Jeidsan A. da C. Pereira

Estatística e Ciência de Dados

Notas e solução dos exercícios

Conteúdo

Prefácio

Esta página contém notas e solução para os exercícios propostos no livro **Estatística e Ciência de Dados**, de autoria de Pedro Alberto Morettin e Júlio da Motta Singer, publicado pela LTC em 2022 [Morettin and Singer, 2022].

É importante destacar que trata-se de um produto não oficial, as anotações e soluções de exercícios aqui apresentadas são de cunho pessoal e não possuem qualquer revisão ou análise por parte dos autores da obra ou da editora. Dessa forma e por se tratar de um produto construído durante o processo de aprendizagem, o conteúdo pode conter erros, tanto no texto em si, como na lógica utilizada para solução dos exercícios.

Dúvidas ou sugestões de melhoria podem ser encaminhadas para o e-mail *jeidsan.pereira@gmail.com*[1].

Pendências

-

[1] mailto:jeidsan.pereira@gmail.com

1

Estatística, Ciência de Dados e Megadados

1.1 Introdução

Atualmente, os termos *Data Science* (**Ciência de Dados**) e *Big Data* (**Megadadas**) são utilizados em profusão como se envolvessem conceitos novos, distintos daqueles com que os estatísticos lidam há cerca de dois séculos [Morettin and Singer, 2022, p. 1].

1.2 Aprendizado com estatística

O **aprendizado supervisionado** está relacionado com metodologia desenvolvida essencialmente para **previsão** e **classificação**. No âmbito da previsão, o objetivo é utilizar **variáveis preditivas** (sexo, classe social, renda, por exemplo) observadas em várias **unidades** (clientes de um banco, por exemplo) para "advinhar" valores de uma **variável resposta numérica** (saldo médio, por exemplo) de novas unidades. O problema de classificação consiste em qual categoria de uma **variável resposta qualitativa** (bons e maus pagadores, por exemplo) as novas unidades são classificadas [Morettin and Singer, 2022, p. 3].

> No **aprendizado não supervisionado**, dispomos apenas um conjunto de dados, sem distinção entre preditoras e respostas, e o objetivo é descrever **associações** e **padrões** entre essas variáveis e **agrupá-las** com o objetivo de identificar características comuns e conjuntos de unidades de investigação ou desenvolver métodos para combiná-las e assim **reduzir sua dimensionalidade** [Morettin and Singer, 2022, p. 3].

> Além de aprendizado supervisionado e não supervisionado, podemos acrescentar um terceiro tipo, denominado **aprendizado com reforço** (*reinforcement learning*), segundo o qual um algoritmo "aprende" a realizar determinadas tarefas por meio de repetições com o fim de maximizar um prêmio sujeito a um valor máximo [Morettin and Singer, 2022, p. 3].

> Embora tanto o aprendizado supervisionado quanto o aprendizado com reforço utilizem um mapeamento entre entradas (*inputs*) e saídas (*outputs*), no primeiro caso a retroalimentação (*feedback*) fornecida ao algoritmo é um conjunto de ações corretas necessárias para a realização de uma tarefa; no aprendizado com reforço, por outro lado, a retroalimentação é baseada num sistema com prêmios e punições como indicativos de ações corretas ou incorretas [Morettin and Singer, 2022, p. 3].

1.3 Aprendizado automático

Jordan [2019 *apud* Morettin and Singer, 2022, p. 4] distingue três tipos de inteligência artificial: i) inteligência artificial imitativa humana; ii) aumento de inteligência; e iii) infraestrutura inteligente.

> De modo informal, a inteligência artificial está relacionada com um esforço para automatizar tarefas intelectuais usualmente realizadas por seres humanos (Chollet, 2018) e consequentemente, intimamente ligada ao desenvolvimento da computação (ou programação de computadores) [Morettin and Singer, 2022, p. 4].

> Convém ressaltar que o objetivo do aprendizado automático não é o mesmo daquele consi- derado na análise de regressão usual, em que se pretende entender como cada variável preditora Ro está associada com a variável resposta. O objetivo do aprendizado automático é selecionar o mo- delo que produz melhores previsões, mesmo que as variáveis selecionadas com essa finalidade não sejam aquelas consideradas numa análise padrão [Morettin and Singer, 2022, p. 5].

1.4 Uma cronologia do desenvolvimento da estatística

Sem notas para esta seção.

1.5 Notação e tipos de dados

Sem notas para esta seção.

1.6 Paradigmas para o aprendizado com estatística

Sem notas para esta seção.

1.7 Este livro

> Independentemente do volume de dados disponíveis para análise, Ciência de Dados é uma atividade multidisciplinar que envolve: i) um problema a ser resolvido com questões claramente especificadas; ii) um conjunto de dados (seja ele volumoso ou não); iii) os meios para sua obtenção; iv) sua organização; v) a especificação do problema original em termos das variáveis desse conjunto de dados; vi) a descrição e resumo dos dados à luz do problema a ser resolvido; vii) a escolha das técnicas estatísticas apropriadas para a resolução desse problema; viii) os algoritmos computacionais necessários para a implementação dessas técnicas; ix) a apresentação dos resultados [Morettin and Singer, 2022, p. 11].

1.8 Conjuntos de dados

Sem notas para esta seção.

1.9 Notas do capítulo

Sem notas para esta seção.

Parte I

Análise Exploratória de Dados

2

Preparação dos dados

2.1 Considerações preliminares

> O ramo da Estatística conhecido como **Análise Exploratória de Dados** se ocupa da organização e resumo dos dados de uma amostra ou, eventualmente, de toda a população e o ramo conhecido como **Inferência Estatística** se refere ao processo de se tirar conclusães sobre uma população com base em uma amostra dela [Morettin and Singer, 2022, p. 21].

2.2 Planilhas de dados

Sem notas para esta seção.

2.3 Contrução de tabelas

Sem notas para esta seção.

2.4 Construção de gráficos

Sem notas para esta seção.

2.5 Notas de capítulo

Sem notas para esta seção.

2.6 Exercícios

Exercício 2.1

O objetivo de um estudo da Faculdade de Medicina da USP foi avaliar a associação entre a quantidade de morfina administrada a pacientes com dores intensas provenientes de lesões medulares ou radiculares e a dosagem dessa substância em seus cabelos. Três medidas foram realizadas em cada paciente, a primeira logo após o início do tratamento e as demais após 30 e 60 dias. Detalhes podem ser obtidos no documento disponível no arquivo morfina.doc.

A planilha morfina.xls, disponível no arquivo morfina foi entregue ao estatístico para análise e contém resumos de características demográficas além dos dados do estudo.

a) Com base nessa planilha, apresente um dicionário com a especificação das variáveis segundo as indicações da Seção 2.2 e construa a planilha correspondente.
b) Com as informações disponíveis, construa tabelas para as variáveis sexo, raça, grau de instrução e tipo de lesão segundo as sugestões da Seção 2.3.

Solução. x

Exercício 2.2

A Figura 2.6 foi extraída de um estudo sobre atitudes de profissionais de saúde com rela- ção a cuidados com infecção hospitalar. Critique-a e reformule-a para facilitar

sua leitura, lembrando que a comparação de maior interesse é entre as diferentes categorias profissionais.

Solução. x

Exercício 2.3

Utilize as sugestões para construção de planilhas apresentadas na Seção 2.2 com a finalidade de preparar os dados do arquivo empresa para análise estatística.

Solução. x

Exercício 2.4

Num estudo planejado para avaliar o consumo médio de combustível de veículos em dife- rentes velocidades foram utilizados 4 automóveis da marca A e 3 automóveis da marca B selecionados ao acaso das respectivas linhas de produção. O consumo (em~L/km) de cada um dos 7 automóveis foi observado em 3 velocidades diferentes (40 km/h, 80 km/h e 110~km/h) Delineie uma planilha apropriada para a coleta e análise estatística dos dados, rotulando-a adequadamente.

Solução. x

Exercício 2.5

Utilizando os dados do arquivo enforco, prepare uma planilha Excel num formato conve niente para análise pelo R. Inclua apenas as variáveis Idade, Altura, Peso, Frequência cardiaca e VO2 no repouso além do quociente VE/VCO2, as correspondentes porcentagens relativa mente ao máximo, o quociente VO2/FC no pico do exercício e data do óbito. Importe a pla nilha Excel que você criou utilizando comandos R e obtenha as características do arquivo importado (número de casos, número de observações omissas etc.)

Solução. x

Exercício 2.6

A Figura 2.7 contém uma planilha encaminhada pelos investigadores responsáveis por um estudo sobre AIDS para análise estatistica. Organize-a de forma a permitir sua análise por meio de um pacote computacional como o R.

Solução. x

Exercício 2.7

A planilha apresentada na Figura 2.8 contém dados de um estudo em que o limiar auditivo foi avaliado nas orelhas direita (OD) e esquerda (OE) de 13 pacientes em 3 ocasiões (Limiar, Teste 1 e Teste 2). Reformate-a segundo as recomendações da Seção 2.2, indicando claramente

a) a definição das variáveis.
b) os rótulos para as colunas da planilha.

Solução. x

Exercício 2.8

A planilha disponível no arquivo cidades contém informações demográficas de 3554 8 municípios brasileiros.

a) Importe-a para permitir a análise por meio do software R, indicando os problemas encontrados nesse processo além de sua solução.
b) Use o comando summary para obter um resumo das variáveis do arquivo.
c) Classifique cada variável como numérica ou alfanumérica e indique o número de observações omissas de cada uma delas.

Solução. x

Exercício 2.9

Preencha a ficha de inscrição do Centro de Estatistica Aplicada (www.ime.usp.br/-cea) com as informações de um estudo em que você está envolvido.

Solução. x

3

Análise de dados de uma variável

3.1 Introdução

3.2 Distribuição de frequências

3.3 Medidas resumo

3.4 *Boxplots*

3.5 Modelos probabilísticos

3.6 Dados amostrais

3.7 Gráficos QQ

3.8 Desvio padrão e erro padrão

3.9 Intervalo de confiança e tamanho da amostra

3.10 Transformação de variáveis

3.11 Notas de capítulo

3.12 Exercícios

4

Análise de dados de duas variáveis

5

Análise de dados de várias variáveis

6

Análise de Regressão

7

Análise de Sobrevivência

Parte II

Aprendizado Supervisionado

8

Regularização e Modelos Aditivos Generalizados

9

Classificação por meio de técnicas clássicas

10

Algoritmos de Suporte Vetorial

11

Árvores e Florestas

12

Redes neurais

Parte III

Aprendizado não Supervisionado

13

Análise de Agrupamentos

14

Redução de dimensionalidade

A

Otimização numérica

B

Noções de simulação

C

Algoritmos para dados aumentados

Bibliografia

Pedro Alberto Morettin and Julio da Motta Singer. *Estatística e Ciência de Dados*. LTC, Rio de Janeiro, 2022.