
Data Science do Zero

Joel Grus

Sumário

Prefácioxi

1. Introdução1

 A Ascensão dos Dados 1

 O Que É Data Science? 1

 Motivação Hipotética: DataSciencester 2

 Encontrando Conectores-Chave 3

 Cientistas de Dados Que Você Talvez Conheça 6

 Salários e Experiência 8

 Contas Pagas 11

 Tópicos de Interesse 11

 Em Diante 13

2. Curso Relâmpago de Python15

 O Básico 15

 Iniciando em Python 15

 Python Zen 16

 Formatação de Espaço em Branco 16

 Módulos 17

 Aritmética 18

 Funções 18

 Strings (cadeias de caracteres) 19

 Exceções 19

 Listas 20

 Tuplas 21

 Dicionários 21

 Conjuntos 24

 Controle de Fluxo 25

Veracidade	25
Não Tão Básico	26
Ordenação	27
Compreensões de Lista	27
Geradores e Iteradores	28
Aleatoriedade	29
Expressões Regulares	30
Programação Orientada a Objeto	30
Ferramentas Funcionais	31
Enumeração (enumerate)	32
Descompactação de Zip e Argumentos	33
args e kwargs	34
Bem-vindo à DataSciencester!	35
Para Mais Esclarecimentos	35
3. Visualizando Dados	37
matplotlib	37
Gráficos de Barra	39
Gráficos de Linhas	43
Gráficos de Dispersão	44
Para Mais Esclarecimentos	47
4. Álgebra Linear	49
Vetores	49
Matrizes	53
Para Mais Esclarecimentos	55
5. Estatística	57
Descrevendo um Conjunto Único de Dados	57
Tendências Centrais	59
Dispersão	61
Correlação	62
Paradoxo de Simpson	65
Alguns Outros Pontos de Atenção sobre Correlação	66
Correlação e Causalidade	67
Para Mais Esclarecimentos	68
6. Probabilidade	69
Dependência e Independência	69
Probabilidade Condicional	70
Teorema de Bayes	72
Variáveis Aleatórias	73

Distribuições Contínuas	74
A Distribuição Normal	75
O Teorema do Limite Central	78
Para Mais Esclarecimentos	80
7. Hipótese e Inferência.....	81
Teste Estatístico de Hipótese	81
Exemplo: Lançar Uma Moeda	81
p -values	84
Intervalos de Confiança	85
P-Hacking	86
Exemplo: Executando um Teste A/B	87
Inferência Bayesiana	88
Para Mais Esclarecimentos	92
8. Gradiente Descendente	93
A Ideia Por Trás do Gradiente Descendente	93
Estimando o Gradiente	94
Usando o Gradiente	97
Escolhendo o Tamanho do Próximo Passo	97
Juntando Tudo	98
Gradiente Descendente Estocástico	99
Para Mais Esclarecimentos	100
9. Obtendo Dados.....	103
stdin e stdout	103
Lendo Arquivos	105
O Básico de Arquivos Texto	105
Arquivos delimitados	106
Extraindo Dados da Internet	108
HTML e Sua Subsequente Pesquisa	108
Exemplo: Livros O'Reilly Sobre Dados	110
Usando APIs	114
JSON (e XML)	114
Usando Uma API Não Autenticada	115
Encontrando APIs	116
Exemplo: Usando as APIs do Twitter	117
Obtendo Credenciais	117
Para Mais Esclarecimentos	120
10. Trabalhando com Dados	121
Explorando Seus Dados	121
Explorando Dados Unidimensionais	121

Duas Dimensões	123
Muitas Dimensões	125
Limpando e Transformando	127
Manipulando Dados	129
Redimensionando	132
Redução da Dimensionalidade	134
Para Mais Esclarecimentos	139
11. Aprendizado de Máquina	141
Modelagem	141
O Que É Aprendizado de Máquina?	142
Sobreajuste e Sub-Ajuste	142
Precisão	145
Compromisso entre Polarização e Variância	147
Recursos Extração e Seleção de Característica	148
Para Mais Esclarecimentos	150
12. K–Vizinhos Mais Próximos	151
O Modelo	151
Exemplo: Linguagens Favoritas	153
A Maldição da Dimensionalidade	156
Para Mais Esclarecimentos	163
13. Naive Bayes	165
Um Filtro de Spam Muito Estúpido	165
Um Filtro de Spam Mais Sofisticado	166
Implementação	168
Testando Nosso Modelo	169
Para Mais Esclarecimentos	172
14. Regressão Linear Simples	173
O Modelo	173
Usando o Gradiente Descendente	176
Estimativa Máxima da Probabilidade	177
Para Mais Esclarecimentos	177
15. Regressão Múltipla	179
O Modelo	179
Mais Suposições do Modelo dos Mínimos Quadrados	180
Ajustando o Modelo	181
Interpretando o Modelo	182
O Benefício do Ajuste	183

Digressão: A Inicialização	183
Erros Padrões de Coeficientes de Regressão	184
Regularização	186
Para Mais Esclarecimentos	188
16. Regressão Logística	189
O Problema	189
A Função Logística	192
Aplicando o Modelo	194
O Benefício do Ajuste	195
Máquina de Vetor de Suporte	196
Para Mais Esclarecimentos	200
17. Árvores de Decisão	201
O Que É uma Árvore de Decisão?	201
Entropia	203
A Entropia de uma Partição	205
Criando uma Árvore de Decisão	206
Juntando Tudo	208
Florestas Aleatórias	211
Para Maiores Esclarecimentos	212
18. Redes Neurais	213
Perceptrons	213
Redes Neurais Feed-Forward	215
Backpropagation	218
Exemplo: Derrotando um CAPTCHA	219
Para Mais Esclarecimentos	224
19. Agrupamento.....	225
A Ideia	225
O Modelo	226
Exemplo: Encontros	227
Escolhendo k	230
Exemplo: Agrupando Cores	231
Agrupamento Hierárquico Bottom-up	233
Para Mais Esclarecimentos	238
20. Processamento de Linguagem Natural	239
Nuvens de Palavras	239
Modelos n-gramas	241
Gramáticas	244

Um Adendo: Amostragem de Gibbs	246
Modelagem de Tópicos	247
Para Mais Esclarecimentos	253
21. Análise de Rede	255
Centralidade de Intermediação	255
Centralidade de Vetor Próprio	260
Multiplicação de Matrizes	260
Centralidade	262
Gráficos Direcionados e PageRank	264
Para Mais Esclarecimentos	266
22. Sistemas Recomendadores	267
Curadoria Manual	268
Recomendando O Que é Popular	268
Filtragem Colaborativa Baseada no Usuário	269
Filtragem Colaborativa Baseada em Itens	272
Para Mais Esclarecimentos	274
23. Bases de Dados e SQL.....	275
CREATE TABLE e INSERT	275
UPDATE	277
DELETE	278
SELECT	278
GROUP BY	280
ORDER BY	282
JOIN	283
Subconsultas	285
Índices	285
Otimização de Consulta	286
NoSQL	287
Para Mais Esclarecimentos	287
24. MapReduce	289
Exemplo: Contagem de Palavras	289
Por que MapReduce?	291
MapReduce Mais Generalizado	292
Exemplo: Analisando Atualizações de Status	293
Exemplo: Multiplicação de Matriz	294
Um Adendo: Combinadores	296
Para Mais Esclarecimentos	296

25. **Vá em Frente e Pratique Data Science299**

- IPython 299
- Matemática 300
- Não Do Zero 300
 - NumPy 301
 - pandas 301
 - scikit-learn 301
 - Visualização 301
 - R 302
- Encontre Dados 302
- Pratique Data Science 303
 - Hacker News 303
 - Carros de Bombeiros 303
 - Camisetas 304
 - E Você? 304

26. **Índice.....305**