Estatística Multivariada

Trabalho No.3

Entregar até o dia 15 de junho de 2023.

1- Dados de alta dimensão são difíceis de explorar e visualizar. Assim, a ideia principal da Análise de Componentes Principais é reduzir o número de variáveis em um conjunto de dados preservando o máximo de informações possível.

Com base nos dados a seguir, qual marca de pizza é a melhor para você? Deve ser difícil avaliar todas as marcas e seus conteúdos nutricionais de uma só vez.

Dados obtidos como:

```
> pizza = read.csv(
file = "http://leg.ufpr.br/~lucambio/CE090/20231S/pizza.csv",
header = TRUE, sep = ",")
```

O conjunto de dados se parece com o seguinte.

```
> head(pizza)
brand    id  mois  prot   fat  ash sodium carb  cal
1     A 14069 27.82 21.43 44.87 5.11   1.77 0.77 4.93
2     A 14053 28.49 21.26 43.89 5.34   1.79 1.02 4.84
3     A 14025 28.35 19.99 45.78 5.08   1.63 0.80 4.95
4     A 14016 30.55 20.15 43.13 4.79   1.61 1.38 4.74
5     A 14005 30.49 21.28 41.65 4.82   1.64 1.76 4.67
6     A 14075 31.14 20.23 42.31 4.92   1.65 1.40 4.67
```

Este quadro de dados contém 300 observações nas 9 variáveis a seguir.

brand marca

id identificador da pizza amostrada

mois quantidade de água

prot quantidade de proteína

fat quantidade de gordura

ash quantidade de cinzas

sodium quantidade de sódio

carb quantidade de carboidratos

cal quantidade de calorias

Estas quantidades referem-se à quantidade correspondente por 100 gramas da amostra.

a) Quantas componentes principais na análise você selecionaria, porquê?

b) Com base nos resultados obtidos é justo dizer que uma componente principal representa pizzas ricas em gordura, cinzas e sódio e pobre em carboidratos, enquanto alguma outra representa pizzas ricas em calorias e pobre em umidade/água. Visualize os resultados em um gráfico de carregamentos.

c) Agora é hora de usar os resultados para decidir qual marca funciona melhor para você. Para fazer isso, apresenta graficamente as componentes principais das amostras de pizza e as cargas variáveis em um único gráfico, que é chamado de biplot na terminologia.

d) Queremos responder quais marcas são mais adequadas segundo a escolha de consumo:

   i) Se você quer uma pizza gordurosa e crocante, qual marca seria uma boa opção?

   ii) Se você quer uma pizza macia e sem gordura, qual marca seria uma boa escolha?

   iii) Se você preferir um equilíbrio de nutrientes, quais marcas podem ser as escolhas adequadas?

2- A Análise Fatorial e a Análise de Componentes Principais são técnicas de redução de dimensionalidade. O principal objetivo da Análise Fatorial não é reduzir a dimensionalidade dos dados. A análise fatorial é uma abordagem útil para encontrar variáveis latentes que não são medidas diretamente em uma única variável, mas sim inferidas de outras variáveis no conjunto de dados. Essas variáveis latentes são chamadas de fatores. Assim, a análise fatorial é um modelo de mensuração de variáveis latentes e sua principal suposição é que existem tais variáveis latentes em nossos dados.

Os dados que usamos aqui são os Recordes Nacionais de Mulheres, representando 55 países em sete eventos diferentes. Podem ser obtidos como a seguir:

```
exemplo = read.csv(
file = "http://leg.ufpr.br/~lucambio/CE090/20231S/NTRforW.csv",
sep = ",")
```

O conjunto de dados se parece com o seguinte.

```
> head(exemplo)
COUNTRY    X1    X2    X3   X4   X5   X6     X7
1 'Argentina' 11.61 22.94 54.50 2.15 4.43 9.79 178.52
2 'Australia' 11.20 22.35 51.80 1.98 4.13 9.08 152.37
3   'Austria' 11.43 23.09 50.62 1.99 4.22 9.34 159.37
4   'Belgium' 11.41 23.04 52.00 2.00 4.14 8.88 157.85
5   'Bermuda' 11.46 23.05 53.30 2.16 4.58 9.81 169.98
6    'Brazil' 11.31 23.17 52.80 2.10 4.49 9.77 168.75
```

A descrição das variáveis é a seguinte.

X1 100m (s)

X2 200m (s)

X3 400m (s)

X4 800m (s)

X5 1500m (min)

X6 3000m (min)

X7 Marathon (min)

a) Faça a transformação das variáveis à mesma unidade de medida, sugerimos que os tempos dos recordes nacionais sejam medidos em segundos.

b) Encontre as cargas fatoriais. As cargas fatoriais originais produzem fatores interpretáveis?

c) Realize a rotação de fatores, que é uma transformação ortogonal dos fatores originais. A rotação de fatores é feita para fins de interpretação. Pretendemos encontrar fatores que representem os recordes de curta distância e longa distância. Portanto, caso seja possível, queremos poder dar nomes relevantes para os dois fatores como segue como fator de velocidade e fator de tolerância ou resistência. Quais as componentes desses fatores?

3- Cluster é um grupo de objetos que pertencem à mesma classe. Em outras palavras, objetos semelhantes são agrupados em um cluster e objetos diferentes são agrupados em outro cluster. Assim, a clusterização é o processo de transformar um grupo de objetos abstratos em classes de objetos semelhantes.

Um cluster de objetos de dados pode ser tratado como um grupo. Ao fazer a análise de cluster, primeiro particionamos o conjunto de dados em grupos com base na similaridade de dados e, em seguida, atribuímos os rótulos aos grupos. A principal vantagem do agrupamento sobre a classificação é que ele é adaptável a mudanças e ajuda a destacar recursos úteis que distinguem diferentes grupos.

Aplicações da Análise de Cluster incluem pesquisas no campo da biologia, na qual pode ser usado para derivar taxonomias de plantas e animais, categorizar genes com funcionalidades semelhantes e obter informações sobre estruturas inerentes às populações. O agrupamento também ajuda na identificação de áreas de uso semelhante da terra em um banco de dados de observação da Terra.

Uma pesquisa de campo foi realizada em pinguins Pygoscelis nidificando em várias ilhas dentro do Arquipélago Palmer perto da Ilha Anvers, Antártica, durante os verões austrais de 2007/08, 2008/09 e 2009/10.

Esta pesquisa for conduzida pela The Palmer Long-Term Ecological Research (LTER). A Estação Palmer é uma das três estações de pesquisa dos Estados Unidos localizadas na Antártica.

Queremos identificar a espécie utilizando bill_length_mm, bill_depth_mm e o comprimento da nadadeira (flipper_length_mm). Podem ser identificadas as espécies de pinguins sem considerar o sexo do animal? Para responder a estas questões faça una Análise de Cluster com a seguinte base de dados:

```
> penguins = read.csv(
file = "http://leg.ufpr.br/~lucambio/CE090/20231S/penguins.csv",
header = TRUE, sep = ",")
```

O conjunto de dados se parece com o seguinte.

```
> str(penguins)
'data.frame': 344 obs. of  9 variables:
$ X                : int  1 2 3 4 5 6 7 8 9 10 ...
$ species          : chr  "Adelie" "Adelie" "Adelie" "Adelie" ...
$ island           : chr  "Torgersen" "Torgersen" "Torgersen" ...
$ bill_length_mm   : num  39.1 39.5 40.3 NA 36.7 39.3 38.9 ...
$ bill_depth_mm    : num  18.7 17.4 18 NA 19.3 20.6 17.8 19.6 ...
$ flipper_length_mm: int  181 186 195 NA 193 190 181 195 193 ...
$ body_mass_g      : int  3750 3800 3250 NA 3450 3650 3625 ...
$ sex              : chr  "male" "female" "female" NA ...
$ year             : int  2007 2007 2007 2007 2007 2007 2007 ...
```

Existem 3 espécies (species) diferentes de pinguins neste conjunto de dados. Ainda foi coletado o sexo (sex) de cada indivíduo amostrado e o peso (body_mass_g).

O culmen é a crista superior do bico de um pássaro, o comprimento e a profundidade do culmen são renomeados como variáveis bill_length_mm e bill_depth_mm para serem mais intuitivos. Para os dados destes pinguins, o comprimento e a profundidade do culmen (bico) são medidos conforme mostrado abaixo.