

Data Science – Princípios e Técnicas

Maio
2024



Data Science

Onde me encontrar:

<https://www.linkedin.com/in/marco-mialaret-junior/>

e

<https://github.com/MatmJr>



Vimos nas aulas passadas

Data Science

Estamos trabalhando com 891 observações de 12 variáveis. Para tornar as coisas um pouco mais explícitas, já que alguns dos nomes das variáveis não são totalmente claros, aqui está o que temos que lidar:

Data Science

Montamos um script que acessa o arquivo e faz o processamento dos dados.

Data Science

Nome da Variável	Descrição
Survived	Sobreviveu (1) ou morreu (0)
Pclass	Classe do passageiro (1 = Upper Class, 2 = Middle Class, 3 = Lower Class)
Name	Nome do passageiro
Sex	Sexo do passageiro
Age	Idade do passageiro
SibSp	Número de irmãos/cônjuges a bordo
Parch	Número de pais/filhos a bordo
Ticket	Número do bilhete
Fare	Tarifa
Cabin	Cabine
Embarked	Porto de embarque C = Cherbourg, Q = Queenstown, S = Southampton

Data Science

A **estatística descritiva** é uma ferramenta essencial nesse processo, utilizando duas abordagens principais:

Data Science

- A abordagem quantitativa, que descreve e resume os dados numericamente.
- A abordagem visual, que ilustra os dados por meio de gráficos e visualizações.

Data Science



Vimos na aula passada:

- A tendência central informa sobre os centros dos dados. Medidas úteis incluem a média, mediana e moda.
- A variabilidade informa sobre a dispersão dos dados. Medidas úteis incluem variância e desvio padrão.

Data Science



Veremos na aula de hoje:

- A correlação (ou variabilidade conjunta) informa sobre a relação entre um par de variáveis em um conjunto de dados.

A Correlação

Data Science

A análise de correlação mede o grau de dependência entre duas ou mais variáveis, ou seja, como uma variável influencia outra. Esta relação pode ser não-causal e é quantificada por coeficientes, como o coeficiente de Pearson, que será o foco desta aula.

Data Science

O coeficiente de Pearson, também chamado de “coeficiente de correlação produto-momento” ou chamado de “ ρ de Pearson”, mede o grau de correlação através do cálculo de direção positiva ou negativa. Este coeficiente, normalmente representado por ρ assume apenas valores entre -1 e 1.

Data Science

A análise de correlação vai retornar três possíveis cenários:

- correlação positiva;
- correlação negativa;
- não há correlação.

Data Science

- Correlação positiva: quando duas variáveis que possuem correlação crescem ou decrescem juntas, ou seja, que possuem uma relação direta;

Data Science

- Correlação negativa: quando duas variáveis que possuem correlação mas quando uma variável cresce a outra decresce, ou vice-versa;

Data Science

- Não ter correlação: quando o crescimento ou decrescimento de uma variável não tem efeito sobre outra variável.

Data Science

$\rho = 0,9$ a 1 (positivo ou negativo): correlação muito forte;
 $\rho = 0,7$ a $0,9$ (positivo ou negativo): correlação forte;
 $\rho = 0,5$ a $0,7$ (positivo ou negativo): correlação moderada;
 $\rho = 0,3$ a $0,5$ (positivo ou negativo): correlação fraca;
 $\rho = 0$ a $0,3$ (positivo ou negativo): não possui correlação.

A Correlação

Data Science



O primeiro passo para usar esse conceito no python é importar as bibliotecas necessárias: ``pandas`` e ``seaborn``. ``pandas`` é essencial para manipular dados, tabelas e dataframes, enquanto ``seaborn`` é ideal para criar visualizações gráficas, especialmente para mapeamentos estatísticos.

Data Science

```
import pandas as pd  
import seaborn as sns
```

```
nums = ["Survived", "Pclass", "Age", "SibSp", "Parch", "Fare"]
```

```
data[nums].corr()
```

Data Science

	Survived	Pclass	Age	SibSp	Parch	Fare
Survived	1.000000	-0.338481	-0.059579	-0.035322	0.081629	0.257307
Pclass	-0.338481	1.000000	-0.413583	0.083081	0.018443	-0.549500
Age	-0.059579	-0.413583	1.000000	-0.249854	-0.175526	0.122692
SibSp	-0.035322	0.083081	-0.249854	1.000000	0.414838	0.159651
Parch	0.081629	0.018443	-0.175526	0.414838	1.000000	0.216225
Fare	0.257307	-0.549500	0.122692	0.159651	0.216225	1.000000

Data Science

Para transformar as colunas categóricas 'Sex', 'Cabin', e 'Title' em números usando LabelEncoder do pacote sklearn, você pode seguir os passos a seguir:

Data Science

```
from sklearn.preprocessing import LabelEncoder
```

```
# Inicializando o LabelEncoder  
label_encoder = LabelEncoder()
```

```
# Colunas a serem transformadas  
columns_to_encode = ['Sex', 'Cabin', 'Title']
```


Data Science

```
# Aplicando LabelEncoder às colunas categóricas
for column in columns_to_encode:
    data[column] = label_encoder.fit_transform(data[column])

# Mostrando o DataFrame transformado
data.head()
```

Data Science



Para visualizarmos a matriz de correlação, vamos utilizar a função `.heatmap()` do pacote `seaborn`, essa função vai nos retornar uma forma gráfica da matriz com uma escala de cor em conjunto com uma escala numérica, as quais vão indicar o grau medido entre as variáveis.

Data Science

```
correlation = data[vars].corr()
```

```
# plot da matriz de correlação
```

```
sns.heatmap(correlation, annot = True, fmt=".1f", linewidths=.6)
```

Dúvidas?



81 98160 7018

marco.junior@pe.senac.br

