**Tópicos em Estatística 2 - 01/2024**
**Aprendizado Supervisionado e Não Supervisionado**
**Prof. George von Borries**
**Atividades para Entrega**

---

## Introdução

**Data de Divulgação: 26/03/24**

**Entrega 1 - Problemas 1 e 2**

**Data de Entrega: até 15/04/24**

1. Escolha uma área de pesquisa de interesse (engenharia, medicina, economia, ecologia, computação ou outra área de interesse). Para cada tipo de problema da lista abaixo, apresente um artigo publicado em revista indexada e indique as características do estudo que o fazem relacionar o artigo ao problema em questão. Indique pontos fortes e fracos de sua formação em estatística para realizar estudos semelhantes.
   (a) Análise Estatística Paramétrica.
   (b) Análise Estatística Não Paramétrica.
   (c) Reconhecimento de Padrões.
   (d) Aprendizado de Máquinas e/ou Estatístico Supervisionado.
   (e) Aprendizado de Máquinas e/ou Estatístico não Supervisionado.

   **Observação 1:** Um estudo pode abordar dois ou mais problemas, entre os listados.

   **Observação 2:** Uma lista de publicações na área de estatística pode ser encontrada em `https://www.stata.com/links/statistics-and-general-journals/`.

   Algumas referências tem acesso livre, como por exemplo:

   - Brazilian Journal of Biometrics: `https://biometria.ufla.br/index.php/BBJ`
   - Chilean Journal of Statistics: `https://soche.cl/chjs/`
   - Significance: `https://academic.oup.com/jrssig?login=false`

   **Observação 3:** Sua resposta deve ser resumida e completa. Anexe ou indique o caminho para acessar cada um dos artigos. Este exercício pode ser realizado em grupos de até 4 alunos.

2. Considere um hipercubo de dimensão $r$ e lados de comprimento $2A$. Dentro deste hipercubo temos uma hiperesfera $r$-dimensional de raio $A$. Encontre a proporção do volume do hipercubo que está fora da hiperesfera e mostre que a proporção tende a 1 a medida que a dimensão $r$ cresce. Escreva um programa `R` para verificar o resultado encontrado. O que este resultado significa?

Teoria da Decisão e Funções Discriminantes

**Data de Divulgação: 26/03/24**

**Entrega 2 - Problemas 3 a 5 - a entrega 2 pode ser feita em grupos de até 4 alunos.**

**Data de Entrega: até 30/04/24**

3. Várias são as possibilidades de geração de variáveis com distribuição normal multivariada $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Diferentes algoritmos seguem três passos principais que são

   - Fazer a decomposição de $\boldsymbol{\Sigma}$ em $\mathbf{D}$
   - Gerar uma matriz $\mathbf{Z}_{n\times p}$ de valores da $\mathrm{N}(0,1)$
   - Gerar $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{Z} * \mathbf{D}$

   Exemplos de algoritmos que seguem estes passos estão no programa **`geranp.R`**, disponível no site do curso e são descritos a seguir.

   **(A) Algoritmo para Geração por Decomposição Espectral**

   - $\boldsymbol{\Sigma}$ positiva definida $\Rightarrow$ $\boldsymbol{\Sigma}^{1/2}$ positiva definida.
   - $\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Gamma}$.
   - A decomposição espectral será $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{\Gamma}$.
   - Se fizermos $\mathbf{Q} = \boldsymbol{\Sigma}^{1/2}$, então a fatorização de $\boldsymbol{\Sigma}$ é $\boldsymbol{\Sigma} = \mathbf{Q}^{\mathrm{T}}\mathbf{Q}$.

     Nota: Lembre que $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}} = \mathbf{I}$ e $\boldsymbol{\Gamma}^{\mathrm{T}} = \boldsymbol{\Gamma}^{-1}$.

   O código `R` será,

```
rmvn.eigen <-
  function(n, mu, Sigma) {
    p <- length(mu)
    ev <- eigen(Sigma, symmetric = TRUE)
    lambda <- ev$values
    V <- ev$vectors
    R <- V %*% diag(sqrt(lambda)) %*% t(V)
    Z <- matrix(rnorm(n*p), nrow = n, ncol = p)
    X <- Z %*% R + matrix(mu, n, p, byrow = TRUE)
    X
  }
```

   **(B) Algoritmo para Geração por Decomposição em Valores Singulares**

   - $\boldsymbol{\Sigma} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}$.
   - $\mathbf{U}$: autovetores de $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\mathrm{T}}$.
   - $\mathbf{S}$: diagonal da raiz dos autovalores de $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\mathrm{T}}$.
   - $\mathbf{V}$: autovetores de $\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{\Sigma}$.

     Como $\boldsymbol{\Sigma}$ simétrica e positiva definida,
     $\mathbf{U} = \mathbf{V}$ e $\boldsymbol{\Sigma}^{1/2} = \mathbf{U}\mathbf{D}^{1/2}\mathbf{V}^{\mathrm{T}} \equiv$ método espectral, porém menos eficiente.

E o código R será então,

```
rmvn.svd <-
function(n, mu, Sigma) {
  p <- length(mu)
  S <- svd(Sigma)
  R <- S$u %*% diag(sqrt(S$d)) %*% t(S$v)
  Z <- matrix(rnorm(n*p), nrow=n, ncol=p)
  X <- Z %*% R + matrix(mu, n, p, byrow=TRUE)
  X  }
```

### (C) Algoritmo para Geração por Decomposição de Cholesky

– $\boldsymbol{\Sigma} = \mathbf{Q}^{\mathrm{T}}\mathbf{Q}$, sendo $\mathbf{Q}$ triangular superior.

e implementando o seguinte código R:

```
rmvn.cholesky <-
  function(n, mu, Sigma) {
    p <- length(mu)
    Q <- chol(Sigma)
    Z <- matrix(rnorm(n*p), nrow=n, ncol=p)
    X <- Z %*% Q + matrix(mu, n, p, byrow=TRUE)
    X
  }
```

O pacote `mvrnorm` do R utiiza estas técnicas e outros algoritmos para gerar valores de variáveis com distribuição normal multivariada.

Problema:

Considere duas classes com distribuição normal multivarida tal que $\omega_1 \sim \mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ com

$$\boldsymbol{\mu} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{e} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

e $\omega_2 \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ com

$$\boldsymbol{\mu} = \begin{bmatrix} -1 \\ 0 \end{bmatrix} \quad \text{e} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

(a) Gere 100 valores para $\omega_1$ e $\omega_2$.

(b) Verifique se os valores gerados seguem distribuição $\mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Lembre que neste caso, o par deve seguir uma distribuição $\chi^2$ e cada variável deve ter distribuição Normal.

(c) Para um determinado $\mu$ na razão de verossimilhança, determine as regiões $\Omega_1$ e $\Omega_2$ na regra de Neyman-Pearson.

(d) Considere diferentes valores $\mathbf{x} = [x_1, x_2]^{\mathrm{T}}$ e utilize a regra de decisão de Bayes para alocar estes valores em $\Omega_1$ ou $\Omega_2$.

4. Considere duas classes com distribuições multivariadas tal que $p(\mathbf{x}|\omega_1) \sim \mathrm{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ e $p(\mathbf{x}|\omega_2) \sim \mathrm{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. Mostre que o logaritmo da razão de verossimilhança é linear em relação ao vetor de características $\mathbf{x}$.

5. (Extra) Pequise sobre pacotes disponíveis no `R` para realizar análise de discriminantes e classificação. Verifique as regras de decisão utilizadas nestes pacotes. Compare os recursos do `R` com procedimento em outra linguagens de programação, como `Python`, `Matlab`, `SAS`, `Julia`.

## Discriminantes Lineares

**Data de Divulgação: 15/04/24**

**Entrega 3 - Problemas 6 a 10 - a entrega 3 pode ser feita em grupos de até 4 alunos.**

**Data de Entrega: até 16/05/24**

6. Estudar o texto de Kneusel (2022) sobre o algoritmo Gradiente Descendente[1] e apresentar um resumo com exemplos em `R`.

### 1. Sobre o Algoritmo Perceptron, Critério de Fisher e Critério de Mínimos Quadrados

7. Estudar e apresentar um exemplo utilizando as notas sobre Perceptron no `R` descritas em `https://rpubs.com/FaiHas/197581`.

8. (a) Estudar e apresentar a função `perceptron` do pacote `mlpack` do `R`.
   (b) Verificar se existe material adicional (instruções em sites, pacotes, artigos) sobre classificação utilizando o algoritmo perceptron no `R` e apresentar exemplos.
   (c) (Extra) Pesquisar e listar as funções referentes ao algoritmo perceptron que estão presentes em alguma outra linguagem, como `Python`, `Matlab`, `SAS`, `Julia`.

9. Apresentar recursos `R` para uso do critério de Fisher na classificação de padrões. Incluir uma breve discussão sobre vantagens e desvantagens da implementação em `R` e um exemplo. (Extra) Compare com os recursos disponíveis em alguma outra linguagem de programação (`Python`, `Matlab`, `SAS`, `Julia`).

10. Apresentar recursos `R` para uso do critério de Mínimos Quadrados na classificação de padrões. Incluir uma breve discussão sobre vantagens e desvantagens da implementação em `R` e um exemplo. (Extra) Compare com os recursos disponíveis em alguma outra linguagem de programação (`Python`, `Matlab`, `SAS`, `Julia`).

[1] Disponível no site do curso.

## 2. Sobre a Regressão Logística

**Data de Divulgação: 24/04/24**

**Entrega 4 - esta entrega é individual.**

**Data de Entrega: até 30/05/24**

11. Pesquisar funções disponíveis em pacotes `R` para classificação utilizando a função logística. Apresentar um pequeno exemplo do uso das funções. Destacar vantagens e desvantagens em relação aos pacotes de Modelos Lineares Generalizados apresentados em aula.

    Exemplos de pacotes para classificação no `R`: `caret`, `class`, `mlpack`.

## Discriminantes Lineares e Não Lineares - Máquinas de Suporte Vetorial

**Data de Divulgação: 13/05/24**

**Entrega 5 - Entrega pode ser em grupos de até 4 alunos.**

**Data de Entrega: até 13/06/24**

12. Revise as notas de aula e estude o Capítulo 9 de James et al. (*with Applications in R* ou *with Applications in Python*), disponível em `https://www.statlearning.com/`. Resolva os exercícios deste capítulo.

13. Escolha uma linguagem de programação (`R`, `Python`, `SAS`, `Matlab`, `Julia`) e apresente um exemplo de classificação com SVM utilizando Kernel Linear e outro utilizando Kernel não Linear.

## Reamostragem e Estimação de Densidades

**Data de Divulgação: 06/06/24**

**Entrega 6 - entrega em grupos de até 4 alunos**

**Data de Entrega: 27/06/24**

**Observação:** Nas questões seguintes, explicar cada passo dos exemplos apresentados.

14. Estudar o pacote `rsample` em https://rsample.tidymodels.org/ e apresentar um exemplo utilizando validação cruzada e Bootstrap.

15. Selecionar ou gerar um conjunto de dados e comparar a classificação após estimação de densidades utilizando os seguintes métodos:

    1. Método do Histograma
    2. Estimação baseada em Núcleos
    3. k-Vizinhos mais Próximos

16. Estudar o pacote `ks` do `R` e apresentar um exemplo.

17. Apresentar um exemplo com classificador LDA e QDA.

Modelos de Mistura Finita / Árvores de Decisão

**Data de Divulgação: 11/06/24**

**Entrega 7 - grupos previamente definidos**

**Data de Entrega: indicado no exercício**

18. Apresentações sobre Classificação em GMM e Árvores de Decisão.

    Exemplo de classificação com mistura de normais - 27/06/24.
    (César, Gabriela, João Vitor, Kevyn)

    Terminologia de árvores de decisão - 27/06/24.
    (Jéssica, Luís Felipe, Maria Luiza, Rayssa)

    Técnica bagging em árvores de decisão - 02/07/24.
    (João Victor, Pedro, Rafael, Vinícius)

    Técnica Random Forest em árvores de decisão - 02/07/24.
    (Bruno, João Alberto, João Pedro, Stefan)