

Aprendizagem de Máquina

Giseldo Neo

Contents

Prefácio	5
Sobre	5
O autor	5
1 Introdução	7
2 Classificação do dado	9
2.1 Dado e informação	9
2.2 Tipo do dado (Qualitativo ou Quantitativo)	9
2.3 Tipo do dado (Preditor, Alvo)	11
2.4 Tabelas	12
3 Estimativas de localização	15
3.1 Média	16
4 Análise exploratória dos dados	17
4.1 Introdução	17
5 Python	23
6 Referências	25
 ## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 -- ## v dplyr 1.1.4 v readr 2.1.5 ## v forcats 1.0.0 v stringr 1.5.1 ## v ggplot2 3.5.0 v tibble 3.2.1 ## v lubridate 1.9.3 v tidyr 1.3.1	

```
## v purrr      1.0.2
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()   masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
```

Prefácio

Este livro visa dar ao leitor uma visão prática e teórica sobre aprendizagem de máquina e estatística com exemplos de programação em R e em Python.

Sobre

Todos os direitos reservados, proibida a redistribuição

O autor

Giseldo Neo é Mestre em Modelagem Computacional e Professor.

Chapter 1

Introdução

Thomas H. Davenport e DJ Patil publicaram uma matéria de opinião em 2012 na conceituada revista de administração e negócios Harvard Business Review afirmando que um dos empregos mais sexy do século 21 seria o cientista de dados (termo cunhado pelo próprio DJ Patil) e deu um exemplo do crescimento em visualizações do linkedin quando foi utilizado o recurso de pessoas que talvez você conheça, recurso este que utilizava a extensa base de dados de conexões do linkedin para fazer um cruzamento do tipo: se uma pessoa qualquer conhece fulano e fulano conhece ciclano então essa pessoa qualquer tem grande chance de conhecer ciclano [Review, 2012]. Porém, acredito que outras pessoas possam ter visões diferentes do que é um emprego sexy, como por exemplo Joel Grus (2016) [Grus, 2016].

Particularmente, a escolha do que é ser sexy foge do escopo desse livro. Porém, existe um relativo consenso de que o campo está em evidência. Muitas pesquisas estão sendo escritas sobre esse tema, com experimentos e achados importantes. Logo, a habilidade em lidar com esses dados e reportá-los é crucial para o profissional que deseja extrair informação útil e construir conhecimento.

Vamos começar pelos conceitos básicos e criar um vocabulário consistente nas seções seguintes.

Chapter 2

Classificação do dado

Vamos classificar o dado em categorias que nos permitirão uma comunicação mais consistente e com menos redundância. Essas classificações facilitarão a nossa comunicação.

2.1 Dado e informação

Dado e informação são coisas diferentes. Os dados são os fatos brutos. Por exemplo, o nome de um estudante e o número do CPF são exemplos de dados brutos.

Informação é quando utilizamos os dados aplicados em um contexto. Por exemplo, os dados do nome e do CPF de um estudante podem fazer parte de uma lista de alunos matriculados em um curso técnico de informática de um Instituto Federal. Os dados organizados agora trazem uma informação associada.

Apesar dessa divisão teórica entre dado e informação, os termos serão usados indiscriminadamente aqui e com o mesmo significado. Também chamaremos no texto dado como variável, ou atributo

2.2 Tipo do dado (Qualitativo ou Quantitativo)

É necessário classificar o dado quanto ao seu tipo pois os algoritmos de aprendizagem de máquina, ou os modelos estatísticos de inferência (termos que serão explicados mais a frente), irão funcionar com determinados tipos de determinadas formas. Logo, com o conhecimento da classificação do tipo do dado poderemos realizar, ou não, as conversões ou tratamentos adicionais que forem necessários.

Utilizaremos dois tipos de dados principais o **qualitativo** e **quantitativo**.

O tipo do dado pode ser: **quantitativo** (também chamado de numérico); **qualitativo** (também chamado de categórico); ou se enquadram na categoria **especial** que engloba outros tipos.

Um dado do tipo **quantitativo** é expresso geralmente como um número. Porém, existem casos em que números inteiros também expressam dados do tipo **qualitativo**, portanto não é só ter número que já podemos classificá-lo como **quantitativo**.

Já o dado do tipo **qualitativo** está relacionado a um valor dentro de um conjunto de itens geralmente finito, porém nem sempre.

Por exemplo, a formação acadêmica de uma determinada pessoa, que pode ser Ensino Fundamental completo, Médio completo ou Superior completo, é um dado do tipo **qualitativo**. Já o salário de um trabalhador é um dado do tipo **quantitativo**.

2.2.1 Quantitativo

O tipo do dado **quantitativo**, também chamado de numérico, ainda pode ser sub classificado como **quantitativo contínuo** ou **quantitativo discreto**.

Um dado **quantitativo contínuo** é quando o dado pode ser qualquer número em um intervalo de números reais - lembrando que o conjunto de números reais engloba os números inteiros -. Geralmente é o resultado de uma medida, por exemplo, a altura de um estudante (por exemplo 1,80 metros) é um dado do tipo **quantitativo contínuo**.

O dado **quantitativo discreto** geralmente é resultado de uma contagem - um número inteiro -, por exemplo, a idade de um estudante (42 anos) é uma contagem, é um dado **quantitativo discreto**.

2.2.2 Qualitativo

Um dado é do tipo **qualitativo** quando representa um valor dentro de um conjunto ou de uma categoria.

O dado **qualitativo** pode ser **qualitativo binário** ou **qualitativo ordinal**, ou nenhuma das duas subcategorias, ou seja **qualitativo somente**.

Um exemplo de **dado qualitativo somente**, é a cor preferida por uma pessoa (por exemplo eu prefiro a cor azul), ou o estado civil de uma pessoa.

O dado do tipo **qualitativo binário** é quando ele somente pode assumir dois valores no universo de valores possíveis. Por exemplo, 0 ou 1, existente ou ausente, true ou false, sim e não, aprovado ou reprovado.

O dado do tipo **qualitativo ordinal** é quando o valor é um elemento de um conjunto que pode ser ordenado, por exemplo, imagine a classificação dos seres humanos entre criança, jovem e adulto. Nesse exemplo, existe uma ordem temporal, o jovem já foi uma criança, o adulto já foi um jovem.

2.2.3 Exemplos

Variável	Tipo do dado
Idade (14, 17, 23)	quantitativo discreto
Doença (Ausente, Presente)	qualitativo binário
Story Points (1, 3, 5, 7 ...)	qualitativo ordinal
Ano (2021, 2022, ...)	quantitativo discreto
Altura (1,79 - 2,05 - ...)	quantitativo contínuo
Estado Civil (Casado, Solteiro)	qualitativo binário
Cores preferidas (Azul, verde, vermelho)	qualitativo somente (nem binário, nem ordinal)

2.3 Tipo do dado (Preditor, Alvo)

Nos modelos de aprendizagem de máquina (quando lidamos com algoritmos classificados como supervisionados) e de inferência estatística o dado também pode ser classificado entre atributo preditor ou atributo alvo. Atributo preditor, são os atributos que serão utilizados para realizar a previsão, geralmente um ou mais atributos. Atributo alvo é o atributo que queremos ‘advinhar (ou dar o melhor chute técnico)’ com os modelos preditivos. Atributo preditor muitas vezes é chamado de variável independente, e atributo alvo de variável dependente.

Col	Tipo do dado (numérico ou categórico)	Tipo do atributo (preditor ou alvo)
IssueKey	categorico somente	-
StoryPoint	numérico discreto	alvo
Created	data	-
Title	texto	preditor
Description	texto	preditor

Ou seja, no modelo proposto, o título e a descrição serão os atributos preditores do atributo alvo, espera-se que os dados do título e da descrição contenham as informações necessárias para a estimativa em Story Points.

2.4 Tabelas

Os dados geralmente são organizados em formato de tabelas. Onde as linhas representam as observações (ou instâncias) e as colunas representam as variáveis.

Vamos utilizar o exemplo de uma empresa que desenvolve software e registra os dados relacionados a seus projetos. Essa empresa mantém o registro de determinada funcionalidade e do tamanho dessa funcionalidade. Cada linha da tabela representa uma funcionalidade (chamada de User Story em projetos que utilizam SCRUM). Cada coluna representa uma informação dessa User Story. As informações que a empresa mantém registro são as variáveis, as colunas da tabela. Uma dessas variáveis é a descrição, outra é uma estimativa que o desenvolvedor atribui do tamanho funcional, chamado Story Point. Essas informações estão dispostas em um arquivo no formato CSV. O código abaixo, carrega esse arquivo e exibe parte de seu conteúdo. Iremos então classificar cada uma das colunas de acordo com o tipo do dado.

Código R

```
df <- read_csv('data/neodataset/7764.csv')

## Rows: 355 Columns: 5
## -- Column specification -----
## Delimiter: ","
## chr  (2): title, description
## dbl  (2): issuekey, storypoints
## dtm  (1): created
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.

head(df)
```

```
## # A tibble: 6 x 5
##   issuekey created          title description storypoints
##   <dbl> <dtm>          <chr>          <chr>          <dbl>
## 1 29688087 2020-01-17 00:50:48 Update templates for web~ "Relates t~          1
## 2 29682716 2020-01-16 19:21:38 Make sure that we Captur~ "This was ~          1
## 3 29644971 2020-01-15 21:17:03 Propose new IA for Brand~ "## Goals\~          1
## 4 29494181 2020-01-10 19:20:50 Cache `node_modules` for~ "# UPDATE ~          1
## 5 29437529 2020-01-09 10:26:51 Disable all remaining un~ "Similar t~          1
## 6 29358963 2020-01-07 08:35:44 Disable unnecessary jobs~ "As discus~          1
```

Código Python

```
import pandas as pd
#pd.set_option('max_columns', None)
df = pd.read_csv('data/neodataset/7764.csv')
df.head()
```

```
##      issuekey  ... storypoints
## 0  29688087  ...             1
## 1  29682716  ...             1
## 2  29644971  ...             1
## 3  29494181  ...             1
## 4  29437529  ...             1
##
## [5 rows x 5 columns]
```

A tabela abaixo não é um exemplo dos dados é a classificação, note que o que era antes coluna virou linha.

Nome da Coluna	Tipo do dado	Observação
Issuekey	categorico somente	Apesar de ser um número, não são realizadas operações no número, ele é um identificador único da User Story
storypoints	numérico discreto	É um número geralmente de 1 á 100
created	data	Data em que a User Story Foi criada
title	texto	Título da User Story
description	texto	Desrição da User Story

A tabela acima apresenta a caracterização dos dados do conjunto de dados neodataset (esse conjunto de dados pode ser baixado em ...). Nessa tabela foram tipificados os dados. É interessante apresentar essa tipificação em estudos científicos e trabalhos de conclusão de curso, quando estamos lidando com conjuntos de dados. Cabe ressaltar que essa tipificação independe da linguagem. Internamente cada linguagem de programação tem seus tipos específicos e que podem ter pequenas diferenças entre as linguagens.

Chapter 3

Estimativas de localização

Muitas vezes é conveniente representar um conjunto de números de uma forma mais simples. Nem sempre temos a possibilidade de lidar com vários números, por limitação ou por falta de praticidade. Por exemplo, imagine uma sala de aula com 5 estudantes, vamos montar uma lista da idade de todos os estudantes nessa sala no R e no Python, duas linguagens de programação comumente utilizadas em análise de dados.

O código a seguir criar uma lista com 5 idades e depois imprime essa mesma lista na linguagem R

```
idades <- c(14,15,16,14,17)
idades
```

```
## [1] 14 15 16 14 17
```

O código a seguir criar uma lista com 5 idades e depois imprime essa mesma lista na linguagem Python

```
idades = [14, 15, 16, 14, 17]
print(idades)
```

```
## [14, 15, 16, 14, 17]
```

Podemos representar essa lista com um número mais simples, que pode resumir ou representar aquela lista original. Para isso, utilizamos as estimativas de localização [Bruce et al., 2020]. As mais comuns são média e mediana.

3.1 Média

A média é calculada dividindo a soma de todos os números da lista pela quantidade de itens. Sua fórmula matemática é apresentada em FIGURA XXX. Onde i é a quantidade de itens da lista e x_i é o i -ésimo item da lista. O termo média também pode ser representado pelo símbolo \bar{X}

No nosso exemplo se fossemos calcular manualmente a média da lista idade, o cálculo seria:

Código R

```
{r} ( 14 + 15 + 16 + 14 + 17 ) / 5
```

Código Python

```
{python} print(( 14 + 15 + 16 + 14 + 17 ) / 5)
```

Porém, podemos utilizar algumas funções que já disponibilizam esse recurso de calcular a média. O código para criar uma lista e verificar a média dessa lista, utilizando as funções, no R e no Python, seria o seguinte:

Código R

```
{r} idades <- c(14, 15, 16, 14, 17) mean(idades)
```

Código Python

```
{python} from statistics import mean idades = [14, 15, 16, 14, 17]  
print(mean(idades))
```

A função `mean`, no R, recebe como parâmetro uma lista de itens e retorna a média dessa lista, no python utilizei a função mesmo nome, porém disponível na biblioteca `statistics` do python.

Chapter 4

Análise exploratória dos dados

4.1 Introdução

Um dos pioneiros na definição da área de análise exploratória de dados (em inglês *Statistical Data Analysis*, ou EDA) foi Tukey (1997) [Tukey et al., 1977]. Tukey (1997) argumenta que seu foco, até aquele momento, estava em desenvolver novas técnicas para inferência. Porém, depois de reflexão, ele chega a conclusão de que o foco dele, e de outros estatísticos, seria melhor aplicado no desenvolvimento de técnicas para a etapa de preparação desses dados. Era nos procedimentos de estruturar os dados que estava o verdadeiro desafio. Problemas, tais como, lidar com dados faltantes ou *outliers*, traziam impactos negativos na inferência e novas técnicas nessa etapa precisavam ser estudadas. Sua recomendação era uma mudança de paradigma e novos estudos, voltando mais para a preparação dos dados. Sua visão é de que isso iria trazer enorme avanços como um todo. O que de fato aconteceu.

Podemos considerar essa necessidade de estudos anterior ao processo de inferência analisando o exemplo criado por Ancobe.

R

```
##  
## Attaching package: 'data.table'  
  
## The following objects are masked from 'package:lubridate':  
##  
##   hour, isoweek, mday, minute, month, quarter, second, wday, week,  
##   yday, year
```

```
## The following objects are masked from 'package:dplyr':
```

```
##
```

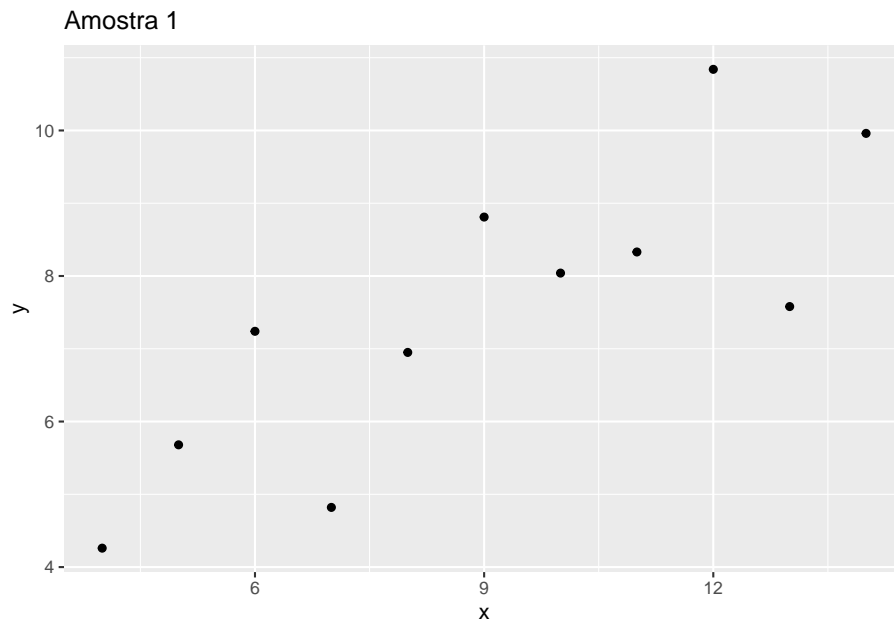
```
##      between, first, last
```

```
## The following object is masked from 'package:purrr':
```

```
##
```

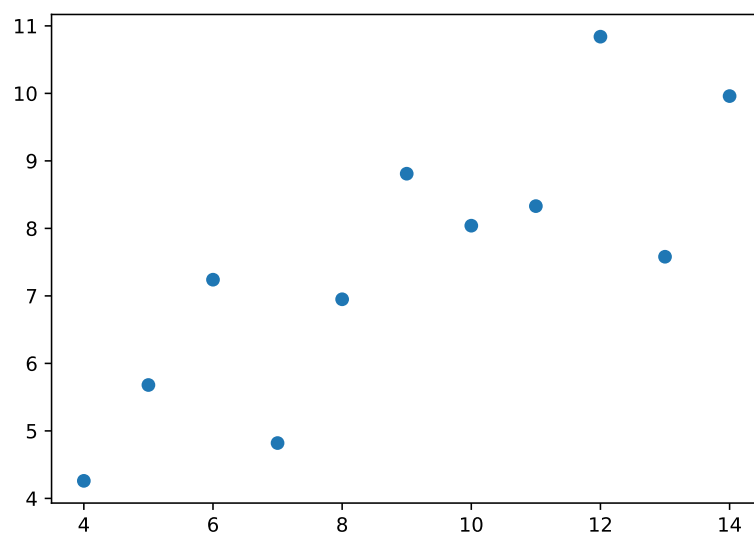
```
##      transpose
```

```
x <- c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5)
y <- c(8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68)
DT = data.table(x, y)
ggplot(DT, mapping = aes(x = x, y = y)) +
  geom_point() +
  labs(title = "Amostra 1")
```



Python

```
import matplotlib.pyplot as plt
x = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
y = [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]
plt.scatter(x, y)
plt.show()
```



Veja a imagem “Amostra 1” acima. Nela visualmente percebemos uma relação linear direta entre as duas variáveis, podemos confirmar isso analisando o gráfico de pontos e o valor da correlação, abaixo.

R

```
cor(x, y)
```

```
## [1] 0.8164205
```

Python

```
from statistics import correlation
print(correlation(x, y))
```

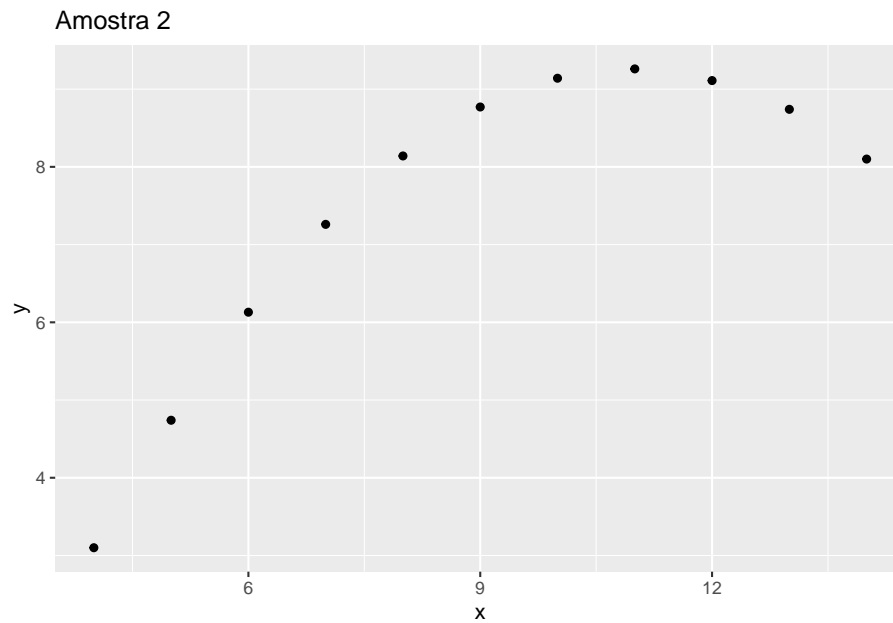
```
## 0.81642051634484
```

Para os dados da Amostra 1, acima temos uma correlação de 0.81.

R

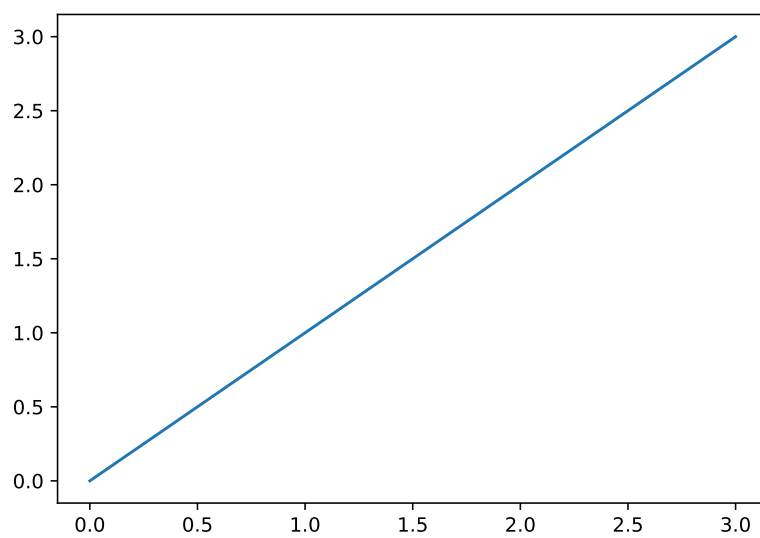
```
x <- c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5)
y <- c(9.14, 8.14, 8.74, 8.77, 9.26, 8.1, 6.13, 3.1, 9.11, 7.26, 4.74)
DT = data.table(x, y)
```

```
ggplot(DT, mapping = aes(x = x, y =y)) +  
  geom_point() +  
  labs(title = "Amostra 2")
```



Python

```
import matplotlib.pyplot as plt  
plt.plot([0, 1, 2, 3])  
plt.show()
```



Na amostra 2 percebemos uma relação em forma de curva, quando verificamos a correlação verificamos o mesmo valor de 0.81 dos dados da amostra 1.

Chapter 5

Python

Com python é possível realizar operações matemáticas, tais como, soma, divisão ou multiplicação, direto no interpretador de código.

```
print(2+2)
```

```
## 4
```


Chapter 6

Referências

Bibliography

Peter Bruce, Andrew Bruce, and Peter Gedeck. *Practical statistics for data scientists: 50+ essential concepts using R and Python*. O'Reilly Media, 2020.

Joel Grus. *Data science do zero*. Alta books, 2016.

Harvard Business Review. "<https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>", 2012.

John W Tukey et al. *Exploratory data analysis*, volume 2. Reading, MA, 1977.