



DEPARTAMENTO DE ESTATÍSTICA

28 março 2024

Lista 1

Prof. Dr. George von Borries

Aluno: Bruno Gondim Toledo

Matrícula: 15/0167636

Tópicos 2

1º/2024

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.

Nome do artigo: **Spatial autocorrelation analysis of multivariate rural insurance data**

Link para o artigo: <https://biometria.ufla.br/index.php/BBJ/article/view/642/385>

Tópico abordado: **Reconhecimento de Padrões.**

O artigo busca utilizar técnicas multivariadas como análise de componentes principais e Análise Exploratória de Dados Espaciais para estudar dependência de seguros rurais de acordo com regiões geográficas brasileiras. A parte computacional do artigo foi feita em Python, mas poderia ser facilmente realizada em R e talvez em SAS também. A graduação e estatística da UnB oferta matérias para capacitação em todas essas três linguagens. A técnica de análise de componentes principais também é estudada na disciplina Análise Multivariada 1 na graduação, enquanto técnicas de estatística espacial é por vezes ofertada, enquanto disciplina optativa. Acredito ser plenamente possível um aluno de graduação de estatística da UnB replicar este artigo.

Nome do artigo: **Ordinal data and residual analysis: Review and application**

Link para o artigo: <https://biometria.ufla.br/index.php/BBJ/article/view/637/380>

Tópicos abordados: **Análise Estatística Paramétrica, Análise Estatística Não Paramétrica e Aprendizado de Máquinas e/ou Estatístico não Supervisionado**

O artigo busca fazer inferência paramétrica para um modelo logístico utilizando estimadores de máxima verossimilhança, seguido de análise de resíduos utilizando o teste não-paramétrico de Kolmogorov-Smirnov. O objetivo do estudo é estudar o dano no fígado de Tambaquis, de acordo com o genoma pré conhecido do animal. Nominalmente, todas as técnicas são ensinadas nas disciplinas obrigatórias da graduação de estatística: Inferência estatística; Análise de dados categorizados; Métodos estatísticos 2. A aplicação computacional também seria possível cursando as matérias obrigatórias CE1 e CE2. Um reforço no aprendizado da técnica de modelagem logística seria possível com o curso optativo oferecido Modelos Lineares Generalizados. Acredito que, respeitada a complexidade do artigo, seria possível um aluno de graduação de estatística da UnB replicar este artigo.

Nome do artigo: **Classification and Analysis of Patients with COVID-19 Using Machine Learning**

Link para o artigo: <https://biometria.ufla.br/index.php/BBJ/article/view/588/362>

Tópico abordado: **Aprendizado de Máquinas e/ou Estatístico não Supervisionado.**

O artigo busca classificar pacientes com COVID-19 utilizando algoritmos de aprendizado estatístico, como SVM e Naive-Bayes, entre outros. Técnicas computacionais sobre utilização desses algoritmos são citados nos cursos de CE2 – R e CE2 – Python, além de Estatística Computacional. Algumas matérias de Tópicos (1 e 2) também abordam estes. Com alguma pesquisa, é possível um aluno de graduação elaborar um artigo como este. Inclusive, eu mesmo cheguei a fazer algo relacionado a pedido do meu estágio, criando um código para realizar processamento de linguagem natural e posterior classificação com alguns algoritmos. Disponível em https://github.com/penasta/NLP/blob/main/funcao_nlp.ipynb

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?

Verificando se aumentar o comprimento A altera o resultado

```
for (A in 1:3){  
  r = 2  
  volume_hiperesfera <- (pi^(r/2) * A^r) / gamma(r/2 + 1)  
  volume_hipercubo <- (2*A)^r  
  prop = (volume_hipercubo-volume_hiperesfera)/volume_hipercubo  
  print("A Proporção do volume do hipercubo que está fora da hiperesfera para r = 2 fixo")  
  print(paste("e A =",A,"é:", prop))  
}
```

```
## [1] "A Proporção do volume do hipercubo que está fora da hiperesfera para r = 2 fixo"  
## [1] "e A = 1 é: 0.214601836602552"  
## [1] "A Proporção do volume do hipercubo que está fora da hiperesfera para r = 2 fixo"  
## [1] "e A = 2 é: 0.214601836602552"  
## [1] "A Proporção do volume do hipercubo que está fora da hiperesfera para r = 2 fixo"  
## [1] "e A = 3 é: 0.214601836602552"
```

É fácil perceber que o comprimento A não altera a proporção do volume do hipercubo que está fora da hiperesfera.

Verificando agora se a dimensão r altera o resultado, conforme solicitado.

```
for (r in seq(2,34,2)){
  A = 1
  volume_hiperesfera <- (pi^(r/2) * A^r) / gamma(r/2 + 1)
  volume_hipercubo <- (2*A)^r
  prop = (volume_hipercubo-volume_hiperesfera)/volume_hipercubo
  print("A Proporção do volume do hipercubo que está fora da hiperesfera para A = 1 fixo")
  print(paste("e r =",r,"é:", prop))
}
```

```
## [1] "A Proporção do volume do hipercubo que está fora da hiperesfera para A = 1 fixo"
## [1] "e r = 2 é: 0.214601836602552"
## [1] "A Proporção do volume do hipercubo que está fora da hiperesfera para A = 1 fixo"
## [1] "e r = 4 é: 0.691574862465958"
## [1] "A Proporção do volume do hipercubo que está fora da hiperesfera para A = 1 fixo"
## [1] "e r = 6 é: 0.919254487811719"
## [1] "A Proporção do volume do hipercubo que está fora da hiperesfera para A = 1 fixo"
## [1] "e r = 8 é: 0.984145655756185"
## [1] "A Proporção do volume do hipercubo que está fora da hiperesfera para A = 1 fixo"
## [1] "e r = 10 é: 0.997509605429807"
## [1] "A Proporção do volume do hipercubo que está fora da hiperesfera para A = 1 fixo"
## [1] "e r = 12 é: 0.999674008113073"
## [1] "A Proporção do volume do hipercubo que está fora da hiperesfera para A = 1 fixo"
## [1] "e r = 14 é: 0.999963423795818"
## [1] "A Proporção do volume do hipercubo que está fora da hiperesfera para A = 1 fixo"
## [1] "e r = 16 é: 0.999996409139551"
## [1] "A Proporção do volume do hipercubo que está fora da hiperesfera para A = 1 fixo"
## [1] "e r = 18 é: 0.999999686638311"
## [1] "A Proporção do volume do hipercubo que está fora da hiperesfera para A = 1 fixo"
## [1] "e r = 20 é: 0.999999975388631"
## [1] "A Proporção do volume do hipercubo que está fora da hiperesfera para A = 1 fixo"
## [1] "e r = 22 é: 0.99999998242752"
## [1] "A Proporção do volume do hipercubo que está fora da hiperesfera para A = 1 fixo"
## [1] "e r = 24 é: 0.99999999884988"
## [1] "A Proporção do volume do hipercubo que está fora da hiperesfera para A = 1 fixo"
## [1] "e r = 26 é: 0.99999999993052"
## [1] "A Proporção do volume do hipercubo que está fora da hiperesfera para A = 1 fixo"
## [1] "e r = 28 é: 0.9999999999961"
## [1] "A Proporção do volume do hipercubo que está fora da hiperesfera para A = 1 fixo"
## [1] "e r = 30 é: 0.9999999999998"
## [1] "A Proporção do volume do hipercubo que está fora da hiperesfera para A = 1 fixo"
## [1] "e r = 32 é: 0.99999999999999"
## [1] "A Proporção do volume do hipercubo que está fora da hiperesfera para A = 1 fixo"
## [1] "e r = 34 é: 1"
```

Desta, percebemos que o aumento da dimensão r faz com que a proporção do volume da hiperesfera em relação ao volume do hipercubo tenda a 1. Isso significa que, conforme a dimensão do hipercubo aumenta, a hiperesfera se torna cada vez mais insignificante em relação ao hipercubo. Isso é um resultado interessante, pois mostra que, conforme a dimensão do espaço aumenta, a hiperesfera se torna cada vez mais rara em relação ao hipercubo.

Generalizar este resultado, seria pensar que a hiperesfera era uma informação, uma característica de interesse no estudo, e conforme aumentamos o tamanho da dimensão, ou seja, a quantidade de features, a informação de interesse “sumiu”.