Métodos Estatísticos Multivariados

Trabalho No.4

Entregar até o dia 29 de junho de 2023.

1- A Análise de Correlação Canônica é um método para explorar as relações entre dois conjuntos multivariados de variáveis (vetores), todos medidos no mesmo indivíduo.

Considere um grupo de representantes de vendas, no qual registramos várias variáveis de desempenho de vendas juntamente com várias medidas de aptidão intelectual e criativa. Podemos querer explorar as relações entre as variáveis de desempenho de vendas e as variáveis de aptidão.

Uma abordagem para estudar as relações entre os dois conjuntos de variáveis é usar a análise de correlação canônica que descreve a relação entre o primeiro conjunto de variáveis e o segundo conjunto de variáveis. Não pensamos necessariamente em um conjunto de variáveis como independente e o outro como dependente, embora essa possa ser uma outra abordagem.

Os dados neste caso vêm de uma empresa que pesquisou uma amostra aleatória de $n = 50$ de seus funcionários na tentativa de determinar quais fatores influenciam o desempenho de vendas.

Duas coleções de variáveis foram medidas:

Performance de vendas:

V1 crescimento das vendas

V2 rentabilidade das vendas

V3 vendas de novas contas

Pontuações de testes como medida de inteligência

V4 criatividade

V5 raciocínio mecânico

V6 raciocínio abstrato

V7 matemática

Existem $p = 3$ variáveis no primeiro grupo relacionadas ao desempenho de vendas e $q = 4$ variáveis no segundo grupo relacionadas aos resultados dos testes.

Para cada grupo, dois terços da amostra foram usados para estimativa e um terço da amostra para teste.

Dados obtidos como:

```
> vendas = read.table("http://leg.ufpr.br/~lucambio/CE090/20221S/sales.txt")
```

O conjunto de dados se parece com o seguinte.

```
> head(vendas)
     V1    V2    V3 V4 V5 V6 V7
1  93.0  96.0  97.8  9 12  9 20
2  88.8  91.8  96.8  7 10 10 15
3  95.0 100.3  99.0  8 12  9 26
4 101.3 103.8 106.8 13 14 12 29
5 102.0 107.8 103.0 10 15 12 32
6  95.8  97.5  99.3 10 14 11 21
```

Responda às seguintes questões:

(a) Encontre as variáveis canônicas, combinações lineares ortogonais das variáveis dentro de cada conjunto que melhor explicam a variabilidade dentro e entre conjuntos.

(b) Encontre p-valor das cargas canônicas utilizando as aproximações F de diferentes testes estatísticos. Quais dimensões canônicas são estatisticamente significativas?

(c) Interpreta as variáveis canônicas.

2- A Análise Discriminante constrói um modelo preditivo para membros de grupos. O modelo é composto por uma função discriminante ou, para mais de dois grupos, um conjunto de funções discriminantes baseadas em combinações lineares das variáveis preditoras que fornecem a melhor discriminação entre os grupos. As funções são geradas a partir de uma amostra de casos para os quais a associação aos grupos é conhecida; as funções podem então ser aplicadas a novos casos que possuem medidas para as variáveis preditoras, mas que possuem associação de grupo desconhecida.

Dados de Insetos. Os dados foram coletados em duas espécies de insetos do gênero Chaetocnema, (a) concinna e (b) heikertlingeri. Variável V1.

Dados obtidos como:

```
> insect = read.table("http://leg.ufpr.br/~lucambio/CE090/20221S/insect.txt")
```

O conjunto de dados se parece com o seguinte.

```
> head(insect)
  V1  V2  V3 V4
1  a 191 131 53
2  a 185 134 50
3  a 200 137 52
4  a 173 127 50
5  a 171 128 49
6  a 160 118 47
```

Três variáveis foram medidas em cada inseto:

V2 largura da 1ª articulação do tarso (pernas)

V3 largura da 2ª articulação do tarso

V4 largura do edeago (órgão reprodutor)

Nosso objetivo é obter uma regra de classificação para identificar as espécies de insetos com base nessas três variáveis. Um entomologista pode identificar essas duas espécies intimamente relacionadas, mas as diferenças são tão sutis que é preciso ter uma experiência considerável para poder dizer a diferença.

Se uma regra de classificação puder ser desenvolvida, essa pode ser uma maneira mais precisa de ajudar a diferenciar essas duas espécies diferentes. Utilize a técnica discriminante para encontrar uma regra de classificação.

3- A mineração de dados é uma etapa crítica na descoberta de conhecimento envolvendo teorias, metodologias e ferramentas para revelar padrões nos dados. É importante entender a lógica por trás dos métodos para que as ferramentas e os métodos tenham um ajuste adequado aos dados e ao objetivo do reconhecimento de padrões. Pode haver várias opções de ferramentas disponíveis para um conjunto de dados.

Quando um banco recebe um pedido de empréstimo, com base no perfil do solicitante, o banco deve tomar uma decisão sobre a aprovação ou não do empréstimo. Dois tipos de riscos estão associados à decisão do banco.

Se o requerente for um bom risco de crédito, ou seja, provavelmente pagará o empréstimo, a não aprovação do empréstimo à pessoa resultará em perda de negócios para o banco. Caso o requerente for um risco de crédito ruim, ou seja, provavelmente não pagará o empréstimo, a aprovação do empréstimo à pessoa resultará em uma perda financeira para o banco.

O objetivo da análise é a minimização do risco e maximização do lucro em nome do banco. Para minimizar a perda do ponto de vista do banco, o banco precisa de uma regra de decisão sobre para quem deve aprovar o empréstimo e quem não deve. Os perfis demográficos e socioeconômicos de um candidato são considerados pelos gerentes de empréstimo antes que uma decisão seja tomada em relação ao seu pedido de empréstimo.

Os dados de crédito alemães contêm dados sobre 20 variáveis e a classificação se um candidato é considerado um risco de crédito bom ou ruim para 1.000 solicitantes de empréstimo. Espera-se que um modelo preditivo desenvolvido com base nesses dados forneça ao gerente do banco uma orientação para tomar a decisão de aprovar ou não um empréstimo a um candidato em potencial com base em seus perfis.

As seguintes abordagens analíticas são sugeridas:

- Regressão logística, a resposta é binária (1 = bom risco de crédito ou 0 = ruim) e diversos preditores são avaliados.
- Análise discriminante.
- Métodos de random forest.

Dados obtidos como:

```
> # Conjunto de dados de treinamento
> Treino = read.csv("http://leg.ufpr.br/~lucambio/CE090/20231S/Training50.csv")
> # Conjunto de dados de teste
> Teste = read.csv("http://leg.ufpr.br/~lucambio/CE090/20231S/Test50.csv")
```

A primeiras linhas e poucas colunas do conjunto de dados de treinamento se parece com o seguinte.

```
> head(treino)[,1:4]
    X Creditability Account.Balance Duration.of.Credit..month.
1 497             1               3                         6
2 756             0               1                        15
3 580             0               1                        42
4 833             0               3                        36
5 602             1               3                        24
6 734             1               1                        15
```

A variável Creditability = 1 significa que o crédito foi atribuído ao candidato, caso contrário, Creditability = 0 negou-se o crédito. As outras variáveis servem de preditoras

Utilizando estas informações aplique modelos discriminantes (LDA, QDA, etc.) nos dados de treinamento para encontrar um modelo preditivo que forneça ao gerente do banco uma orientação para tomar a decisão de aprovar ou não um empréstimo a um candidato em potencial com base em seus perfis. Os dados de teste servem para avaliar as probabilidades de erros de classificação (Misclassification rate).