

$$B_{\mathbb{J}_k}(\mathbf{x}_i; \boldsymbol{\theta}_k) = \prod_{j \in \mathbb{J}} G(x_{s_j,i}; \gamma_j, c_j)^{\frac{n_{k,j}(1+n_{k,j})}{2}} \left[1 - G(x_{s_j,i}; \gamma_j, c_j)\right]^{(1-n_{k,j})(1+n_{k,j})}.$$

Formulação Matemática

$$G(x_i;\gamma,c) = \frac{1}{1+e^{-\gamma(x_i-c)}}$$

Vamos considerar agora um STR-Tree em uma árvore completamente crescida, ou cheio, com profundidade d, $K = 2d$, nós terminais (folhas) definido como

$$y_i = \sum_{k=1}^{K} (\alpha_{K+k-2} + \beta_{K+k-2}\mathbf{z}_i) B_k(\mathbf{x}_i;\boldsymbol{\theta}_k) + \epsilon_i.$$

O vetor de parâmetros
$\boldsymbol{\psi} = (\alpha_{K-1},...,\alpha_{2K-2},\beta_{K-1},...,\beta_{2K-2},\boldsymbol{\theta}_1',...,\boldsymbol{\theta}_K')' \in \mathbb{R}^r$ possui
$r = (p+1)K + 2N$ elementos onde N é o número ne nós intermediários.

Exemplo

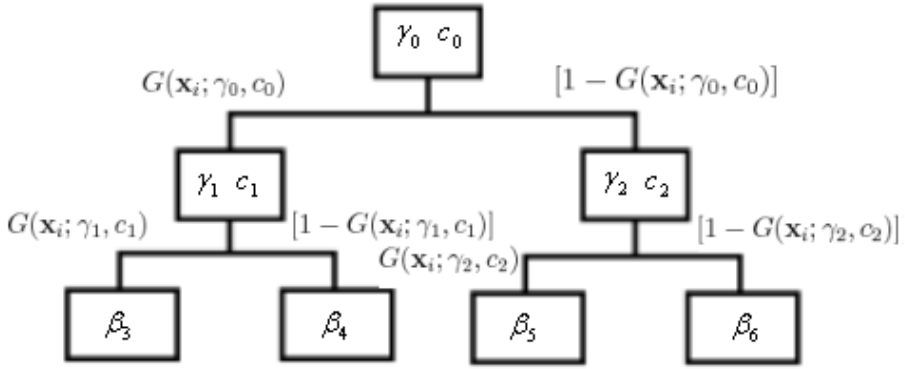

Figura: Exemplo Árvore

$$
\begin{aligned}
y_i \quad &= \{(\alpha_3+\beta_3\mathbf{z}_i)G(\mathbf{x}_i;\gamma_1,c_1)+(\alpha_4+\beta_4\mathbf{z}_i) \\
&[1-G(\mathbf{x}_i;\gamma_1,c_1)]\}G(\mathbf{x}_i;\gamma_0,c_0)+ \\
&+\{(\alpha_5+\beta_5\mathbf{z}_i)G(\mathbf{x}_i;\gamma_2,c_2)+ \\
&+(\alpha_6+\beta_6\mathbf{z}_i)[1-G(\mathbf{x}_i;\gamma_2,c_2)]\} \\
&[1-G(\mathbf{x}_i;\gamma_0,c_0)]+\epsilon_i
\end{aligned}
$$

Exemplo

$$
\begin{aligned}
y_i \quad &= (\alpha_3 + \beta_3 \mathbf{z}_i) G(\mathbf{x}_i; \gamma_0, c_0) G(\mathbf{x}_i; \gamma_1, c_1) + \\
&+ (\alpha_4 + \beta_4 \mathbf{z}_i) G(\mathbf{x}_i; \gamma_0, c_0)[1 - G(\mathbf{x}_i; \gamma_1, c_1)] \\
&+ (\alpha_5 + \beta_5 \mathbf{z}_i)[1 - G(\mathbf{x}_i; \gamma_0, c_0)] G(\mathbf{x}_i; \gamma_2, c_2) \\
&+ (\alpha_6 + \beta_6 \mathbf{z}_i)[1 - G(\mathbf{x}_i; \gamma_0, c_0)][1 - G(\mathbf{x}_i; \gamma_2, c_2)] + \epsilon_i
\end{aligned}
$$

Com isso podemos deduzir que

$$
\begin{aligned}
B_1(\mathbf{x}_i; \boldsymbol{\theta}_1) &= G(\mathbf{x}_i; \gamma_0, c_0) G(\mathbf{x}_i; \gamma_1, c_1) \\
B_2(\mathbf{x}_i; \boldsymbol{\theta}_2) &= G(\mathbf{x}_i; \gamma_0, c_0)[1 - G(\mathbf{x}_i; \gamma_1, c_1)] \\
B_3(\mathbf{x}_3; \boldsymbol{\theta}_3) &= [1 - G(\mathbf{x}_i; \gamma_0, c_0)] G(\mathbf{x}_i; \gamma_2, c_2) \\
B_4(\mathbf{x}_4; \boldsymbol{\theta}_4) &= [1 - G(\mathbf{x}_i; \gamma_0, c_0)][1 - G(\mathbf{x}_i; \gamma_2, c_2)]
\end{aligned}
$$

Especificação do modelo

Assim

$$
\begin{aligned}
y_i \quad &= (\alpha_3 + \beta_3)\mathbf{z}_i B_1(\mathbf{x}_i;\boldsymbol{\theta}_1) + (\alpha_4 + \beta_4)\mathbf{z}_i B_2(\mathbf{x}_i;\boldsymbol{\theta}_2) \\
&+(\alpha_5 + \beta_5)\mathbf{z}_i B_3(\mathbf{x}_3;\boldsymbol{\theta}_3) + (\alpha_6 + \beta_6)\mathbf{z}_i B_4(\mathbf{x}_4;\boldsymbol{\theta}_4) + \epsilon_i
\end{aligned}
$$

$$
y_i = \sum_{k=1}^{4}(\alpha_{K+k-2} + \beta_{K+k-2}\mathbf{z}_i)B_k(\mathbf{x}_i;\boldsymbol{\theta}_k) + \epsilon_i.
$$

Levando em consideração que nossa variável dependente seja dicotômica, podemos escrever o modelo na forma de uma Regressão Logística, utilizando como função de ligação a logito, denominado STLR-Tree, como se segue

$$
\log\left[\frac{\pi(\mathbf{z}_i)}{1-\pi(\mathbf{z}_i)}\right] = \sum_{k=1}^{K}(\alpha_{K+k-2} + \beta_{K+k-2}\mathbf{z}_i)B_k(\mathbf{x}_i;\boldsymbol{\theta}_k),
$$

Especificação do Modelo

$$\pi(\mathbf{z}_i) = \frac{e^{\sum_{k=1}^{K}(\alpha_{K+k-2}+\beta_{K+k-2}\mathbf{z}_i)B_k(\mathbf{x}_i;\boldsymbol{\theta}_k)}}{1+e^{\sum_{k=1}^{K}(\alpha_{K+k-2}+\beta_{K+k-2}\mathbf{z}_i)B_k(\mathbf{x}_i;\boldsymbol{\theta}_k)}}. \tag{2}$$

Por fim, temos a função de log-verossimilhança do STLR-Tree.

$$\begin{aligned} l(\boldsymbol{\beta}) \quad = \sum_{i=1}^{n}\Bigg[& y_i \sum_{k=1}^{K}(\alpha_{K+k-2}+\beta_{K+k-2}\mathbf{z}_i)B_k(\mathbf{x}_i;\boldsymbol{\theta}_k) - \\ & - \log\left(1+e^{\sum_{k=1}^{K}(\alpha_{K+k-2}+\beta_{K+k-2}\mathbf{z}_i)B_k(\mathbf{x}_i;\boldsymbol{\theta}_k)}\right)\Bigg] \end{aligned}$$

Especificação do Modelo

- Escolha das variáveis relevantes

No caso o STLR-Tree será restrito para variáveis contínuas independentes. Para a escolha dos elementos de $\mathbf{z}_i$ destacam-se três métodos:

1. por critérios de informação, AIC e BIC, os quais já foram descritos, sendo o melhor modelo aquele que minimiza tais critérios;
2. aproximação polinomial do modelo;
3. outra opção é dada através de técnicas não paramétricas, porém esta classe é computacionalmente dispendiosa, principalmente para um grande número de observações.

Especificação do Modelo

- Escolha do nó a ser dividido

Estamos testando a linearidade do modelo, que nada mais é do que assumi-lo como

$$\log\left[\frac{\pi(\mathbf{z}_i)}{1-\pi(\mathbf{z}_i)}\right] = \sum_{k=1}^{K}(\alpha_{K+k-2} + \beta_{K+k-2}\mathbf{z}_i),$$

excluindo-se a parte não-linear $B_k(\mathbf{x}_i;\boldsymbol{\theta}_k)$.

Especificação do Modelo

Mais uma vez para o exemplo em que temos $d = 1, K = 2$ e $N = 1$ e fazendo uma reparametrização para podemos deixar o modelo em uma forma reduzida

$$\log\left[\frac{\pi(\mathbf{z}_i)}{1-\pi(\mathbf{z}_i)}\right] = \phi_0 + \lambda_0 G(\mathbf{x}_i; \gamma_0, c_0),$$

onde $\phi_0 = \alpha_2 + \beta_2 \mathbf{z}_i$ e $\lambda_0 = [\alpha_1 - \alpha_2 + \beta_1 \mathbf{z}_i - \beta_2 \mathbf{z}_i] G(\mathbf{x}_i; \gamma_0, c_0)$
Deve-se testar a hipótese de significância da primeira divisão

$$\begin{cases} H_0 : \gamma_0 = 0 \\ H_1 : \gamma_0 > 0 \end{cases}$$

Especificação do Modelo

Porém, sob H_0 temos que enfrentar um problema de especificação do modelo, pois os parâmetros γ e c podem assumir diferentes valores sem alterar a função de verossimilhança.

- Solução: Expansão de Taylor de terceira ordem em torno de $\gamma = 0$ e assim, após manipulações algébricas, podemos reescrever o modelo como

$$\log\left[\frac{\pi(\mathbf{z}_i)}{1-\pi(\mathbf{z}_i)}\right] = \alpha_0 + \alpha_1\mathbf{z}_i x_{s_0 t} + \alpha_2\mathbf{z}_i x_{s_0 t}^2 + \alpha_3\mathbf{z}_i x_{s_0 t}^3,$$

Assim podemos reescrever a hipótese de nulidade dos parâmetros,
$H_0 : \alpha_i = 0, \quad i = 1, 2 \text{ e } 3$

Especificação do Modelo

A seqüência de teste de Razão de Verossimilhança comparando o modelo sob H_0 contra o modelo sob H_1 é realizada através da estatística de teste abaixo. Como há equivalência entre a soma dos quadrados dos resíduos de uma regressão linear Gaussiana e a função desvio, podendo assim, esta última, ser utilizada.

$$LM = \frac{\left[D(\mathbf{y};\hat{\boldsymbol{\mu}}^0) - D(\mathbf{y};\hat{\boldsymbol{\mu}})\right]/3(p+1)}{D(\mathbf{y};\hat{\boldsymbol{\mu}})/[N-3(p+1)-(p+1)]} \xrightarrow{a} F_{3(p+1),N-4(p+1)}.$$

Especificação do Modelo

- Escolha das variáveis de transição

Aplicar os testes de RV para cada uma das variáveis explicativas e selecionar a variável xs_0t que gere o menor *p-valor*, sob um nível de significância α. Sabe-se que $s_0 \in \mathbb{S} = \{1, 2, ..., m\}$ é o conjunto dos índices dos elementos em $\mathbf{x}_i$.

Estimação

- Máxima Verossimilhança

Fixando os parâmetros não-lineares, γ e c, o vetor de parâmetros, $\boldsymbol{\beta}$, é estimado por Máxima Verossimilhança, que necessita da utilização do processo iterativo de Newton-Raphson.

$$\beta^{(m+1)} = [\mathbf{B}(\boldsymbol{\theta})'\mathbf{W}^{(m)}\mathbf{B}(\boldsymbol{\theta})]^{-1}\mathbf{B}(\boldsymbol{\theta})'\mathbf{W}^{(m)}\mathbf{z}^{(m)}$$

onde $\mathbf{z} = \mathbf{B}(\boldsymbol{\theta})\beta^{(m)} + \mathbf{W}^{-1}(\mathbf{y} - \boldsymbol{\pi})$ e

$$\mathbf{B}(\boldsymbol{\theta}) = \begin{pmatrix} B_1(\mathbf{x}_1;\boldsymbol{\theta}_1) & \dots & B_K(\mathbf{x}_1;\boldsymbol{\theta}_K) \\ \vdots & \ddots & \vdots \\ B_1(\mathbf{x}_N;\boldsymbol{\theta}_1) & \dots & B_K(\mathbf{x}_N;\boldsymbol{\theta}_K) \end{pmatrix}$$

Estimação

Após estimados os parâmetros da parte linear, faz-se a estimação de γ e c também por Máxima Verossimilhança usando as estimativas de $\boldsymbol{\beta}$ nos cálculos.
A estimação dos parâmetros lineares e não-lineares seguem basicamente o seguinte:

1. Encontra-se os valores iniciais dos parâmetros não-lineares, γ e c, através de uma busca em *grid*, ao maximizar a log-verossimilhança concentrada;
2. Aplica-se os valores encontrados no item anterior e estima-se por Máxima Verossimilhança os parâmetros lineares, $\boldsymbol{\beta}$;
3. As estimativas de $\boldsymbol{\beta}$ são usadas na estimação de γ e c também por Máxima Verossimilhança;
4. Alternar os dois passos anteriores.

Ciclo de modelagem

- Criação da Primeira profundidade (a partir de $d = 0$)

Escolher as variável $x_{s_0 i}$ que gere o menos *p-valor*. Dado $s_0 \in \mathbb{S} = \{1, 2, ..., m\}$ é feita a estimação do vetor de parâmetros, $\psi = (\gamma_0, c_0, \beta_1, \beta_2)'$ conforme os métodos de estimação especificados anteriormente e testa-se as hipóteses

$$\begin{cases} H_{01} : \beta_1 = 0 \\ H_{02} : \beta_2 = 0 \\ H_{03} : \beta_1 - \beta_2 = 0 | \beta_1, \beta_2 \neq 0 \end{cases}$$

Ciclo de modelagem

Após a profundidade d=0, o nível de significância igual a α, é penalizado da seguinte maneira

$$\alpha(d, n) = \frac{\alpha}{n^d}$$

a partir daí, na primeira profundidade ($d = 1$), são aplicados dois testes ($n = 2$) e o nível de significância passa a valer $\frac{\alpha}{2}$. A evolução de α de acordo com a profundidade e o número de testes se dá a partir do nó raíz como: $\alpha, \frac{\alpha}{2}, \frac{\alpha}{3}, \frac{\alpha}{4^2}, \frac{\alpha}{5^2}, \frac{\alpha}{6^3}, \frac{\alpha}{7^3}, \frac{\alpha}{8^4}, ...$

Ciclo de modelagem

- Criação da Primeira profundidade (a partir de $d = 1$)

Selecionar o par de combinações entre o índice da variável de transição em $\mathbb{S} = \{1, 2, ..., m\}$ e o número do nó em $\mathbb{D} = \{1, 2\}$ que minimize o *p-valor*. Sendo assim, estima-se os parâmetros e testa-se a significância da divisão através das hipóteses

$$\begin{cases} H_{01} : \beta_{2j_1+1} = 0 \\ H_{02} : \beta_{2j_1+2} = 0 \\ H_{03} : \beta_{2j_1+1} - \beta_{2j_1+2} = 0 | \beta_{2j_1+1}, \beta_{2j_1+2} \neq 0. \end{cases} \tag{3}$$

Caso os dois nós da primeira profundidade gerem mais dois nós filhotes teremos na segunda profundidade 4 nós que serão: $2j_1 + 1, 2j_1 + 2, 2j_2 + 1$ e $2j_2 + 2$. Por outro lado, se nenhum dos dois nós gerarem divisões significativas paramos o crescimento da árvore e fazemos a avaliação do ajuste.

Ciclo de modelagem

- Criação da k-ésima profundidade

Para cada combinação $\{j_k; s_{j_k}\}$, de nó e variável de transição, é aplicado o teste de RV confrontando-os com o modelo apenas linear. As variáveis de transição pertencem ao conjunto $\mathbb{S} = \{1, 2, ..., m\}$ enquanto que os nós estão em $\mathbb{D}_k = \{2^k - 1, 2^k, ..., 2^{k+1} - 2\}$ e seleciona-se $j_k \in \mathbb{D}_k$ e $s_{j_k} \in \mathbb{S}$ que gere o menor *p-valor*. Com isso em mãos estima-se os parâmetros do modelo.

Sumário

Bases

- E-mail/Spam: contém dados de 4601 mensagens de e-mail, em que a variável dependente é igual a 0 se a mensagem foi considerada um e-mail de fato ou 1 caso tinha sido caracterizada como um *spam*;
- Doenças Cardíacas na África do Sul (DCAS): possui informações de 462 indivíduos homens com idades entre 15 e 64 anos;
- Fraude/Irregularidade no Consumo de Energia Elétrica: a empresa possui cerca de 452 mil clientes inspecionados em baixa tensão com perfis de consumo de energia diferentes, distribuídos em 2 regiões de estudo (Leste e Oeste). Essas regiões estão subdivididas e foi uma dessas subdivisões que selecionamos nesse exemplo. Ela possui 2430 clientes que são classificados através de uma variável binária com o valor 0 para os clientes normais e 1 para os supostos clientes irregulares.

Bases

- E-mail/Spam: contém dados de 4601 mensagens de e-mail, em que a variável dependente é igual a 0 se a mensagem foi considerada um e-mail de fato ou 1 caso tinha sido caracterizada como um *spam*;
- Doenças Cardíacas na África do Sul (DCAS): possui informações de 462 indivíduos homens com idades entre 15 e 64 anos;
- Fraude/Irregularidade no Consumo de Energia Elétrica: a empresa possui cerca de 452 mil clientes inspecionados em baixa tensão com perfis de consumo de energia diferentes, distribuídos em 2 regiões de estudo (Leste e Oeste). Essas regiões estão subdivididas e foi uma dessas subdivisões que selecionamos nesse exemplo. Ela possui 2430 clientes que são classificados através de uma variável binária com o valor 0 para os clientes normais e 1 para os supostos clientes irregulares.

Bases

- E-mail/Spam: contém dados de 4601 mensagens de e-mail, em que a variável dependente é igual a 0 se a mensagem foi considerada um e-mail de fato ou 1 caso tinha sido caracterizada como um *spam*;
- Doenças Cardíacas na África do Sul (DCAS): possui informações de 462 indivíduos homens com idades entre 15 e 64 anos;
- Fraude/Irregularidade no Consumo de Energia Elétrica: a empresa possui cerca de 452 mil clientes inspecionados em baixa tensão com perfis de consumo de energia diferentes, distribuídos em 2 regiões de estudo (Leste e Oeste). Essas regiões estão subdivididas e foi uma dessas subdivisões que selecionamos nesse exemplo. Ela possui 2430 clientes que são classificados através de uma variável binária com o valor 0 para os clientes normais e 1 para os supostos clientes irregulares.

E-mail/Spam

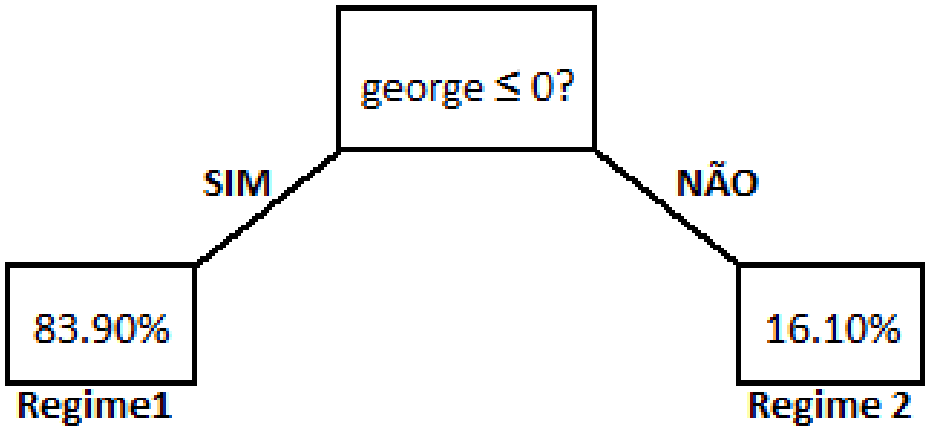

Figura: Estrutura do modelo - Spam

Observado (y)	Predito (ŷ)		
	0	1	
0	1142	70	1212
1	105	683	788
	1247	753	2000

Figura: Tabela de Classificação - Spam

E-mail/Spam

	Tx. de acertos Total	Tx. de acertos para 1 (Sensitividade)	Tx. de acertos para 0 (Especificidade)
GAM	95.20%	92.26%	97.11%
STLR-Tree	91.25%	86.68%	94.22%
Reg. Logística	91.05%	86.80%	93.81%
CART	89.95%	80.46%	96.12%
k-NN	89.21%	81.32%	94.15%
Análise Discrim.	87.75%	78.17%	93.98%

Figura: Comparação das Taxas de Acerto - Spam

Tx. de acertos Total	Tx. de acertos para 1 (Sensitividade)	Tx. de acertos para 0 (Especificidade)
GAM (0.952)	GAM (0.922)	GAM (0.971)
STLR-Tree (0.912)	Reg. Logística (0.86)	CART (0.961)
Reg. Logística (0.910)	STLR-Tree (0.866)	STLR-Tree (0.942)
CART (0.899)	k-NN (0.813)	k-NN (0.941)
k-NN (0.892)	CART (0.804)	Análise Discrim. (0.939)
Análise Discrim. (0.877)	Análise Discrim. (0.781)	Reg. Logística (0.938)

Figura: Métodos de Classificação Ordenados por Taxas de Acerto - Spam

E-mail/Spam

O ajuste a esses dados gerou uma árvore com 2 nós terminais e profundidade 1, contra um CART com 13 nós terminais e profundidades igual a 7

	GAM	Regressão Logística	Análise Discrimina	STLR-Tree (REGIME 1)	STLR-Tree (REGIME 2)
Intercepto	-1.3979	-1.577	-	-0.391	-0.014
our	0.0110	0.010	0.250	-3.153	11.598
over	0.0184	0.018	0.243	-10.772	-7.960
remove	0.0350	0.042	0.414	9.870	12.162
internet	0.0157	0.020	0.243	-5.669	-1.185
free	0.0141	0.015	0.257	3.111	2.307
business	0.0130	0.012	0.142	0.037	-0.962
hp	-0.0313	-0.025	-0.205	2.401	-5.604
hpl	0.0020	-0.019	-0.122	1.802	0.145
george	-0.0428	-0.060	-0.210	-2.879	-328.586
1999	-	-	-0.148	-	-
remove	-0.0164	-	-0.149	-7.036	13.064
edu	-0.0198	-0.014	-0.168	-14.609	-7.917
char_!	0.0291	-0.034	0.266	3.451	-18.344
char_$	0.0819	0.027	0.328	19.250	14.834
CAPMAX	-0.0051	0.096	0.125	1.528	0.014
CAPTOT	0.0002	-0.003	0.282	0.017	0.017

Figura: Coeficientes - Spam

E-mail/Spam

$$
\begin{aligned}
Regime \quad 1 = \quad & (-0.39 - 3.15our - 10.77over + 9.87remove - \\
- \quad & 5.67internet + 3.11free + 0.04business + 2.4hp + \\
+ \quad & 1.8hpl - 2.88george - 7.04remove - 14.61edu + \\
+ \quad & 3.45char\_! + 19.25char\_\$ + \\
+ \quad & 1.53CAPMAX + 0.02CAPTOT)G(george; 0, 50)
\end{aligned}
$$

$$
\begin{aligned}
Regime \quad 2 = \quad & (-0.01 + 11.6our - 7.96over + 12.16remove - \\
- \quad & 1.18internet + 2.31free - 0.96business - 5.6hp + \\
+ \quad & 0.15hpl - 328.59george + 13.06remove - 7.92edu - \\
+ \quad & 18.34char\_! + 14.83char\_\$ + \\
+ \quad & 0.01CAPMAX + 0.02CAPTOT)[1 - G(george; 0, 50)]
\end{aligned}
$$

DCAS

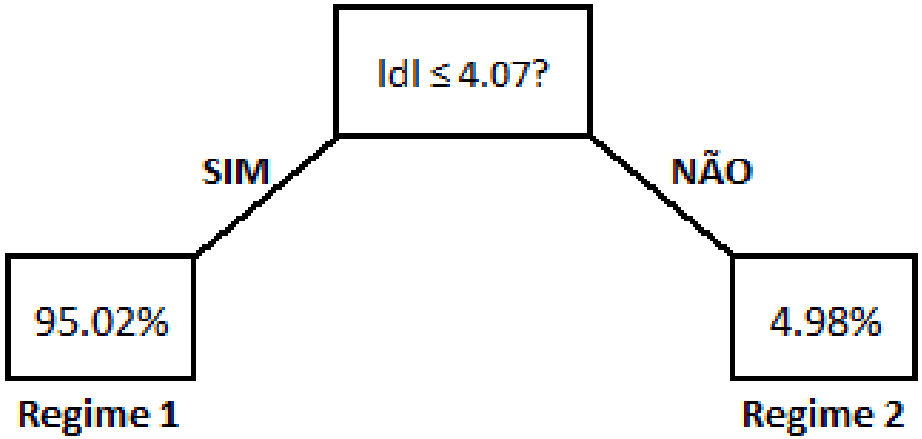

Figura: Estrutura do modelo - DCAS

Observado (y)	Predito (ŷ)		
	0	1	
0	268	34	302
1	85	75	160
	353	109	462

Figura: Tabela de Classificação - DCAS

DCAS

	Tx. de acertos Total	Tx. de acertos para 1 (Sensitividade)	Tx. de acertos para 0 (Especificidade)
GAM	78.35%	60.63%	87.75%
STLR-Tree	74.24%	46.88%	88.74%
CART	72.94%	50.63%	84.77%
Reg. Logística	70.78%	47.50%	83.11%
Análise Discrim.	69.05%	71.88%	67.55%
K-NN	66.23%	56.52%	70.37%

Figura: Comparação das Taxas de Acerto - DCAS

Tx. de acertos Total	Tx. de acertos para 1 (Sensitividade)	Tx. de acertos para 0 (Especificidade)
GAM (0.783)	Análise Discrim. (0.718)	STLR-Tree (0.887)
STLR-Tree (0.742)	GAM (0.606)	GAM (0.877)
CART (0.729)	K-NN (0.565)	CART (0.847)
Reg. Logística (0.707)	CART (0.506)	Reg. Logística (0.831)
Análise Discrim. (0.690)	Reg. Logística (0.47)	K-NN (0.703)
K-NN (0.662)	STLR-Tree (0.468)	Análise Discrim. (0.675)

Figura: Métodos de Classificação Ordenados por Taxas de Acerto - DCAS

DCAS

O ajuste a esses dados gerou um modelo com 2 nós terminais e uma profundidade, contra um CART com 6 nós terminais e profundidades igual a 3

	GAM	Regressão Logística	Análise Discrimina	STLR-Tree (REGIME 1)	STLR-Tree (REGIME 2)
Intercepto	-4.1713	-4.3762	-	130.491	-3.629
sbp	0.0067	0.0051	0.104	-3.583	0.002
tobacco	0.0027	0.0033	0.389	2.664	0.081
ldl	0.0029	0.0026	0.393	6.386	-0.004
alcohol	-0.0003	-0.0001	0.016	-8.911	0.002
k-NN	0.0440	0.0514	0.563	-3.926	0.051

Figura: Coeficientes - DCAS

DCAS

$$
\begin{aligned}
\textit{Regime} \quad 1 = \quad & (130.491 - 3.583\textit{sbp} + 2.664\textit{tobacco} + 6.386\textit{ldl} - \\
- \quad & 8.911\textit{alcohol} - 3.926\textit{age})G(\textit{ldl}; 4.0757, 50)
\end{aligned}
$$

$$
\begin{aligned}
\textit{Regime} \quad 2 = \quad & (-3.629 + 0.002\textit{sbp} + 0.081\textit{tobacco} - 0.004\textit{ldl} + \\
+ \quad & 0.002\textit{alcohol} + 0.051\textit{age})[1 - G(\textit{ldl}; 4.0757, 50)]
\end{aligned}
$$

Consumo de Energia Elétrica

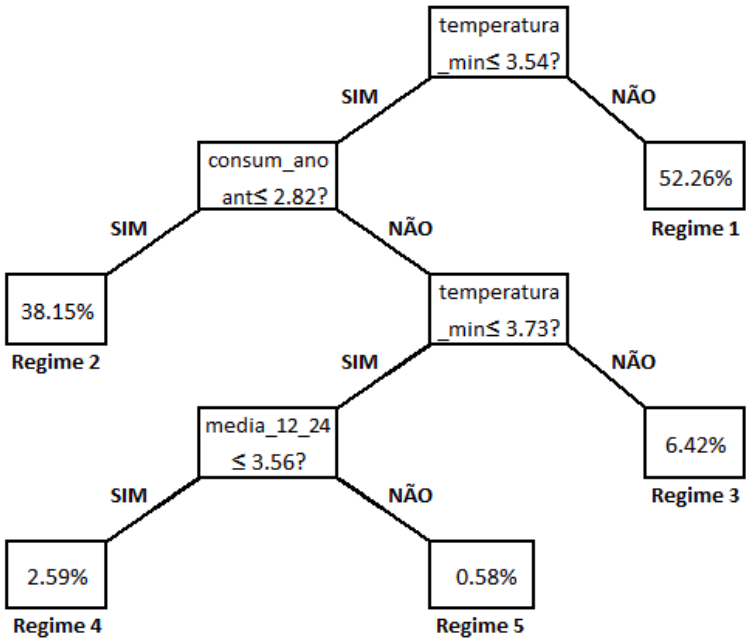

Figura: Estrutura do modelo - Consumo de Energia

Observado (y)	Predito (ŷ)		
	0	1	
0	775	440	1215
1	397	818	1215
	1172	1258	2430

Figura: Tabela de Classificação - Consumo de Energia

Consumo de Energia Elétrica

	Tx. de acertos Total	Tx. de acertos para 1 (Sensitividade)	Tx. de acertos para 0 (Especificidade)
Redes Neurais	71.77%	69.38%	74.16%
GAM	66.58%	64.20%	68.97%
STLR-Tree	65.56%	67.33%	63.79%
Análise Discrim.	60.16%	58.85%	61.48%
CART	59.92%	29.71%	90.12%
Reg. Logística	59.79%	57.20%	62.39%
K-NN	54.94%	56.90%	53.41%

Figura: Comparação das Taxas de Acerto - Consumo de Energia

Tx. de acertos Total	Tx. de acertos para 1 (Sensitividade)	Tx. de acertos para 0 (Especificidade)
Redes Neurais (0.717)	Redes Neurais (0.693)	CART (0.901)
GAM (0.665)	STLR-Tree (0.673)	Redes Neurais (0.741)
STLR-Tree (0.655)	GAM (0.642)	GAM (0.689)
Análise Discrim. (0.601)	Análise Discrim. (0.588)	STLR-Tree (0.637)
CART (0.599)	Reg. Logística (0.57)	Reg. Logística (0.623)
Reg. Logística (0.597)	K-NN (0.56)	Análise Discrim. (0.614)
K-NN (0.549)	CART (0.297)	K-NN (0.534)

Consumo de Energia Elétrica

Sua estrutura foi amaior dentre todas as estruturas dos demais exemplos, tendo 5 nós terminais e profundidade igual a 4. A estrutura do CART tem 4 nós terminais e profundidade 3

	GAM	Regressão Logística	Análise Discrimina	STLR-Tree (REGIME 1)	STLR-Tree (REGIME 2)	STLR-Tree (REGIME 3)	STLR-Tree (REGIME 4)	STLR-Tree (REGIME 5)
Intercepto	0.52	-1.23	-	-2.38	-0.54	-3.86	33.85	-16.84
consumo	1.47	-	0.14	6.40	33.76	7.82	-35.36	-0.69
consumo_ano_ant	0.27	-	0.19	-1.48	-23.65	-1.65	-32.45	-0.23
consumo_ano_base	0.00	0.00	0.05	35.20	-6.64	1.03	15.22	-0.01
media_3	-0.83	-	-0.44	0.28	0.46	-0.11	0.98	0.98
media_6	-2.76	-1.55	-0.55	0.14	-7.70	-0.03	-19.19	-2.61
media_12	0.98	1.13	0.48	-0.78	-0.12	-3.32	6.09	1.51
media_12_24	-0.45	-0.77	-0.36	0.03	-1.54	6.50	40.83	-0.06
indic_trimestral_1	1.24	0.60	0.35	-	-	-	-	-
indic_trimestral_2	1.04	0.61	0.27	-	-	-	-	-
indic_anual	-0.34	-	0.10	-1.51	0.15	-5.74	-75.25	-0.72
indic_ajuste	-0.19	-	-0.34	-0.55	3.56	1.96	35.18	-0.24
indic_tendencia	0.14	-	-0.08	2.79	-4.65	-0.92	-25.79	-0.15
temperatura_min	-	5.27	1.34	-1.62	3.62	0.60	-18.00	1.90
temperatura_max	-	-	0.01	0.95	-0.59	-0.36	-29.10	0.34
carga	-	-2.76	-0.70	0.59	0.74	-4.70	56.38	-1.23

Figura: Coeficientes - Consumo de Energia

Sumário

Conclusão

- Os resultados obtidos pelo STLR-Tree foram bastante satisfatórios, sendo ele detentor da segunda melhor taxa de acerto total nas aplicações: *Spam* e *DCAS*, perdendo apenas para a classificação obtida pelo GAM. Ambos provavelmente conseguiram captar uma relação não-linear entre as variáveis que os demais não fizeram. Na aplicação *Consumo de Energia* o modelo ficou atrás apenas atrás de Redes Neurais e GAM para a mesma taxa. Atentamos para o fato deste último exemplo ter sido trabalhado e ajustado especificamente em um estudo sobre Redes Neurais.

Conclusão

- Além disso, o modelo se apresentou bastante parcimonioso nos dois primeiros exemplos, não passando de dois nós terminais na estrutura das árvores. Como comparação, temos o CART que se estruturou com 13 nós terminais para os exemplos da base *Spam* e 6 nós terminais para a base *DCAS*. Já no último exemplo, *Consumo de Energia*, o número de nós terminais do STLR-Tree foi maior do que o CART, 5 e 4 nós respectivamente. Entretanto o primeiro foi bastante superior em todas as taxas de acertos avaliadas (total, para 1 (sucesso) e para 0 (fracasso)), obtendo um acerto total de 65.56% contra 59.92%.
- Concluímos com isto, que o modelo se adaptou bem as alterações realizadas, mostrando-se uma alternativa a ser utilizada para classificação, devendo ainda ter seu desempenho computacional melhorado, objetivando minimizar o tempo de duração de suas aplicações.

Conclusão

- Além disso, o modelo se apresentou bastante parcimonioso nos dois primeiros exemplos, não passando de dois nós terminais na estrutura das árvores. Como comparação, temos o CART que se estruturou com 13 nós terminais para os exemplos da base *Spam* e 6 nós terminais para a base *DCAS*. Já no último exemplo, *Consumo de Energia*, o número de nós terminais do STLR-Tree foi maior do que o CART, 5 e 4 nós respectivamente. Entretanto o primeiro foi bastante superior em todas as taxas de acertos avaliadas (total, para 1 (sucesso) e para 0 (fracasso)), obtendo um acerto total de 65.56% contra 59.92%.
- Concluímos com isto, que o modelo se adaptou bem as alterações realizadas, mostrando-se uma alternativa a ser utilizada para classificação, devendo ainda ter seu desempenho computacional melhorado, objetivando minimizar o tempo de duração de suas aplicações.