Análise Multivariada - Atividade Final

Ricardo Limongi

2025-05-01

Contents

Introdução

Nesta atividade avaliativa, você irá aplicar três técnicas de análise multivariada a um conjunto de dados de saúde para responder perguntas de pesquisa. As técnicas a serem utilizadas são:

1. **Análise Discriminante Linear (LDA)**
2. **Análise de Cluster**
3. **Análise Fatorial**

O objetivo é integrar as técnicas para obter insights sobre os padrões nos dados e auxiliar na tomada de decisão clínica.

Base de Dados: Doença Cardíaca

Nesta atividade, utilizaremos o conjunto de dados "Heart Disease" do UCI Machine Learning Repository, que contém informações de pacientes com suspeita de doença cardíaca.

```
# Carregar os pacotes necessários
pacotes_necessarios <- c(
  # Para manipulação de dados
  "tidyverse", "dplyr", "readr",

  # Para análise discriminante
  "MASS", "caret", "klaR",

  # Para análise de cluster
  "cluster", "factoextra", "NbClust",

  # Para análise fatorial
  "psych", "corrplot", "lavaan", "semPlot",

  # Para visualizações
  "ggplot2", "gridExtra", "psych",

  # Para visualização de texto com repulsão
  "ggrepel", "tidyverse", "magrittr"
)

# Definir o mirror do CRAN
options(repos = c(CRAN = "https://cloud.r-project.org"))

# Instalar e carregar pacotes se necessário
for (pacote in pacotes_necessarios) {
  if (!require(pacote, character.only = TRUE)) {
    install.packages(pacote)
    library(pacote, character.only = TRUE)
  } else {
    library(pacote, character.only = TRUE)
  }
}
```

```r
# Carregar os dados do repositório UCI
url <- "https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data"

# Nomes das colunas baseados na documentação do UCI
colunas <- c("age", "sex", "cp", "trestbps", "chol", "fbs", "restecg",
             "thalach", "exang", "oldpeak", "slope", "ca", "thal", "target")

# Carregar os dados
dados_heart <- read.csv(url, header = FALSE, sep = ",",
                        col.names = colunas, na.strings = "?")

# Transformar variáveis categóricas em fatores
dados_heart$sex <- factor(dados_heart$sex, levels = c(0, 1),
                          labels = c("Female", "Male"))
dados_heart$cp <- factor(dados_heart$cp, levels = c(1, 2, 3, 4),
                         labels = c("Typical Angina", "Atypical Angina",
                                    "Non-anginal Pain", "Asymptomatic"))
dados_heart$fbs <- factor(dados_heart$fbs, levels = c(0, 1),
                          labels = c("False", "True"))
dados_heart$restecg <- factor(dados_heart$restecg, levels = c(0, 1, 2),
                              labels = c("Normal", "ST-T abnormality",
                                         "LV hypertrophy"))
dados_heart$exang <- factor(dados_heart$exang, levels = c(0, 1),
                            labels = c("No", "Yes"))
dados_heart$slope <- factor(dados_heart$slope, levels = c(1, 2, 3),
                            labels = c("Upsloping", "Flat", "Downsloping"))
dados_heart$thal <- factor(dados_heart$thal, levels = c(3, 6, 7),
                           labels = c("Normal", "Fixed Defect", "Reversible Defect"))

# A variável alvo (target) indica a presença de doença cardíaca
# 0 = ausência, 1-4 = presença (vários graus)
dados_heart$target <- ifelse(dados_heart$target > 0, 1, 0)
dados_heart$target <- factor(dados_heart$target, levels = c(0, 1),
                             labels = c("Healthy", "Disease"))

# Converter a variável ca para fator após tratar valores ausentes
dados_heart$ca <- as.numeric(dados_heart$ca)
dados_heart$ca <- factor(dados_heart$ca)

# Verificar dados carregados
glimpse(dados_heart)
```

```
## Rows: 303
## Columns: 14
## $ age      <dbl> 63, 67, 67, 37, 41, 56, 62, 57, 63, 53, 57, 56, 56, 44, 52, 5~
## $ sex      <fct> Male, Male, Male, Male, Female, Male, Female, Female, Male, M~
## $ cp       <fct> Typical Angina, Asymptomatic, Asymptomatic, Non-anginal Pain,~
## $ trestbps <dbl> 145, 160, 120, 130, 130, 120, 140, 120, 130, 140, 140, 140, 1~
## $ chol     <dbl> 233, 286, 229, 250, 204, 236, 268, 354, 254, 203, 192, 294, 2~
## $ fbs      <fct> True, False, False, False, False, False, False, False, False,~
## $ restecg  <fct> LV hypertrophy, LV hypertrophy, LV hypertrophy, Normal, LV hy~
## $ thalach  <dbl> 150, 108, 129, 187, 172, 178, 160, 163, 147, 155, 148, 153, 1~
## $ exang    <fct> No, Yes, Yes, No, No, No, No, Yes, No, Yes, No, No, Yes, No, ~
```

```
## $ oldpeak  <dbl> 2.3, 1.5, 2.6, 3.5, 1.4, 0.8, 3.6, 0.6, 1.4, 3.1, 0.4, 1.3, 0~
## $ slope    <fct> Downsloping, Flat, Flat, Downsloping, Upsloping, Upsloping, D~
## $ ca       <fct> 0, 3, 2, 0, 0, 0, 2, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0~
## $ thal     <fct> Fixed Defect, Normal, Reversible Defect, Normal, Normal, Norm~
## $ target   <fct> Healthy, Disease, Disease, Healthy, Healthy, Healthy, Disease~
```

```
# Verificar valores ausentes
sum(is.na(dados_heart))
```

```
## [1] 6
```

```
# Remover linhas com valores ausentes
dados_heart_clean <- na.omit(dados_heart)

# Verificar os dados após limpeza
summary(dados_heart_clean)
```

```
##       age            sex                      cp         trestbps
##  Min.   :29.00   Female: 96   Typical Angina  : 23   Min.   : 94.0
##  1st Qu.:48.00   Male  :201   Atypical Angina : 49   1st Qu.:120.0
##  Median :56.00                Non-anginal Pain: 83   Median :130.0
##  Mean   :54.54                Asymptomatic    :142   Mean   :131.7
##  3rd Qu.:61.00                                       3rd Qu.:140.0
##  Max.   :77.00                                       Max.   :200.0
##       chol          fbs                  restecg       thalach      exang
##  Min.   :126.0   False:254   Normal          :147   Min.   : 71.0   No :200
##  1st Qu.:211.0   True : 43   ST-T abnormality:  4   1st Qu.:133.0   Yes: 97
##  Median :243.0               LV hypertrophy  :146   Median :153.0
##  Mean   :247.4                                      Mean   :149.6
##  3rd Qu.:276.0                                      3rd Qu.:166.0
##  Max.   :564.0                                      Max.   :202.0
##     oldpeak             slope     ca                   thal
##  Min.   :0.000   Upsloping  :139   0:174   Normal           :164
##  1st Qu.:0.000   Flat       :137   1: 65   Fixed Defect     : 18
##  Median :0.800   Downsloping: 21   2: 38   Reversible Defect:115
##  Mean   :1.056                     3: 20
##  3rd Qu.:1.600
##  Max.   :6.200
##      target
##  Healthy:160
##  Disease:137
##
##
##
##
```

Descrição das Variáveis

- **age**: Idade em anos
- **sex**: Sexo (1 = masculino; 0 = feminino)
- **cp**: Tipo de dor torácica (1 = angina típica; 2 = angina atípica; 3 = dor não-anginal; 4 = assintomático)
- **trestbps**: Pressão arterial em repouso (em mm Hg)

- **chol**: Colesterol sérico (em mg/dl)
- **fbs**: Açúcar no sangue em jejum > 120 mg/dl (1 = verdadeiro; 0 = falso)
- **restecg**: Resultados eletrocardiográficos em repouso
- **thalach**: Frequência cardíaca máxima alcançada
- **exang**: Angina induzida por exercício (1 = sim; 0 = não)
- **oldpeak**: Depressão ST induzida por exercício em relação ao repouso
- **slope**: Inclinação do segmento ST de pico do exercício
- **ca**: Número de vasos principais coloridos por fluoroscopia (0-3)
- **thal**: Resultado do teste de estresse com tálio (3 = normal; 6 = defeito fixo; 7 = defeito reversível)
- **target**: Diagnóstico de doença cardíaca (0 = ausência; 1 = presença)

Perguntas de Pesquisa

Você deverá aplicar as técnicas multivariadas para responder às seguintes perguntas de pesquisa:

Pergunta Principal: Como podemos identificar, agrupar e caracterizar pacientes com diferentes perfis de risco cardiovascular, integrando métodos de análise multivariada?

Perguntas Específicas:

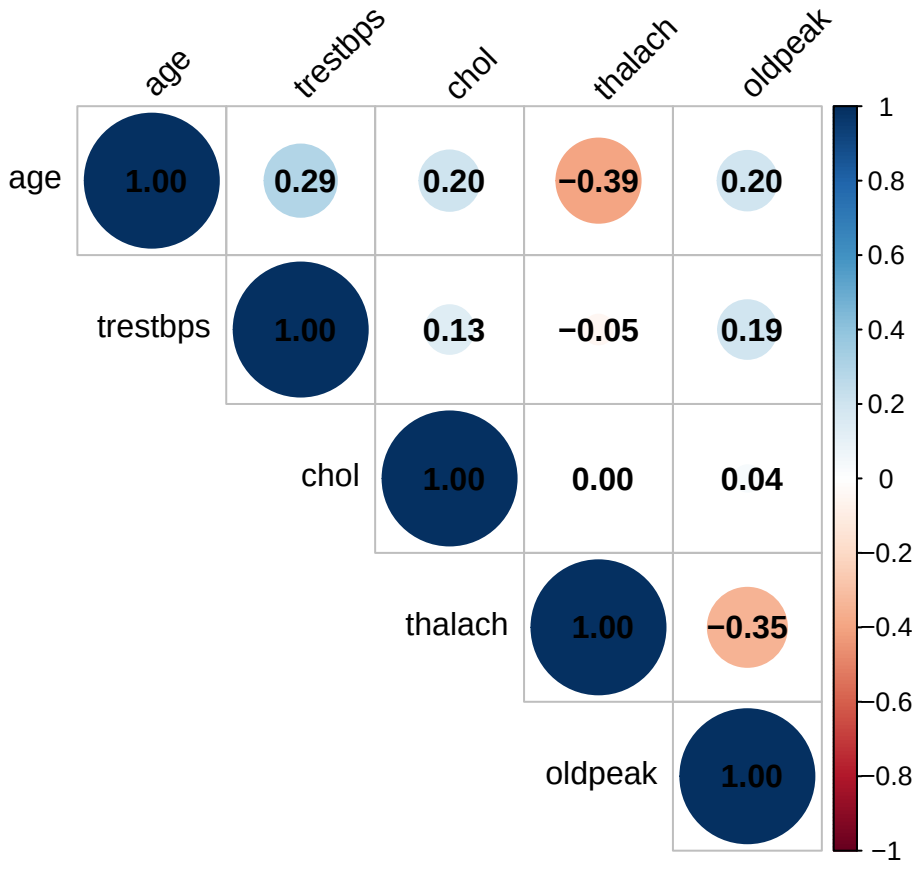
1. **Análise Discriminante (LDA):** Quais variáveis são mais relevantes para discriminar entre pacientes com e sem doença cardíaca? É possível criar um modelo de classificação eficaz usando essas variáveis?
2. **Análise de Cluster:** É possível identificar subgrupos naturais (clusters) de pacientes com perfis de risco cardiovascular semelhantes? Como esses grupos se relacionam com o diagnóstico de doença cardíaca?
3. **Análise Fatorial:** Quais fatores latentes (não diretamente observáveis) podem ser identificados? Como esses fatores se relacionam com o risco cardiovascular?
4. **Integração:** Como as três técnicas podem ser combinadas para fornecer uma visão mais completa do perfil de risco cardiovascular dos pacientes e auxiliar na tomada de decisão clínica?

Parte 1: Análise Discriminante Linear (LDA)

Aplique a Análise Discriminante Linear para identificar as variáveis que melhor discriminam pacientes com e sem doença cardíaca, e criar um modelo de classificação.

```
# Separar variáveis numéricas e categóricas
var_num <- select_if(dados_heart_clean, is.numeric)
var_cat <- select_if(dados_heart_clean, is.factor)

# Verificar correlações entre variáveis numéricas
cor_matrix <- cor(var_num)
corrplot(cor_matrix, method = "circle", type = "upper",
         tl.col = "black", tl.srt = 45, addCoef.col = "black")
```




```r
# Dividir dados em treino (70%) e teste (30%)
set.seed(123)
indices_treino <- createDataPartition(dados_heart_clean$target, p = 0.7, list = FALSE)
dados_treino <- dados_heart_clean[indices_treino, ]
dados_teste <- dados_heart_clean[-indices_treino, ]
```

Exercício 1.1: Aplicar Análise Discriminante Linear

Aplique a LDA para discriminar entre pacientes saudáveis e com doença cardíaca.

```r
# TAREFA: Construa o modelo LDA
# Dica: Use a função lda() do pacote MASS para criar o modelo
# Selecione as variáveis mais relevantes baseando-se na análise de correlação

# TAREFA: Analise os coeficientes e médias por grupo
# Dica: Examine modelo_lda$scaling e modelo_lda$means

# TAREFA: Visualize a função discriminante
# Dica: Use as funções ldahist() do pacote MASS ou ggplot2

# TAREFA: Faça predições no conjunto de teste
# Dica: Use a função predict()

# TAREFA: Calcule e visualize a matriz de confusão
# Dica: Use as funções table() ou confusionMatrix() do pacote caret
```

```
# TAREFA: Calcule métricas de desempenho (acurácia, sensibilidade, especificidade)
```

Exercício 1.2: Interpretação da LDA

Com base nos resultados da Análise Discriminante Linear, responda às seguintes perguntas:

1. Quais variáveis mais contribuem para a discriminação entre pacientes com e sem doença cardíaca?
2. Qual a acurácia do modelo LDA na classificação de novos pacientes?
3. Quais as implicações clínicas desses resultados para o diagnóstico de doença cardíaca?

Parte 2: Análise de Cluster

Aplique técnicas de cluster para identificar subgrupos naturais de pacientes com perfis de risco cardiovascular semelhantes.

```
# Selecionar apenas variáveis numéricas para o clustering

# Verificar a estrutura dos dados selecionados
```

Exercício 2.1: Determinação do Número Ideal de Clusters

Determine o número ideal de clusters usando diferentes métodos.

```
# TAREFA: Determine o número ideal de clusters
# Dica: Use os métodos do cotovelo e silhueta
# Utilize as funções do pacote factoextra (fviz_nbclust)

# Método do cotovelo

# Método da silhueta
```

Exercício 2.2: Aplicação do K-means

Aplique o algoritmo K-means com o número ideal de clusters determinado.

```
# TAREFA: Aplique o algoritmo K-means
# Dica: Use a função kmeans() com o número de clusters determinado anteriormente

# TAREFA: Adicione a informação de cluster ao dataset original

# TAREFA: Analise a relação entre os clusters e o diagnóstico de doença cardíaca
# Dica: Use table() para criar uma tabela de contingência

# Proporção de doença em cada cluster
```

Exercício 2.3: Caracterização dos Clusters

Caracterize os clusters identificados em termos das variáveis originais.

```
# TAREFA: Calcule as estatísticas descritivas para cada cluster
# Dica: Use group_by() e summarise() do dplyr

# Exibir o perfil dos clusters

# TAREFA: Visualize as características de cada cluster
# Dica: Use boxplots ou heatmaps para comparar os clusters

# Idade e FC máxima

# Colesterol e Pressão

# Visualizar juntos

# Distribuição da doença por cluster

# TAREFA: Interprete os resultados e nomeie cada cluster com base em suas características
```

Exercício 2.4: Interpretação da Análise de Cluster

Com base nos resultados da Análise de Cluster, responda às seguintes perguntas:

1. Quantos clusters foram identificados e quais são suas principais características?
2. Como os clusters se relacionam com o diagnóstico de doença cardíaca?
3. Quais as implicações clínicas dessa segmentação para o manejo de pacientes cardíacos?

Parte 3: Análise Fatorial

Aplique a Análise Fatorial para identificar fatores latentes que expliquem os padrões de correlação observados nos dados.

```
# Selecionar variáveis contínuas relevantes para a análise fatorial


# Verificar a matriz de correlação
```

Exercício 3.1: Adequação dos Dados para Análise Fatorial

Verifique se os dados são adequados para a Análise Fatorial.

```
# TAREFA: Verifique a adequação dos dados para análise fatorial
# Dica: Use o teste KMO e o teste de esfericidade de Bartlett

# Teste KMO (Kaiser-Meyer-Olkin)

# Teste de esfericidade de Bartlett
```

```
# TAREFA: Determine o número adequado de fatores
# Dica: Use o critério de Kaiser (autovalores > 1) e o scree plot

# Análise paralela (alternativa mais estável ao fa.parallel)
```

Exercício 3.2: Aplicação da Análise Fatorial

Aplique a Análise Fatorial com o número adequado de fatores.

```
# TAREFA: Execute a análise fatorial
# Dica: Use a função fa() do pacote psych com rotação varimax ou oblimin

# TAREFA: Visualize as cargas fatoriais
# Dica: Use print(modelo_fa$loadings, cutoff=0.3)

# Alternativa à função fa.diagram() que pode causar problemas

# TAREFA: Interprete os fatores identificados
# Dica: Examine quais variáveis têm cargas altas em cada fator
```

Exercício 3.3: Cálculo dos Escores Fatoriais

Calcule os escores fatoriais e adicione-os ao dataset.

```
# TAREFA: Calcule os escores fatoriais
# Dica: Use a função factor.scores() para calcular os escores

# TAREFA: Adicione os escores fatoriais ao dataset original

# TAREFA: Analise a relação entre os fatores e o diagnóstico de doença cardíaca
# Dica: Compare os escores fatoriais entre pacientes com e sem doença cardíaca

# Visualização dos escores por diagnóstico

# Comparação estatística
```

Exercício 3.4: Interpretação da Análise Fatorial

Com base nos resultados da Análise Fatorial, responda às seguintes perguntas:

1. Quais fatores latentes foram identificados e como podem ser interpretados?
2. Como esses fatores se relacionam com o risco cardiovascular?
3. Quais as implicações clínicas desses fatores para a compreensão da doença cardíaca?

Parte 4: Integração das Técnicas

Nesta parte, você deverá integrar os resultados das três técnicas multivariadas para obter insights mais profundos sobre os padrões presentes nos dados.

Exercício 4.1: Combinação das Análises

Combine os resultados das três técnicas para criar uma visão integrada do perfil de risco cardiovascular dos pacientes.

```
# TAREFA: Crie um dataset integrado com os resultados das três análises
# Dica: Combine os clusters, escores fatoriais e predições da LDA


# TAREFA: Analise as relações entre os resultados das diferentes técnicas
# Dica: Examine como os clusters se relacionam com os fatores e com a classificação da LDA

# Visualizar clusters no espaço dos fatores

# Relação entre clusters e predição LDA

# TAREFA: Visualize essas relações
# Dica: Use gráficos de dispersão, heatmaps ou outros tipos de visualização apropriados

# Visualização integrada
```

Exercício 4.3: Relatório Final

Escreva um relatório final (máximo 1000 palavras) integrando os resultados das três análises e respondendo à pergunta principal de pesquisa. O relatório deve incluir:

1. Uma breve introdução ao problema de pesquisa e às técnicas utilizadas
2. Os principais resultados de cada técnica
3. Como esses resultados se complementam e o que revelam sobre o perfil de risco cardiovascular dos pacientes
4. Implicações para a prática clínica e para a gestão em saúde
5. Limitações da análise e sugestões para pesquisas futuras

Boa análise! “‘