5 de novembro de 2023

ANÁLISE DE DISCRIMINANTE LINEAR

ME731 - ANÁLISE MULTIVARIADA

Ana Julia Cunha e Silva - RA: 236038

Súmario

Lista de Tabelas

Lista de Figuras

1 Introdução

Este relatório tem como propósito apresentar, a técnica de discriminação multivariada análise de disciminante linear, com certo nível de detalhe teórico. Além de apresentar, este relatório irá fazer uma abordagem prática exemplificada com um banco de dados e o software RStudio, além de descrever e discutir vantagens e desvantagens da técnica em questão. A estrutura do documento se resume em, análise teórica, análise completa de um banco de dados(composta por análise Descritiva, Exploratória e Inferêncial), discussão sobre as vantagens e desvantagens do método e por fim conclusões sobre os resultados das análises do banco de dados.

No Rstudio foram utilizados os pacotes *MASS* e *tidyverse* para fazer as análises e tratamento dos dados e o pacote *palmerpenguins* para obter o banco de dados **penguins**. Todos os valores utilizados para fazer as tabelas da análise dos dados foram arrendodados com até 4 casas decimais.

2 Técnicas de Discriminação

A análise de discriminante pode ser explicada de forma resumida como uma técnica de agrupamento, consistindo na separação de algumas observações para alocação em grupos pré-definidos, assim como explicado em [1]. De acordo com [2], a análise de discriminante tem natureza exploratória, pois seu propósito é investigar as diferenças quando as relações causais não são bem entendidas.

Existem algumas outras Técnicas de discriminação, como a análise de discriminante quadrática. Entretanto, dentre todas as possíveis, a análise de discriminante linear foi a escolhida para analisar o banco de dados.

2.1 Regras de Discriminação

Seja Π_j a j-ésima população(ou grupo), nossos dados possuem Π_i populações, com $i = 1, \ ... \ , \ n$ queremos alocar uma observação x a uma das i populações, para isso precisamos definir alguma regra pra alocação(ou "discriminação"). Uma regra de discriminação, de modo geral, é definida pela separação do espaço amostral $\mathbb{R}^p$ em regiões R_j e então se $x \ \epsilon R_j$ alocamos x para o Π_j grupo. (Explicado em [1])

Existem várias regras de discriminação, em [1] são citadas duas famosas, a regra por Máxima Verossimilhança e a regra de Bayes. A regra de Bayes, se baseia na probabilidade condicional(assim como o próprio teorema de bayes) enquanto a regra de verossimilhança se baseia na função de verossimilhança.

- **Regra de Discriminação por Máxima Verossimilhança**: Seja $f_j(x)$ a função densidade da população Π_j, com $j = 1, \ ... \ , n$. A regra de discriminação por máxima verossimilhança aloca a observação x para o grupo com a maior função de verossimilhança, por isso o nome.

- **Regra de Discriminação de Bayes** : Seja π_j a probabilidade inicial de que uma observação seja originária da população Π_j. A regra de discriminação de Bayes aloca a observação x para o grupo Π_i onde o produto $\pi_i f_i(x)$ é o maior dentre todos os outros produtos, $i = 1, \ ... \ , \ n$.

Note que as regras (Bayes e Máxima Verossimilhança) são equivalentes quando $\pi_i = \frac{1}{n}$, ou seja, ambas provém os mesmos resultados.

2.2 Erros de Classificação

Vistas as regras de discriminação, é importante compreender um pouco sobre os erros de classificação, para prevenir que eles aconteçam e atrapalhem a modelagem. Todas as informações apresentadas a seguir nesse tópico 2.2 dizem respeito ao caso de duas populações.

Seja $C(i|k)$ o custo de alocar uma observação x, que pertence ao grupo i, no grupo k, ou seja o custo do erro de alocação. Assim conseguimos construir a seguinte tabela:

Tabela 1: Custo de Erro de Alocação

	População Correta	
População Alocada	**Grupo i**	**Grupo k**
Grupo i	0	$C(i\|k)$
Grupo k	$C(k\|i)$	0

Seguindo esse raciocínio, temos que o custo esperado do erro de alocação é dado por:

$$E(Custo\ do\ Erro\ da\ Alocacao) = P(i|x)C(i|k)\pi_k + P(k|x)C(k|i)\pi_i \tag{1}$$

Portanto, com isso podemos dizer que $x \,\epsilon\, \Pi_i \;\Leftrightarrow\; \frac{\pi_i f_i(x)}{C(i|k)} \;>\; \frac{\pi_k f_k(x)}{C(k|i)}$, caso contrário $x \,\epsilon\, \Pi_k$, assim sendo esta uma outra possível regra de alocação.

Quanto a probabilidade de erro de alocação é dada por:

$$P(Erro\ de\ Alocacao\ em\ qualquer\ grupo) \;=\; \pi_i \int_{\Pi_j} f_i(x)dx \;+\; \pi_j \int_{\Pi_i} f_j(x)dx \tag{2}$$

2.3 Função de Discriminate Linear de Fisher

Supondo que ambas as populações Π_i e Π_k tem variância igual a $\sum$, ou seja $\sum_i \;=\; \sum_k \;=\; \sum$, Sir Ronald Fisher propôs uma função para discriminante. Seja x_o uma observação do vetor de observações $\mathbf{x} \;=\; [x_1,\; x_2,\; ...\;, x_p]$ da variável $\mathbf{X}$ e seja μ_i e μ_k as médias de X_i e X_k a função de discriminante linear de Fisher é dada por:

$$(\mu_i \;-\; \mu_k)^T \sum^{-1} x_o \;=\; \frac{1}{2}(\mu_i \;-\; \mu_k)^T \sum^{-1} (\mu_i \;+\; \mu_k) \;+\; ln(\frac{C(i|k)\pi_k}{C(k|i)\pi_i}) \tag{3}$$

3 Vantagens

Pode ser utilizado como método de redução de dimensão, tem fácil implementação e fácil interpretação. Muito útil em várias áreas do conhecimento e industria, como marketing, saúde e economia. Assim por ter tantas aplicações em tantas áreas diferentes, tem grande vantagem na aplicação dentre alguns métodos de redução de dimensionalidade, como por exemplo a análise fatorial.

4 Desvantagens

Os requisitos de Homocedasticidade(variância constante) das populações e as variáveis contínuas são restrições relativamente fortes, sendo assim um empecilho para a aplicação do método em dados de natureza diferente. Variáveis categóricas não são bem contemplados nesse método, sendo mais uma restrição para o tipo de dado que pode ser aplicado.

Por se tratar de um método de Aprendizado Supervisionado, precisamos ter os grupos pré-definidos para aplicar o método. Este fato pode ser tanto benéfico quanto maléfico, dependendo apenas do propósito inicial do estudo.

5 Análise do Banco de Dados

5.1 Análise Descritiva e Diagnóstico

O banco de dados **penguins** do pacote *palmerpenguins* conta com informações de alguns pinguins de três ilhas do Arquipélago Palmer na Antártida. Ao todo são 344 observações e 8 variáveis, sendo elas:

- **species**: As 3 espécies de pinguins Adélie, Chinstrap e Gentoo.
- **island**: As 3 ilhas Biscoe, Dream e Torgersen.
- **bill-length-mm**: Medida em milimetros do comprimento do bico do pinguim.
- **bill-depth-mm**: Medida em milimetros da profundidade do bico do pinguim.
- **flipper-length-mm**: Medida em milimetros do comprimento da nadadeira do pinguim.
- **body-mass-g**: Medida em gramas do peso(massa corporal) do pinguim.
- **sex**: Sexo do pinguim.
- **year**: O ano em que o dado foi coletado, de 2007 a 2009.

Abaixo segue uma tabela com as estatísticas básicas das variáveis numéricas.

Tabela 2: Estatísticas Básicas dos Dados

	Min	Max	Média	Mediana
bill-length-mm	32.10	59.60	43.92	44.45
bill-depth-mm	13.10	21.50	17.15	17.30
flipper-length-mm	172.0	231.0	200.9	197.0
body-mass-g	2700	6300	4202	4050

Além das informações da tabela, foi constatado 2 valores faltantes em cada uma das variáveis. Outras informações relevantes é a frequência de cada espécie (Adelie: 152, Chinstrap: 68 e Gentoo: 124), ilha (Biscoe: 168, Dream: 124 e Torgersen: 52), sexo (feminino: 165, mascuino: 168 e 11 valores faltantes) e por fim ano (2007: 110, 2008: 114 e 2009: 120).

Para as próximas análises iremos desconsiderar todos os pinguins que tiverem informação faltante nas variáveis que serão usadas, sendo assim usaremos uma amostra de 342 observações. As variáveis **body-mass-g**, **island**, **year** e **sex** não serão utilizadas, porém será aplicado um reescalonamento na variavel **body-mass-g** para que mostre por kg, sendo chamada.

Na figura abaixo temos uma matriz gráfica com as dispersões das variáveis que serão utilizadas, com cada espécie representada por uma cor diferente. verde = Chinstrap, laranja = Adelie, roxo = Gentoo.

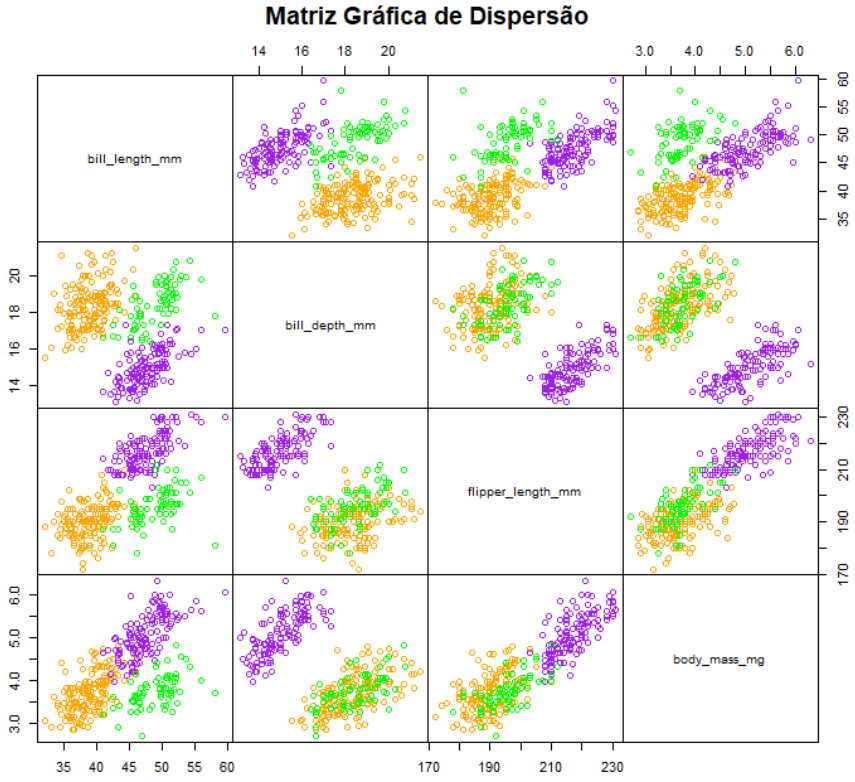

Figura 1: Matriz de Dispersão das Variáveis - por Espécie

A partir da figura, podemos ver que há uma distinção suave entre os grupos quando relacionamos o comprimento do bico com qualquer uma das outras variáveis. Entretanto, a distinção de pinguins Adelie e Chinstrap não é tão simples quando relacionamos a profundidade do bico e o comprimento da nadadeira, diferentemente da espécie Gentoo que é claramente identificada.

Para testar se os dados são homocedasticos sera usado o *Teste de Bartlett*(informações em [3]). A hipótese nula é de que todas as variâncias são iguais, ou seja variância constante(homocedasticidade). O resultado do teste aplicado(o p-valor) é *0.1475*, ou seja não há evidencias para rejeitar a hipóteses nula, portanto temos um indicativo

de que o método de análise de discriminante linear é, no mínimo, razoável para os dados.

5.2 Análise Inferêncial e Exploratória

A aplicação do modelo sera feita duas vezes, uma utilizando todos os dados e outra utilizando uma amostra treino dos dados, que conta com 231 observações do banco original e será utilizada para criação do modelo, enquanto as outras 111 observações foram designadas para uma amostra de teste que será utilizada para previsão. A separação dessas amostras ocorreu por aleatorização, com 70% de chance de ir para a amostra treino e 30% para teste.

Após aplicar o modelo de discriminante linear podemos obter alguns dados interessantes que são mostrados nas tabelas abaixo, como a probabilidade a priori de pertencer a uma espécie e a média de cada grupo em cada variável. Cada tabela a seguir se refere a um modelo, o "modelo total"conta com todos os dados enquanto o "modelo treinado"é o que foi feito com o conjuto de treino.

Tabela 3: Resumo 1 - Modelo Total

		Média do Grupo			
Espécie	**Prob. a Priori**	**bill-length-mm**	**bill-depth-mm**	**flipper-length-mm**	**body-mass-kg**
Adelie	0.442	38.791	18.346	189.954	3.701
Chinstrap	0.199	48.834	18.421	195.824	3.733
Gentoo	0.360	47.505	14.982	217.187	5.076

Em ambas as tabelas vemos que as probabilidades a priori são maiores para Adelie e menores pra Chinstrap. Apesar da separação do banco em dois conjuntos, as características dos dados não sofrem grandes mudanças, seguindo exatamente o que queriamos para o conjunto de treino. Todas as médias mudaram menos de 1.0 .

Tabela 4: Resumo 1 - Modelo Treinado

		Média do Grupo			
Espécie	**Prob. a Priori**	**bill-length-mm**	**bill-depth-mm**	**flipper-length-mm**	**body-mass-kg**
Adelie	0.433	38.730	18.345	190.460	3.723
Chinstrap	0.190	48.434	18.234	194.591	3.626
Gentoo	0.377	47.415	15.031	217.598	5.080

Já nas tabelas abaixo, ainda explorando os resultados dos modelos, temos os coeficientes de discriminação linear, também chamados de "preditores", e a proporção de variabilidade explicada por cada função. Assim temos uma visão mais critica de quais variáveis são mais influentes.

Tabela 5: Resumo 2 - Modelo Total

		Coeficiente de Discriminate			
Função	**Proporção Explicada**	**bill-length-mm**	**bill-depth-mm**	**flipper-length-mm**	**body-mass-kg**
LD1	0.866	0.0883	-1.0373	0.0862	1.2995
LD2	0.134	-0.4179	-0.0210	0.0135	1.7114

Com o banco completo cerca de 86.6% da variabilidade é explicada com a primeira função de discriminante(LD1) e 13.4% com a segunda(LD2). A LD1 tem a massa corporal em kg e a profundidade em mm do bico como os principais preditores(os mais impactantes, sendo o primeiro positivamente e o segundo negativamente). Já a LD2 não se diferencia muito da LD1 no preditor, também conta com a massa corporal como preditor mais importante, porém o comprimento do bico é o outro preditor principal(sendo esse o menor valor).

Tabela 6: Resumo 2 - Modelo Treinado

		Coeficiente de Discriminate			
Função	**Proporção Explicada**	**bill-length-mm**	**bill-depth-mm**	**flipper-length-mm**	**body-mass-kg**
LD1	0.8729	0.0786	-1.0448	0.0917	1.2672
LD2	0.1271	-0.4084	0.0213	0.0144	1.6998

Na tabela acima, que se refere ao modelo treinado, vemos que LD1 explica 87.29% da variabilidade dos dados enquanto a LD2 explica os outros 12.71% para o conjunto de treino. Os coeficientes de discriminante(preditores) mais relevantes seguem os mesmos do modelo com o banco total.

Para uma abordagem mais visual, seguindo a mesma relação de cores e espécies ja apresentada e os modelos apresentados, nas figuras abaixo vemos a relação entre as funções de discriminante e as especies. Assim poderemos ver se os modelos estão conseguindo separar bem mesmo os grupos.

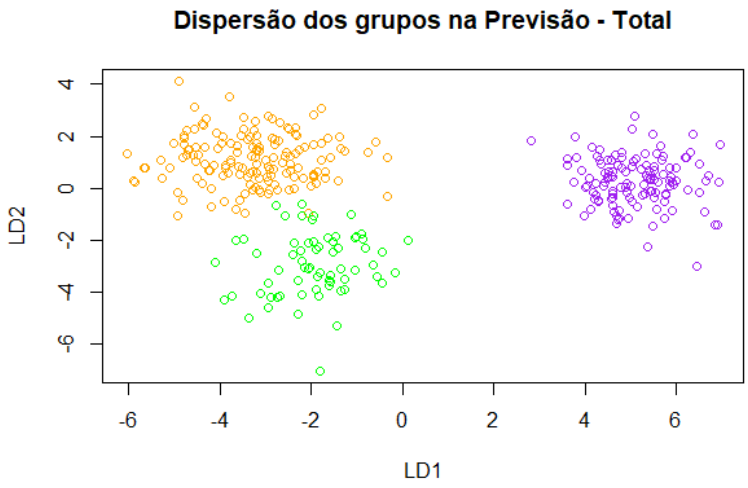

Figura 2: Dispersão das Funções Discriminante - Modelo Treinado

Na figura acima, a que se trata do modelo com todos os dados, vemos que LD1 consegue separar bem a espécie gentoo(roxo), enquanto a LD2 fica responsável por separar as espécies Chinstrap(verde) e Adelie(laranja). Sendo assim conseguimos ver nitidamente que existem 3 grupos distintos, mesmo que Adelie e Chinstrap estejam quase se misturando.

Por fim, na figura abaixo, a que se trata do modelo treinado, vemos um comportamento semelhante ao modelo com todos os dados, LD1 separando Gentoo, LD2 separando as outras duas espécies. Entretanto vemos que LD2 consegue separar um pouco melhor as espécies, sem que nenhum pinguim seja alocado entre muito próximo dos pinguins do outro grupo. Sendo assim, parece ter tido um desempenho melhor.

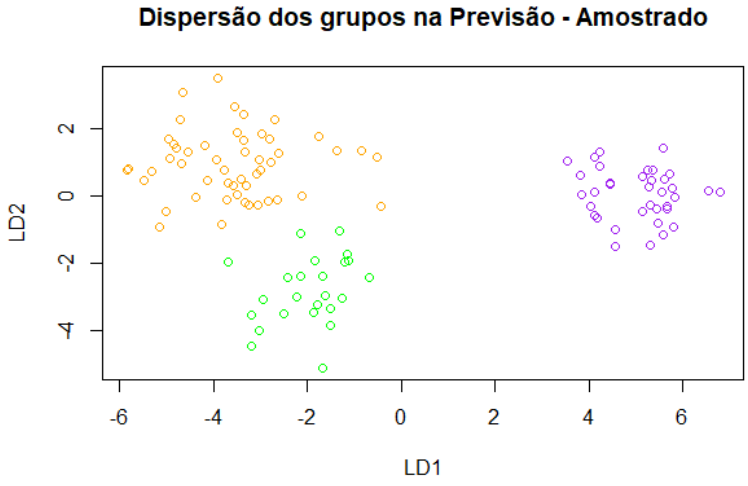

Figura 3: Dispersão das Funções Discriminante - Total

A matriz de confusão nada mais é do que a quantidade de erros na previsão. A acurácia é calculada pela divisão da soma dos elementos da diagonal com a soma de todos os elementos da matriz de confusão. Assim, abaixo segue as duas Matrizes de confusão das previsões, uma usando todos os dados e a outra usando a amostra de teste:

Tabela 7: Matriz de Confusão - Banco Completo

	População Correta		
População Alocada	**Adelie**	**Chinstrap**	**Gentoo**
Adelie	150	3	0
Chinstrap	1	65	0
Gentoo	0	0	123

No modelo usando todos os dados, podemos ver que houveram alguns erros na previsão, 3 pinguins Chinstrap foram alocados como Adelie, e 1 Adelie foi alocado como Chinstrap. Por causa disso, a acurácia do modelo ficou em 0.988, o que é bem alto e mostra que o modelo captou muito bem as características dos grupos.

Tabela 8: Matriz de Confusão - Modelo Treinado

	População Correta		
População Alocada	**Adelie**	**Chinstrap**	**Gentoo**
Adelie	51	0	0
Chinstrap	0	24	0
Gentoo	0	0	36

O modelo que foi amostrado em treino e teste teve um desempenho ainda melhor, não houveram erros de alocação então todos os pinguins foram alocados em suas respectivas espécies. A acurácia foi máxima, ou seja atingiu o 1.000, o que mostra que o modelo capturou perfeitamente as características das espécies.

5.3 Discussão dos Resultados

Sobre os modelos, temos duas formas visuais de analisar, cada uma com informação específica. Pelos gráficos de dispersão entre as funções de discriminante lineares, podemos ver que ambos os modelos conseguiram separar os dados em 3 conjuntos específicos(as espécies). Já pelas tabelas, vimos a proporção da variância explicada, vimos que a variável mais impactante foi a que se referia ao peso corporal em quilogramas, para ambas as funções.

Em relação as previsões dos modelos, aquele que usou todos os dados acabou tendo o pior desempenho, entretanto ainda assim foi um ótimo desempenho com acurácia 0.988 e apenas 4 erros de alocação. O outro modelo conseguiu atingir a perfeição com a acurácia 1.000, ou seja nenhum erro.

6 Conclusão

A análise de discriminante linear se mostrou um ótimo método, de facil implementação e entendimento, entretanto possui condições fortes(homocedasticidade e possivelmente normalidade) para os dados que será aplicada. É uma otima tecnica de aprendizado supervisionado e que mostrou resultados incriveis nos diagnósticos e implementações.

Pelo diagnóstico dos dados vimos que é sim aplicável o método, a partir da estatística do teste de Bartlett para homocedasticidade. Já pelo diagnostico dos modelos de previsão vimos que ambos tiveram ótimos desempenhos, capturaram muito bem as características das espécies. Entretanto deve-se levar em consideração que o modelo que houve a separação de amostras, teve uma quantidade menor de dados para teste, sendo isso uma possível causa da previsão perfeita(acurácia = 1.000).

A escolha do método se mostrou muito boa para os dados, visto que mesmo com algumas variáveis categóricas(o que poderia ser uma desvantagem para a técnica), que não foram consideradas para os modelos, o fato de já sabermos os grupos ajudou na escolha. Como a intenção era identificar as 3 diferentes espécies, a técnica serviu perfeitamente.

7 Referências Bibliográficas

[1] Härdle, W. K. e Simar, L. (2019). Applied Multivariate Statistical Analysis. Quinta Edição, Springer Nature.

[2] Johnson, R. A. e Wichern, D. W. (2007). Applied Multivariate Statistical Analysis. Sexta Edição, Prentice Hall, Nova Jersey.

[3] Arsham, H. e Lovric, M. (2011). International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg.