Análise Discriminante

Danielle Ribeiro, Luccas Martins, Marlon Alves e Matheus Alves

A Análise Discriminante é uma técnica da estatística multivariada utilizada para discriminar e classificar objetos baseando-se na separação dos mesmos em duas ou mais classes. A primeira etapa do processo é a discriminação ou separação, considerada a parte exploratória da análise e que consiste em procurar características em comum entre os objetos que servirão para alocá-los em diferentes grupos previamente definidos. Ou seja, um conjunto de regras é definido para que os elementos de uma população sejam agrupados de forma com que os grupos menores possuam caracteristicas mais semelhantes entre si. É um método utilizado para prever a probabilidade de um elemento pertencer a uma determinada classe com base em uma ou várias variáveis preditoras que podem ser contínuas ou categóricas.

Função Discriminante

O objetivo da Análise Discriminante é, através de funções matemáticas, classificar um elemento X com base em medidas de um número p de características, buscando minimizar a probabilidade de classificar inadequadamente este elemenento em uma população π_i, quando realmente pertence a população $\pi_j, (i \neq j)\ i, j = 1, 2, ..., g$. A análise é feita através de uma combinação linear de características observadas que apresenta o maior poder de discriminação entre as populações. Essa função é denominada de função discriminante e pode ser linear ou quadrática.

A função discriminante tem a propriedade de minimizar as probabilidades de má classificação, quando as populações são normalmente distribuídas, o que não ocorre quando a média e a variância da distribuição não são conhecidas necessitando então da estimação desses parâmetros. Na grande maioria dos processos, as informações da população não são conhecidas e por isso a função discriminante é obtida a partir de amostras de treinamento. Além disso, ela tem como objetivo separar as regiões de alocação, ou seja, a reta traçada pela função discriminante define uma fronteira entre os grupos onde cada elemento da população em análise será alocado. Porém, é importante ressaltar que no mundo real a fronteira entre regiões não está exatamente definida e sempre haverá superposição, isto é, erro de classificação.

Regras de Classificação

O principal objetivo da Análise Discriminante é proporcionar uma boa classificação que deve resultar em pequenos erros, havendo pouca probabilidade de má classificação. Para isso acontecer é necessário que regras de classificação sejam seguidas. São elas:

- Considerar as probabilidades a priori e os custos de má classificação
- Considerar se as variâncias das populações são iguais ou não

Quando a regra de classificação assume que as variâncias das populações são iguais, as funções discriminantes são ditas lineares e quando não, são funções discriminantes quadráticas.

A Análise Discriminante possui, basicamente, dois tipos de função discriminante. São elas, a função discriminante linear de Fisher e a função discriminante de Anderson. Abaixo estão definidas cada uma delas.

Função Discriminante Linear de Fisher

É uma combinação linear de características originais e se caracteriza por produzir a separação máxima entre duas populações. Considerando que μ_i e Σ são parâmetros conhecidos e respectivamente, os vetores de médias e a matriz de covariâncias comum das populações π_i. Demonstra-se que a função linear do vetor aleatório X que produz separação máxima entre duas populações é dada por:

$$D(X) = L'.X = [\mu_1 - \mu_2].\Sigma^{-1}.X$$

em que,

$$X = \lfloor X_1 X_2 ... X_p \rfloor \text{ e } \pi = [\pi_1, \pi_2]$$

$L =$ vetor discriminante;
$X =$ vetor aleatório de características das populações;
$\mu =$ vetor de médias p-variado;
$\Sigma =$ matriz comum de covariâncias das populações π_1 e π_2;

O valor da função discriminante de Fisher para uma dada observação x_0 é:

$$D(x_0) = [\mu_1 - \mu_2]'.\Sigma^{-1}.x_0$$

O ponto médio entre as duas médias populacionais univariadas μ_1 e μ_1 é:

$$m = \frac{1}{2}[\mu_1 - \mu_2]'.\Sigma^{-1}.[\mu_1 + \mu_2], \text{ ou seja}$$

$$m = \frac{1}{2}[D(\mu_1) + D(\mu_2)]$$

A regra de classificação baseada na função discriminante de Fisher é:

$$\text{Alocar } x_0 \text{ em } \pi_1 \text{ se } D(x_0) = [\mu_1 - \mu_2]'.\Sigma^{-1}.x_0 \geq m$$

$$\text{Alocar } x_0 \text{ em } \pi_2 \text{ se } D(x_0) = [\mu_1 - \mu_2]'.\Sigma^{-1}.x_0 < m$$

Assumindo que as populações π_1 π_2 têm a mesma matriz de covariâncias Σ pode-se então estimar uma matriz comum de covariâncias S_c:

$$S_c = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\right].S_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\right].S_2$$

em que,

$S_c =$ estimativa da matriz comum de covariâncias Σ;
$n_1 =$ número de observações da população μ_1;
$n_2 =$ número de observações da população μ_2;
$S_1 =$ estimativa matriz de covariâncias da população π_1;
$S_2 =$ estimativa matriz de covariâncias da população π_2;

A função discriminante linear amostral de Fisher é obtida substituindo-se os parâmetros μ_1, μ_2 e Σ pelas respectivas quantidades amostrais $\bar{x}_1$, $\bar{x}_2$ e S_c:

$$D(x) = \hat{L}'.x = [\bar{x}_1 - \bar{x}_2]'.S_c^{-1}.x$$

em que,

$D(x) =$ função discriminante linear amostral de Fisher;
$\hat{L}' =$ estimativa do vetor disriminante;
$\bar{x}_1 =$ média amostral da população π_1;
$\bar{x}_2 =$ média amostral da população π_2.

Função Discriminante de Anderson

Sejam $\pi_1, \pi_2, \cdots, \pi_g$ um grupo de g populações. Para obter as funções discriminantes para esse grupo de populações é necessário respeitar alguns pressupostos. São eles:

- As populações apresentam algum tipo de distribuição
- Existe uma probabilidade de ocorrência a priori para cada população no grupo
- Existe um custo de má classificação

Após verificar os pressupostos acima, os classificadores são desenvolvidos com o objetivo de alocar uma observação X em uma das g populações, sendo $g \geq 2$. Para desenvolver um classificador também é necessário respeitar alguns pressupostos para o modelo da função discriminate. São eles:

- As g populações apresentam distribuição normal multivariada
- As p_i probabilidades a priori de ocorrência das populações são iguais e $\sum_{i=1}^{n} p_i = 1$
- As populações apresentam custos iguais de má classificação

Por fim, para obter a função discriminante de Anderson, considerando que as g populações apresentam distribuição normal multivariada a função discriminante é dada por:

$$D_i(\tilde{x}) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}[\tilde{x} - \tilde{\mu}_i]'\Sigma_i^{-1}[\tilde{x} - \tilde{\mu}_i] + \ln(p_i),\ i = 1, 2, ..., g$$

em que,

$D_i(\tilde{x}) =$ função discriminante da população i o vetor aleatório $\tilde{x}$;
$\Sigma_i =$ matriz de covariância da população i
$\tilde{x} =$ vetor aleatório de características;
$\tilde{\mu}_i =$ vetor de médias da população i;
$p_i =$ probabilidade de ocorrência da população i.

Supondo igualdade das matrizes de covariâncias, então os componentes constantes para todo i podem ser retirados e a função discriminante é:

$$D_i(\tilde{x}) = \tilde{L}_i'.\tilde{x} - \frac{1}{2}.\tilde{L}_i. + \ln(p_i)$$

$D_i(\tilde{x}) =$ função discriminante da população i o vetor aleatório $\tilde{x}$;
$\tilde{L}_i =$ vetor aleatório discriminante da população i
$\tilde{x} =$ vetor aleatório de características;
$\tilde{\mu}_i =$ vetor de médias da população i;
$p_i =$ probabilidade de ocorrência da população i.

sendo que,

$$\tilde{L}_i = \Sigma^{-1}.\mu_i$$

$\Sigma =$ matriz comum de covariâncias das g populações.

A regra de classificação para alocar um indivíduo x é a seguinte: classificar x em π_i se e somente se

$$D_i(\tilde{x}) = max\Big(D_1(\tilde{x}), D_2(\tilde{x}), \cdots, D_g(\tilde{x})\Big)$$

em que,

$D_i(\tilde{x}) =$ valor da função discriminante da população i para o vetor de características $\tilde{x}$;
$D_1(\tilde{x}) =$ valor da função discriminante da população 1 para o vetor de características $\tilde{x}$;
$D_2(\tilde{x}) =$ valor da função discriminante da população 2 para o vetor de características $\tilde{x}$;
$D_g(\tilde{x}) =$ valor da função discriminante da população g para o vetor de características $\tilde{x}$;

A próxima etapa é analisar a igualdade das matrizes de covariâncias das populações. Se as matrizes Σ_i são iguais, então a função discriminante é dita linear de Anderson, caso contrário é dita função discriminante quadrática de Anderson. Segundo VARELLA (2004), através de análises realizadas, a acurácia do modelo gerado com ambos os tipos de função discriminante não é a mesma. Portanto, a função que apresentar maior acurácia de classificação nas amostras de teste é considerada a mais adequada. Alguns estudos apontam que na maioria dos casos a função linear apresenta melhor resultado. E segundo HOFFBECK & LANDGREBE (1996), em situações em que o número de observações utilizadas no treinamento do classificador é limitado, a estimativa de uma covariância comum para todas as populações pode resultar numa melhor classificação, devido a redução dos parâmetros a serem estimados.

Esta estimativa da matriz comum de covariâncias amostral é dada pela seguinte fórmula:

$$S_c = \frac{\sum_{i=1}^{g}(n_i - 1).S_i}{\sum_{j=1}^{g}(n_j - 1)}$$

onde,

$S_c =$ estimativa da matriz comum de covariâncias das g populações.

Como a escolha da melhor função discriminante depende dos dados e da análise da acurácia do modelo, recomenda-se testar os dois modelos de funções discriminantes, isto é o modelo linear e o modelo quadrático, para que ambos sejam analisados e o melhor modelo de classificação seja escolhido.

Para isso, a próxima etapa é avaliar a acurácia de classificação por meio do coeficiente kappa κ, obtido a partir da matriz de erros da classificação. Para obter a matriz de erros, o indicado é utilizar 25% do total de observações e os restantes 75% são usados para obter as funções discriminantes. A matriz de erros tem dimensão $g \times g$, onde g é o número de populações envolvidas na análise discriminante. Nas colunas da matriz estão contidas as informações das observações de referência, enquanto nas linhas as informações das observações classificadas. Logo, na diagonal encontram-se o número de observações corretamente classificadas.

A partir dessa matriz é possível calcular os erros de omissão, de comissão, a exatidão global e o coeficiente kappa e por fim decidir qual melhor função discriminante para a análise em questão.

A exatidão global é determinada pela seguinte expressão:

$$\text{EG} = \frac{N_c}{N_t}.100 \tag{1}$$

em que,

EG = exatidão global, %;
$N_c =$ número de observações corretamente classificadas;
$N_t =$ número total de observações.

O coeficiente kappa é estimado pela seguinte expressão:

$$\hat{\kappa} = \frac{n\sum_{i=1}^{c} X_{ii} - \sum_{i=1}^{c} X_{i+}X_{+i}}{n^2 - \sum_{i=1}^{c} X_{i+}X_{+}i} \tag{2}$$

em que,

$c =$ número de classes na matriz de erros;
$X_{ii} =$ valores na linha i e na coluna i;
$X_{i+} =$ total da linha i;
$X_{+i} =$ total da coluna i;
$n =$ número total de observações.

Os coeficientes kappas são comparados pelo teste Z determinado pela seguinte expressão (CONGALTON & MEAD, 1983):

$$Z = \frac{\tilde{\kappa}_1 - \tilde{\kappa}_2}{\sqrt{\tilde{\sigma}_1 + \tilde{\sigma}_2}} \tag{3}$$

em que,

$Z =$ valor Z calculado;
$\kappa_1 =$ estimativa do coeficiente Kappa do classificador 1;
$\kappa_2 =$ estimativa do coeficiente Kappa do classificador 2;
$\tilde{\sigma}_1 =$ estimativa da variância do Kappa do classificador 1;
$\tilde{\sigma}_2 =$ estimativa da variância do Kappa do classificador 2.

Se o valor Z calculado para o teste for maior que o valor Z tabelado, diz-se que o resultado foi significativo e rejeita-se a hipótese nula ($H_0 : K_1 = K_2$) concluindo-se que os dois classificadores são estatisticamente diferentes. O valor de Z tabelado ao nível de 5% de probabilidade é igual a 1,96.

Apesar dos modelos lineares e quadráticos serem os mais conhecidos e utilizados, existem outros métodos para realizar uma análise discriminante. Os métodos citados anteriormente e tratados neste estudo são métodos que precisam respeitar o pressuposto de normalidade, mas para tentar sanar o problema de casos em que isso não acontece, é possível aplicar os seguintes métodos: análise discriminante de mistura; análise discriminante flexível; e análise discriminante regularizada.

Exemplo

Para a realização do exemplo, usamos a base de dados do RStudio chamada *irís* que possui 150 observações e 5 variáveis.

Passo 1 - Preparando os dados

Neste passo, a base de dados será separada em 2 subconjuntos, treino e teste. Para isso será utilizada a função *CreateDataPartition* do pacote *caret*, pois ela garante que as amostras resultantes manteram as mesmas proporções dos dados como um todo. O conjunto de dados de treino será usado para construir o modelo preditivo e o de teste para avaliar a qualidade deste modelo.

```r
library(tidyverse)
library(MASS)
require(htmltools)
library(caret)
library(mvnormtest)
library(klaR)
```

```r
set.seed(12042021)
ind_treino = createDataPartition(iris$Species, p = 0.75, list = F)

treino = iris[ind_treino,]
teste = iris[-ind_treino,]

prop.table(table(iris$Species))
```

```
##
##     setosa versicolor  virginica
##  0.3333333  0.3333333  0.3333333
```

```r
prop.table(table(treino$Species))
```

```
##
##     setosa versicolor  virginica
##  0.3333333  0.3333333  0.3333333
```

```r
prop.table(table(teste$Species))
```

```
##
##     setosa versicolor  virginica
##  0.3333333  0.3333333  0.3333333
```

Passo 2 - Checando os pressupostos

Neste passo, os pressupostos necessários para a aplicação da análise discriminante linear serão checados. Porém, para efeitos didáticos, vamos supor que esses pressupostos foram aceitos.

Normalidade (Teste da normalidade multivariada)

```r
U = t(as.matrix(treino[,-5]))
mshapiro.test(U)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.97867, p-value = 0.06553
```

Igualdade de variâncias (Teste de Bartlett)

```r
bartlett.test(treino[,-5])
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  treino[, -5]
## Bartlett's K-squared = 228.36, df = 3, p-value < 2.2e-16
```

Passo 3 - Aplicando a análise discriminante linear (LDA)

O *prior probabilities of groups* mostra a probabilidade de escolhermos aleatoriamente uma observação de alguma categoria da base de dados total. Como temos 50 observações de cada categoria, era esperado que o resultado fosse próximo de 33.3% para cada categoria. O *group means* mostra a média de cada categoria para as diferentes variáveis numéricas apresentadas na base. *Coefficients of linear discriminants* mostra os coeficientes para cada discriminante, com a combinação linear do primeiro discriminante linear sendo:

$$(0,54 * Sepal.Length) + (2,02 * Sepal.Length) + (-1.59 * Sepal.Length) + (-3.70 * Sepal.Length)$$

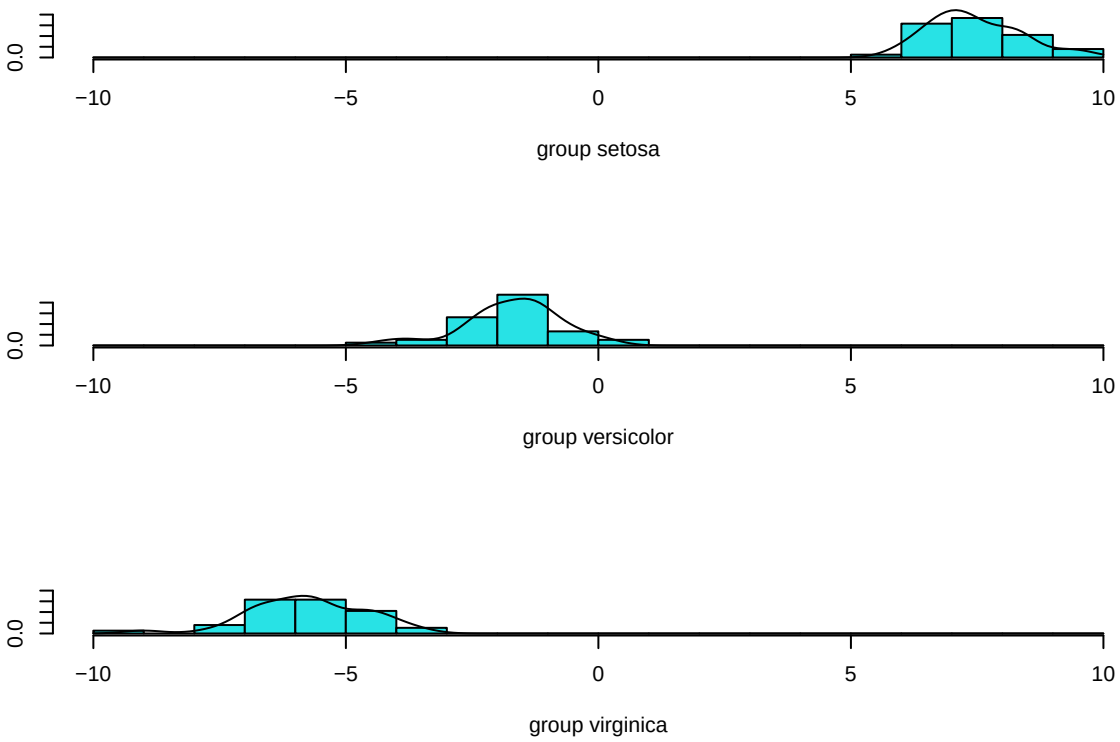
A *Proportion of trace* descreve a proporção das variâncias entre classes que é explicada pelos discriminantes lineares. Neste caso, 98.99% da variância é explicada pelo LD1.

```r
modelo1 = lda(Species~.,treino)
modelo1
```

```
## Call:
## lda(Species ~ ., data = treino)
##
## Prior probabilities of groups:
##     setosa versicolor  virginica
##  0.3333333  0.3333333  0.3333333
##
## Group means:
##            Sepal.Length Sepal.Width Petal.Length Petal.Width
## setosa         5.015789    3.434211     1.457895    0.250000
## versicolor     5.844737    2.763158     4.189474    1.305263
## virginica      6.589474    2.981579     5.557895    2.042105
##
## Coefficients of linear discriminants:
##                      LD1        LD2
## Sepal.Length  0.5424166 -0.5600061
## Sepal.Width   2.0269083 -1.9076612
## Petal.Length -1.5873412  1.1215009
## Petal.Width  -3.7006372 -2.7219083
##
## Proportion of trace:
##    LD1    LD2
## 0.9899 0.0101
```

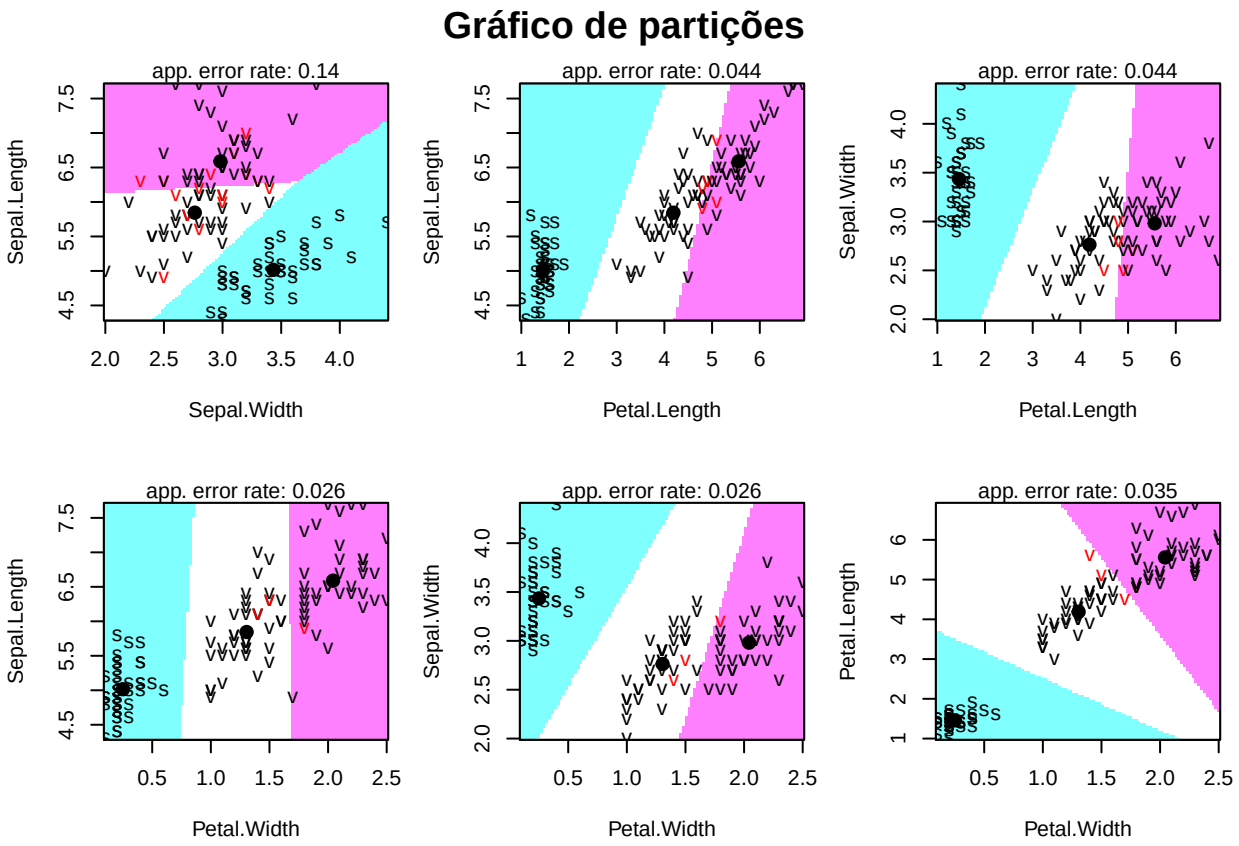
Com um gráfico simples é possível ver que o grupo *setosa* apresentou um comportamento bem distinto em relação ao valor discriminante. Em contra partida os grupos *versicolor* e *virginica* apresentam sobreposições proximo ao valor -4. Isso pode indicar uma maior dificuldade em distinguir observações desses grupos

```
plot(modelo1, dimen = 1, type = "b")
```



Através do gráfico de partiçõees é possivel ver como cada combinação de duas variáveis auxiliam na classificação dos objetos. Cada uma dar cores representa uma das classes, e as observações que são classificadas de acordo com a cor da região que se encontram.

```
partimat(Species ~ ., data=treino, method="lda", main = "Gráfico de partições")
```



Passo 4 - Predição

Como pode-se observar, todas as observações foram previstas corretamente, pois fora da diagonal da matriz de confusão, todos os valores são 0. Além disso, o modelo tem um acurácia de 100%, o que significa que possui um desempenho excelente.

```
predict1 = predict(modelo1, teste)

matriz_confusao1 = confusionMatrix(predict1$class, teste$Species)
matriz_confusao1$table
```

```
##             Reference
## Prediction   setosa versicolor virginica
##   setosa         12          0         0
##   versicolor      0         12         0
##   virginica       0          0        12
```

```
kappa1 = kappa(matriz_confusao1$table)
acuracia1 = matriz_confusao1$overall[1]
EG1 = sum(diag(matriz_confusao1$table)) / dim(teste)[1] * 100
resultados1 = data_frame(Acuracia = acuracia1, Kappa = kappa1$coef, EG = EG1)

resultados1
```

```
## # A tibble: 1 x 3
##   Acuracia Kappa    EG
##      <dbl> <dbl> <dbl>
## 1        1     1   100
```

Análise discriminante quadrática

Diferente da LDA, a QDA não assume covariâncias iguais entre as classes.Também não encontra combinações lineares de variáveis independentes, mas sim função quadrática dessas variáveis.

Aplicando QDA

O *prior probabilities of groups* mostra a probabilidade de escolhermos aleatoriamente uma observação de alguma categoria da base de dados total. Como temos 50 observações de cada categoria, era esperado que o resultado fosse próximo de 33.3% para cada categoria. O *group means* mostra a média de cada categoria para as diferentes variáveis numéricas apresentadas na base. Diferente da LDA, não possui combinações lineares.

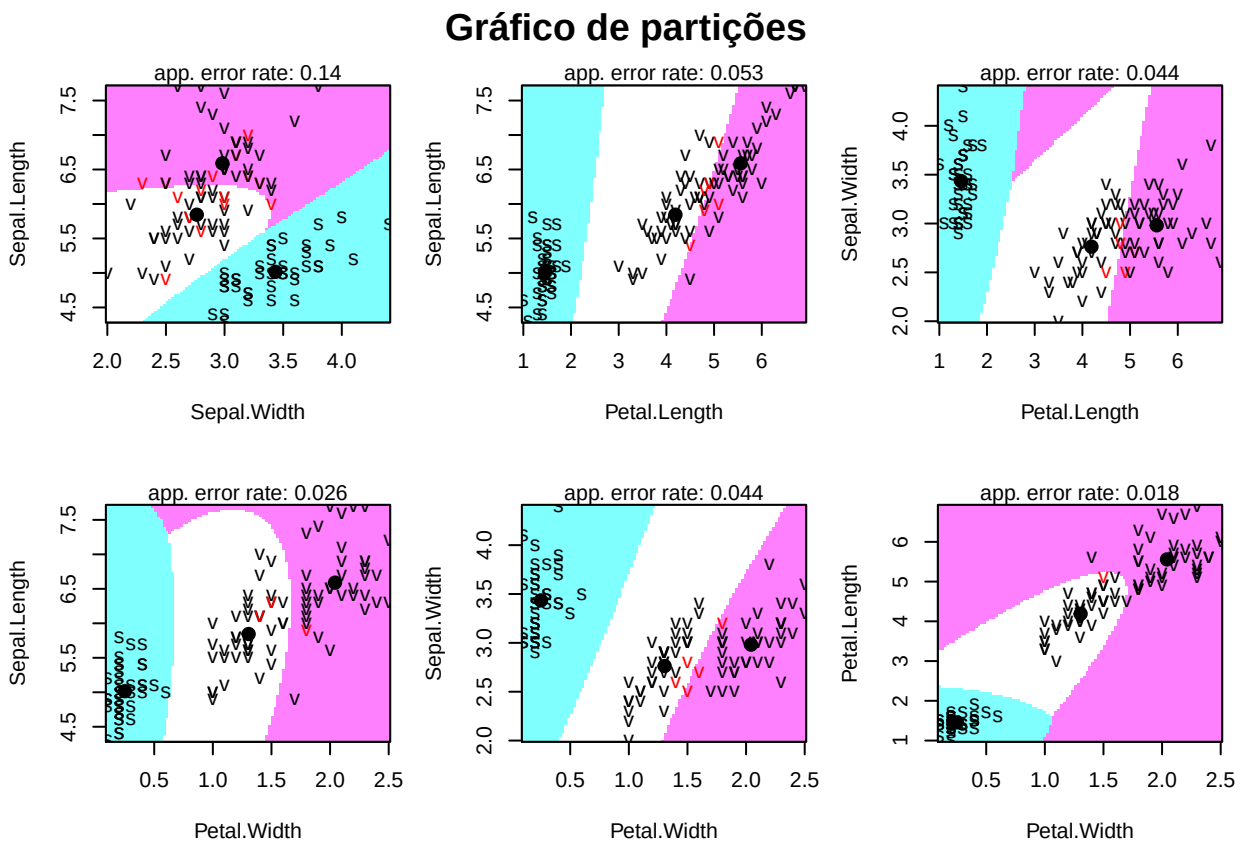
```
modelo2 = qda(Species~.,treino)

modelo2
```

```
## Call:
## qda(Species ~ ., data = treino)
##
## Prior probabilities of groups:
##     setosa versicolor  virginica
##  0.3333333  0.3333333  0.3333333
##
## Group means:
##            Sepal.Length Sepal.Width Petal.Length Petal.Width
## setosa         5.015789    3.434211     1.457895    0.250000
## versicolor     5.844737    2.763158     4.189474    1.305263
## virginica      6.589474    2.981579     5.557895    2.042105
```

Através do gráfico de partições é possivel ver como cada combinação de duas variaveis auxiliam na classificação dos objetos. Cada uma das cores representa uma das classes, e as observações que são classificadas de acordo com a cor da região que se encontram.

```
partimat(Species ~ ., data=treino, method="qda", main = "Gráfico de partições")
```



Predição

Como pode-se observar, praticamente todas as observações foram previstas corretamente, com apenas uma sendo prevista errada, pois está fora da diagonal da matriz de confusão. Além disso, o modelo tem um acurácia de 97.22%, o que significa que possui um desempenho excelente.

```
predict2 = predict(modelo2, teste)

matriz_confusao2 = confusionMatrix(predict2$class, teste$Species)
matriz_confusao2$table
```

```
##             Reference
## Prediction   setosa versicolor virginica
##   setosa         12          0         0
##   versicolor      0         11         0
##   virginica       0          1        12
```

```
kappa2 = kappa(matriz_confusao2$table)
acuracia2 = matriz_confusao2$overall[1]
EG2 = sum(diag(matriz_confusao2$table)) / dim(teste)[1] * 100
resultados2 = data_frame(Acuracia = acuracia2, Kappa = kappa2$coef, EG = EG2)

resultados2
```

```
## # A tibble: 1 x 3
##   Acuracia Kappa    EG
##      <dbl> <dbl> <dbl>
## 1    0.972 0.958  97.2
```

Comparando os modelos

```
Z = (kappa1$coef - kappa2$coef) / sqrt(kappa1$coefmat[2]^2 + kappa2$coefmat[2]^2)
```

Com um valor de $Z = 1.0154$, tem-se que, considerando, um nível de significância de 5%, há evidências a favor da igualdade entre os modelos.

Referências

JOHNSON, R. A.; WICHERN, D. W. Applied multivariate statistical analysis. 4th ed. Upper Saddle River, New Jersey: Prentice-Hall, 1999, 815 p.

VARELLA, C.A.A. Análise Discriminante. Rio de Janeiro.

VARELLA, C.A.A. Estimativa da produtividade e do estresse nutricional da cultura do milho usando imagens digitais. 2004. 92 f. Tese (Doutorado em Engenharia Agrícola) – Universidade Federal de Viçosa, Viçosa, 2004.