3
Modelos e metodologias comparadas

Este capítulo tem o propósito de listar algumas das alternativas existentes na literatura que envolve classificação, e serão utilizadas neste trabalho sendo comparadas ao modelo STLR-Tree. A maioria delas está resumida e outras bem detalhadas em (14), que ilustra com muitos exemplos suas aplicações e comparações. A seção referente à Regressão Logística não foi colocada neste capítulo, pois a mesma aparece bem detalhada no capítulo 2.

3.1
Classification and Regression Trees (CART)

Uma breve revisão da estrutura em árvore, seguindo o algoritmo CART (*Classification and Regression Trees*) em (2), onde foram unificados todos os métodos de árvores de regressão e classificação existentes no período, será feita sobre sua formulação matemática, a fim de melhor entender a estrutura do STLR-Tree apresentada posteriormente.

A distinção entre as árvores de classificação e regressão é feita de acordo com o tipo de variável dependente. Quando a variável é contínua, utiliza-se árvores de regressão e no caso de variáveis categóricas, árvores de classificação. Por não fazerem suposições sobre componentes aleatórias e sobre a forma funcional do modelo, tão pouco assumirem a existência de modelos probabilísticos, tal como acontece nos modelos estatísticos de regressão e classificação, as árvores são tidas como métodos não-paramétricos para tais fins.

De fácil entendimento, as árvores particionam de forma recursiva o espaço das covariáveis, $\mathbb{X}$. Sua estrutura é simples e usualmente são representadas e ajustadas em um gráfico que cresce de um nó inicial (ou nó raiz), que é determinado como posição 0, 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 $2j + 2$, e assim progressivamente, até que os nós geradores não sejam mais divididos, quando passam a ser chamados de nós terminais.

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

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

Seguindo (17), como foi citado e (7) 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, $H(\mathbf{x}_i; \boldsymbol{\psi})$ de $\mathbf{x}_i$ e definida pelo vetor de parâmetros $\boldsymbol{\psi} \in \mathbb{R}^r$ onde r é o número total de parâmetros do modelo.

A partição é usualmente definida por um conjunto de hiperplanos ortogonais aos eixos das variáveis explicativas, chamada de *variável de transição* (em inglês: *split variable*).

No contexto apresentado em (2), $H(\mathbf{x}_i; \boldsymbol{\psi})$ 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

$$f(\mathbf{x}_i) \approx H(\mathbf{x}_i; \boldsymbol{\psi}) = \sum_{j=1}^{K} \beta_j I_j(\mathbf{x}_i; \boldsymbol{\theta}_j) \tag{3-1}$$

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 $\boldsymbol{\psi} = (\beta_1, ..., \beta_K, \boldsymbol{\theta}'_1, ..., \boldsymbol{\theta}'_K)'$.

Neste trabalho, será considerada uma regressão logística linear por partes em cada folha, que representa um novo regime, e a transição entre os regimes é feita de forma suave. Nessa linha e dentro do contexto dos modelos lineares generalizados destacam-se os trabalhos de (5) e (4) que discutem as função $H(\mathbf{x}_i; \boldsymbol{\psi})$ para uma árvore de regressão Poisson e em Regressão Logística. Já o primeiro propõe a diferença entre funções desvio para a divisão dos nós e crescimento da árvore.

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

No exemplo mais simples que se pode apresentar de uma árvore em que temos apenas uma profundidade ($d = 1$) e $K = 2$ nós terminais, a equação que explica a relação entre y_i e $\mathbf{x}_i$ é dada por

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

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

Um exemplo numérico apresentado em (7) é mostrado na figura 3.1, a seguir. Logo após, na figura 3.2, apresentamos a divisão no espaço das covariáveis, $\mathbb{X} \subseteq \mathbb{R}^2$ e a tabela 3.1 com as sentenças lógicas e o correspondente valor da variável dependente estimada, $\hat{y}$.

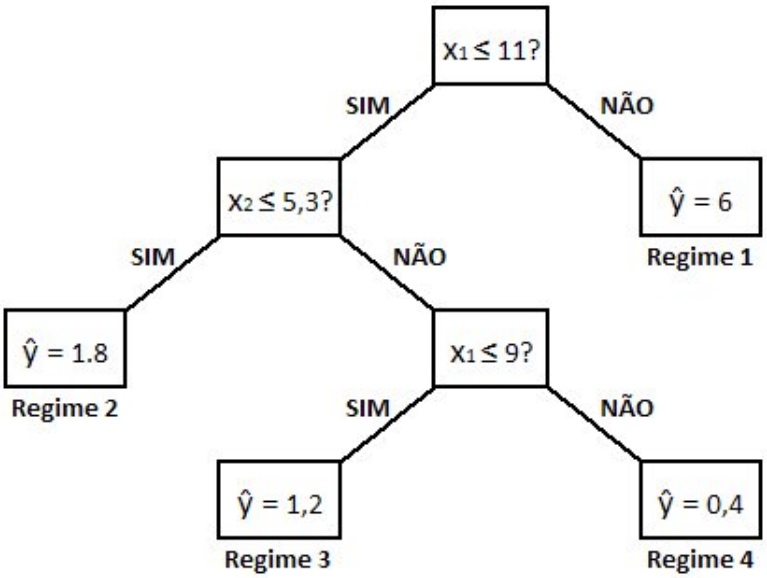

Figura 3.1: Estrutura do modelo. Exemplo em (7)

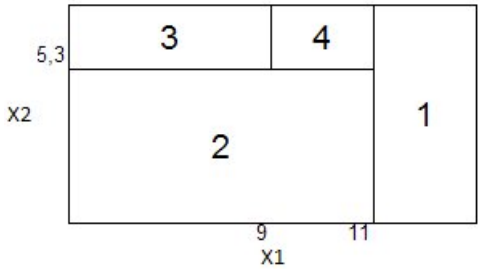

Figura 3.2: Divisão do espaço das covariáveis

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 3.1: Divisão do espaço das covariáveis

Algoritmo de Crescimento

Consiste na escolha de um nó a ser dividido, conseqüentemente uma variável de transição (x_{s_j}) e um limiar (c_j), e, de forma iterativa, estima-se os parâmetros dos modelos contidos em cada nó gerador. A seleção dos elementos citados e a estimação dos parâmetros são feitos simultaneamente.

Basicamente busca-se, a partir do nó raiz, 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

$$\hat{\beta}_1^{MQ} = \frac{\sum_{i=1}^{n} y_i I_0(\mathbf{x}_i; s_0, c_0)}{\sum_{i=1}^{n} I_0(\mathbf{x}_i; s_0, c_0)} \tag{3-2}$$

$$\hat{\beta}_2{}^{MQ} = \frac{\sum_{i=1}^{n} y_i[1 - I_0(\mathbf{x}_i; s_0, c_0)]}{\sum_{i=1}^{n}[1 - I_0(\mathbf{x}_i; s_0, c_0)]} \tag{3-3}$$

A divisão do nó gerado na posição 1 é feita da mesma maneira, através da busca por x_{s_1} e c_1 que minimizam

$$\begin{aligned} SQ^{Arv2} \quad &= \sum_{i=1}^{n} \{y_i - \beta_2[1 - I_0(\mathbf{x}_i; s_0, c_0)] - [\beta_3 I_1(\mathbf{x}_i; s_1, c_1) \\ &+\beta_4[1 - I_1(\mathbf{x}_i; s_1, c_1)] I_0(\mathbf{x}_i; s_0, c_0)\}^2 \end{aligned}$$

e assim sucessivamente até que não se tenha ganhos com a divisão. Em (2) é proposto um critério de parada, declarando como nó terminal aquele que contenha 5 observações ou menos.

Além disso, a fim de diminuir a complexidade da árvore que pode crescer mais que o necessário,mesmo utilizando-se o critério de parada, existe uma técnica que determina o corte de algumas folhas e por essa razão é chamada de *Podagem*.

Uma função, sugerida em (2), e apresentada em (7), cumpre o papel de

avaliar a necessidade de se reespecificar o modelo na tentativa de melhorar seu poder de previsão é

$$R^*(N, \alpha) = \sum_{i=1}^{N} R_i + |\alpha| \, N, \tag{3-4}$$

onde R_i é uma medida da qualidade do ajuste na i-ésima folha em uma árvore com N folhas, em que α é o parâmetro que penaliza a árvore pelo seu tamanho. Como um exemplo da R_i, suponha que ela seja calculada após a divisão do nó gerador da posição 1, assim o melhor modelo é o que maximiza

$$R(Arv_2) = SQ(Arv_1) - SQ(Arv_2).$$

3.2 Generalized Additive Models (GAM)

Proposto por (12) os quais posteriormente estenderam o trabalho em (13), trata-se de modelos de regressão não paramétricos desenvolvidos após o estudo sobre *Modelos Aditivos* em (27).

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, $\sum f_j(\mathbf{x}_j)$. Trata-se de uma generalização ainda maior do que os MLG´s, como se mostra abaixo, figura 3.3, na estrutura dos modelos encontrada em (11).

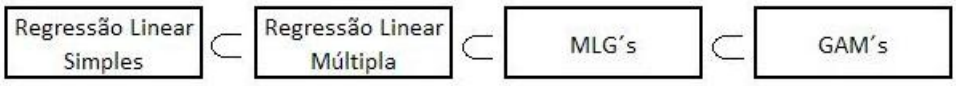

Figura 3.3: Hierarquia dos modelos

Os GAM´s são considerados modelos *semi-paramétricos*, pois, assim como os MLG´s, são paramétricos no que diz respeito a distribuição de probabilidade da variável dependente, a qual deve ser especificada, porém alguns preditores podem ser modelados de forma não-paramétrica através de termos lineares e polinomiais de outros preditores, podendo, desta maneira, mensurar relações não-lineares entre a variável dependente e as variáveis explicativas. Essa é, sem dúvida, sua maior vantagem.

Assim como os MLG´s a relação entre a média da variável dependente e, no caso dos GAM´s, as função suavizadas das variáveis explicativas é feita por uma função de ligação tendo como principal hipótese que aquelas sejam funções aditivas entre as covariáveis e as componentes sejam suaves. Desta

maneira, a forma de um GAM é apresentada da seguinte maneira

$$\mathbb{E}(\mathbf{y}|x_1, ..., x_p) = f_0 + f_1(x_1) + f_2(x_2) + ... + f_p(x_p) \tag{3-5}$$

onde o preditor linear dos MLG´s, $\eta = \sum \beta_j \mathbf{x}_j$, é substituído por $\eta = \sum f_j(\mathbf{x}_j)$, $j = 1, ..., p$.

Cada uma dessas funções é ajustada através de um diagrama de dispersão suavizado (*scatterplot smoother*) e utilizando um algoritmo se realiza a estimação das p funções simultaneamente. Conforme apresentado em (13), um diagrama de dispersão suavizado é uma função s de $\mathbf{x}$ e $\mathbf{y}$, com mesmo domínio que os valores em $\mathbf{x} : s = \mathbf{S}(\mathbf{y}|\mathbf{x})$, a qual tem como principais atributos a descrição visual da relação entre a variável dependente e as covariáveis, além da estimação da relação entre as mesmas, que nada mais é que o ajuste da reta suavizada, $f(x)$, que sintetize a dependência entre $\mathbf{y}$ e $\mathbf{x}$. Tal reta deve ser tal que minimize $\sum_{i=1}^{n}[y_i - f(x_i)]^2$.

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 - SQRP* (em inglês, *Penalized Residual Sum of Squares - PRSS*), denotada por

$$SQRP(f, \lambda) = \sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \int_a^b f''(t)^2 dt, \quad a \leq x_1 \leq ... \leq b \tag{3-6}$$

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

Quanto maior for λ ($\lambda \to \infty$) o termo que penaliza a SQRP é dominante, forçando $\int_a^b f''(t)^2 dt = 0$, sendo assim a reta ajustada por Mínimos Quadrados. Caso contrário, $\lambda \to 0$, a solução tende para qualquer função que faça a interpolação dos dados. Alguns métodos de seleção automática de λ são apresentados em (14).

Uma maneira intuitiva de escolher λ é através da determinação dos graus de liberdade (gl) para o suavizador, no caso o Cubic Smoothing Splines, e utilizar uma otimização numérica para determinar o valor do parâmetro que retorne tal número.

Os graus de liberdade de um suavizador são dados por

$$gl_\lambda = trace(\mathbf{S}_\lambda) \tag{3-7}$$

onde $\mathbf{S}_\lambda$ é um operador linear suavizado e a soma de seus autovalores definem os graus de liberdade. Por exemplo, quando usamos um suavizador com 4 gl, significa que para cada x_j o parâmetro λ_j é escolhido tal que

$trace[\mathbf{S}_j(\lambda_j)] - 1 = 4$.

3.2.1 Regressão Logística Aditiva

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

$$\log\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right] = f_0 + f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) + ... + f_p(\mathbf{x}_p). \tag{3-8}$$

As função de ligação utilizada, $g[\pi(x)]$, é a logito. Além dela os GAM´s também admitem as demais funções de ligação: probito, logística e, obviamente, a identidade.

Para especificar o modelo utiliza-se os critérios de forma semelhante àquela feita para os MLG´s na seção 2.2.1. Apenas a estimação é feita de forma diferente.

Estimação do modelo de Regressão Logística Aditiva

Para estimar uma Regressão Logística Aditiva sabendo que, dada a forma do modelo apresentada anteriormente, temos para apenas uma variável dependente, X, o modelo

$$\log\left[\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)}\right] = f(x)$$

em que

$$\mathbb{P}(Y=1|X=x) = \frac{e^{f(x)}}{1+e^{f(x)}}.$$

Para tal, o método da *Máxima Verossimilhança Penalizada*, apresentado em detalhes em (14), é alocado. Tal método segue o mesmo princípio apresentado anteriormente onde se deve maximizar a log-verossimilhança, guardadas as devidas alterações com a inclusão do termo penalizador como segue

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

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

Representando $f(x) = \sum_{j=1}^{n} N_j(x)\theta_j$, chamado *natural spline*, em que $N_j(x)$ é um conjunto N-dimensional de funções bases, temos a primeira e

segunda derivadas representadas na forma matricial por

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta} = \mathbf{N}'(\mathbf{y} - \mathbf{p}) - \lambda \boldsymbol{\Omega} \theta,$$

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta \partial \beta'} = -\mathbf{N}'\mathbf{W}\mathbf{N} - \lambda \boldsymbol{\Omega},$$

onde $\{\mathbf{N}\}_{ij} = N_j(x_i)$ e $\{\boldsymbol{\Omega}_N\}_{jk} = \int N_j''(t) N_k''(t) dt$.

Assim os valores de θ e das funções ajustadas são obtidos iterativamente por meio de

$$\begin{aligned} \theta^{m+1} &= (\mathbf{N}'\mathbf{W}\mathbf{N} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{N}'\mathbf{W}(\mathbf{N}\theta^m + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p})) \\ &= (\mathbf{N}'\mathbf{W}\mathbf{N} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{N}'\mathbf{W}\mathbf{z} \end{aligned}$$

$$\begin{aligned} \mathbf{f}^{m+1} &= (\mathbf{N}'\mathbf{W}\mathbf{N} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{N}'\mathbf{W}(\mathbf{f}^m + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p})) \\ &= \mathbf{S}_{\lambda,w}\mathbf{z}. \end{aligned}$$

3.3 k-Nearest Neighbor

Primeiramente apresentado em (9) e posteriormente desenvolvido e teoricamente provado em (6), trata-se de um classificador não-paramétrico, para o qual não é necessário um modelo a ser ajustado, onde, 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

$$\hat{Y}(x_{(i)}) = \frac{1}{k} \sum_{x_0 \in N_k(x_{(i)})} y_0, \tag{3-9}$$

onde $N_k(x_{(i)})$ é a vizinhança de $x_{(i)}$ definida pelos k pontos amostrais próximos de x_0 na amostra de treinamento. Tal procedimento nada mais é do que encontrar as k observações mais próximas do novo ponto, x_0, e tirar a média dos valores de suas variáveis dependentes.

A figura 3.4 ilustra um exemplo de como a variação no valor de k influencia na classificação. Para $k = 5$ classificaríamos o novo ponto como sendo um círculo e caso o número aumentasse para $k = 15$, por exemplo, a

classificação mudaria para um quadrado.

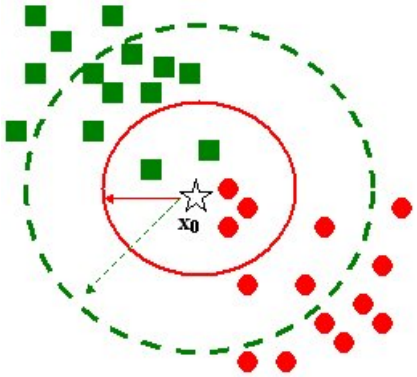

Figura 3.4: Exemplo: k-Nearest Neighbor

3.4 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.

Utiliza-se a técnica através da *Função Discriminante Linear de Fisher* (em inglês, *Fisher Discriminant Linear - FDL*) que, conforme apresentado em (16), transforma as observações multivariadas $\mathbf{x}$ em observações univariadas y, tal que os y´s das populações P_1 e P_2 fossem separados o máximo possível.

Assim sendo, a função discriminante de Fisher tem a forma da combinação linear $\hat{\mathbf{y}} = \hat{\mathbf{a}}'\mathbf{x}$, onde $\hat{\mathbf{a}}' = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_p^{-1}$ e $\mathbf{S}_p = \frac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2}$. Considerando os estimadores S e $\bar{\mathbf{x}}$ referentes a Σ e $\boldsymbol{\mu}$.

A expressão $\hat{\mathbf{y}} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_p^{-1}\mathbf{x}$ maximiza a razão

$$\frac{(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)^2}{s_y^2} = \frac{(\hat{\mathbf{a}}'\bar{\mathbf{x}}_1 - \hat{\mathbf{a}}'\bar{\mathbf{x}}_2)^2}{\hat{\mathbf{a}}'\mathbf{S}_p\hat{\mathbf{a}}} = \frac{(\hat{\mathbf{a}}'\mathbf{d})^2}{\hat{\mathbf{a}}'\mathbf{S}_p\hat{\mathbf{a}}}$$

onde $\mathbf{d} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$. O resultado e prova da maximização são encontrados em (16) (pp. 610).

Para uma nova observação, x_0, a regra de alocação em uma das populações discriminadas pela função é a seguinte

– Alocação em P_1 se:

$$\hat{y}_0 = \hat{\mathbf{a}}'\mathbf{x}_0 \geq \frac{1}{2}\,\hat{\mathbf{a}}'(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) = \frac{\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2}{2};$$

– Alocação em P_2 se:

$$\hat{y}_0 < \frac{\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2}{2}.$$

3.4.1
Estimação das Probabilidades de Classificação Incorreta

Em se tratando das populações P_1 e P_2, podem ser cometidos dois tipos de erros. Segundo (23) são eles:

– Erro 1: elementos provenientes da população 1 que são classificados como pertencentes à população 2;

– Erro 2: elementos provenientes da população 2 que são classificados como pertencentes à população 1.

Desta maneira define-se $\mathbb{P}(Erro\ 1) = p(2|1)$ e $\mathbb{P}(Erro\ 2) = p(1|2)$.

Uma forma de visualizar tais erros é através da *Matriz de Confusão*, que é um artifício semelhante à Tabela de Classificação, como se pode notar na tabela 3.2

Tabela 3.2: Matriz de Confusão

	População Classificada		
População Real	pop 1	pop 2	
pop 1	n_{11}	n_{12}	N_1
pop 2	n_{21}	n_{22}	N_2

Seus elementos são:

– n_{11}: itens de P_1 classificados corretamente em P_1;

– n_{12}: itens de P_1 classificados incorretamente em P_2;

– n_{21}: itens de P_2 classificados incorretamente em P_1;

– n_{22}: itens de P_2 classificados corretamente em P_2;

– N_1: total de itens em P_1;

– N_2: total de itens em P_2.

A Taxa Aparente de Erro (APER) definida em (16) é dada por

$$APER = \frac{n_{12} + n_{21}}{N_1 + N_2}$$

Ainda com a tabela 3.2 calculamos as estimativas das probabilidades dos erros, dadas por: $\hat{p}(1|2) = \frac{n_{21}}{N_2}$ e $\hat{p}(2|1) = \frac{n_{12}}{N_1}$. Quanto menores elas forem, melhor será a função de discriminação.

A avaliação e construção da matriz de confusão serão feitas neste trabalho através do *Método da Ressubstituição* (ver (23)) em que os escores de cada elemento amostral observado de P_1 e P_2 são calculados, sendo a regra de discriminação utilizada para classificar os $N = N_1 + N_2$ elementos da amostra conjunta. Assim os mesmos elementos amostrais participam da estimação da regra de classificação e da estimação dos erros. Outros dois métodos utilizados são: *Método Holdout* e o *Método de Lachenbruch*, extensamente debatidos em (16) e (23).