

Elogios ao *Guia do Python para Data Science*, tradução da segunda edição

Existem muitos livros sobre data science disponíveis, mas a obra de Jake VanderPlas é excepcional. Ele trata de um assunto muito amplo e complexo dividindo-o de forma didática, com boa escrita e exercícios que fazem você aplicar os conceitos rapidamente.

— Celeste Stinger, Site Reliability Engineer (SRE).

A experiência e a paixão que Jake VanderPlas mostra ao repassar seu conhecimento são inegáveis. Esta edição recém-atualizada dá exemplos claros e simples que o ajudarão a se preparar e usar ferramentas essenciais para data science e aprendizado de máquina.

Se estiver pronto para se aprofundar nas principais técnicas de ferramentas Python para ter insights a partir dos seus dados, então este livro é para você.

— Anne Bonner, fundadora e CEO, Content Simplicity

O *Guia do Python para Data Science* é uma das minhas principais recomendações para estudantes da área há anos. A segunda edição consegue melhorar um livro que já é muito bom e completo, com notebooks Jupyter interessantes e que permitem executar suas receitas durante a leitura.

— Noah Gift, Executive in Residence (EIR) da Universidade Duke e fundador do Pragmatic AI Labs

Esta versão atualizada é uma ótima introdução às bibliotecas que fazem do Python uma excelente linguagem para data science e programação científica, sendo apresentadas de maneira acessível e com ótimos exemplos ao longo de todo o livro.

— Allen Downey, autor de *Pense em Python* e *Think Bayes*

O Guia do Python para Data Science é ótimo para leitores em processo de aprendizado em data science. Os exemplos práticos são completos e escritos de forma acessível, garantindo o aprendizado do leitor sobre como armazenar, manipular e obter insights a partir de um conjunto de dados.

— William Jamir Silva, engenheiro de software sênior, Adjust GmbH

Jake VanderPlas tem um histórico de desmembrar os principais conceitos e ferramentas de Python para os alunos de data science, e na segunda edição do *Guia do Python para Data Science* ele cumpriu seu objetivo mais uma vez. Neste livro, ele fornece uma visão geral de todas as ferramentas necessárias para começar, além de algumas explicações do funcionamento interno das coisas, fazendo tudo isso de forma muito acessível.

— Jackie Kazil, criadora do driver Mesa Library e líder em data science

TRADUÇÃO DA SEGUNDA EDIÇÃO

Guia do Python para Data Science

*Ferramentas Essenciais
para Trabalhar com Dados*

Jake VanderPlas



ALTA BOOKS
GRUPO EDITORIAL

Rio de Janeiro, 2025

Guia do Python para Data Science - Tradução da 2ª Edição

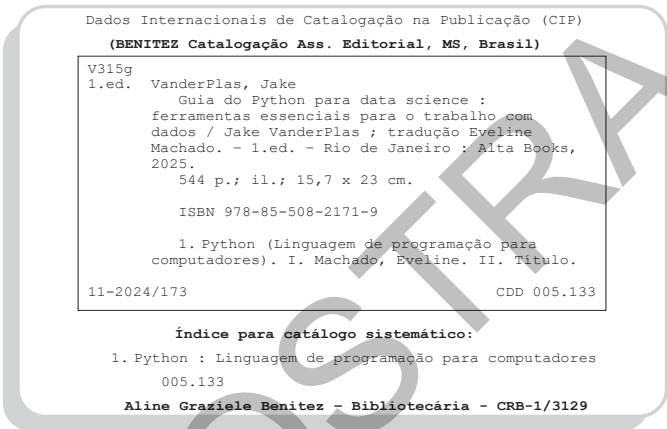
Copyright © 2025 Starlin Alta Editora e Consultoria Eireli.

Copyright © 2022 Jake VanderPlas.

ISBN: 978-85-508-2171-9

Authorized Portuguese translation of the English edition of Python Data Science Handbook, 2nd Edition ISBN 9781098121228 © 2022 Jake VanderPlas. This translation is published and sold by permission of O'Reilly Media, Inc., which owns or controls all rights to publish and sell the same. PORTUGUESE language edition published by Grupo Editorial Alta Books Ltda., Copyright © 2025 by STARLIN ALTA EDITORA E CONSULTORIA LTDA.

Impresso no Brasil – 1ª Edição, 2025 – Edição revisada conforme o Acordo Ortográfico da Língua Portuguesa de 2009.



Todos os direitos estão reservados e protegidos por Lei. Nenhuma parte deste livro, sem autorização prévia por escrito da editora, poderá ser reproduzida ou transmitida. A violação dos Direitos Autorais é crime estabelecido na Lei nº 9.610/98 e com punição de acordo com o artigo 184 do Código Penal.

O conteúdo desta obra fora formulado exclusivamente pelo(s) autor(es).

Marcas Registradas: Todos os termos mencionados e reconhecidos como Marca Registrada e/ou Comercial são de responsabilidade de seus proprietários. A editora informa não estar associada a nenhum produto e/ou fornecedor apresentado no livro.

Material de apoio e erratas: Se parte integrante da obra e/ou por real necessidade, no site da editora o leitor encontrará os materiais de apoio (download), errata e/ou quaisquer outros conteúdos aplicáveis à obra. Acesse o site www.altabooks.com.br e procure pelo título do livro desejado para ter acesso ao conteúdo.

Suporte Técnico: A obra é comercializada na forma em que está, sem direito a suporte técnico ou orientação pessoal/exclusiva ao leitor.

A editora não se responsabiliza pela manutenção, atualização e idioma dos sites, programas, materiais complementares ou similares referidos pelos autores nesta obra.

Produção Editorial: Grupo Editorial Alta Books
Diretor Editorial: Anderson Vieira
Vendas Governamentais: Cristiane Mutús
Gerência Comercial: Claudio Lima
Gerência Marketing: Andréa Guatiello

Produtora Editorial: Isabella Gibara
Tradução: Eveline Machado
Copidesque: Aline Amaral
Revisão: Denise Himpel
Diagramação: Daniel Vargas e Joyce Matos
Revisão Técnica: Jhonatan Pereira
(Desenvolvedor de Software)



Rua Viúva Cláudio, 291 – Bairro Industrial do Jacaré
CEP: 20.970-031 – Rio de Janeiro (RJ)
Tels.: (21) 3278-8069 / 3278-8419
www.altabooks.com.br – altabooks@altabooks.com.br
Ouvidoria: ouvidoria@altabooks.com.br



Sumário

Prefácio.....	xv
O Que É Data Science?	xv
Para Quem É este Livro?	xvi
Por que Python?	xvii
Estrutura do Livro	xvii
Considerações Sobre a Instalação	xviii
Convenções Usadas neste Livro	xix
Usando Exemplos de Código	xix
PARTE I Jupyter: Além do Python Comum	
CAPÍTULO 1: Começando com IPython e Jupyter.....	3
Executando o Shell IPython	3
Executando o Jupyter Notebook	3
Ajuda e Documentação no IPython	4
Acessando a Documentação com ?	5
Acessando o Código-fonte com ??	6
Explorando Módulos com o Autocompletar	7
Atalhos de Teclado no Shell IPython	9
Atalhos de Navegação	10
Atalhos da Entrada de Texto	10
Atalhos do Histórico de Comandos	10
Atalhos Diversos	12
CAPÍTULO 2: Recursos Interativos Aprimorados.....	13
Comandos Mágicos do IPython	13
Executando o Código Externo: %run	13
Cronometrando a Execução do Código: %timeit	14
Ajuda nas Funções Mágicas: ?, %magic e %lsmagic	15
Histórico de Entradas e Saídas	15
Objetos In e Out do IPython	15
Atalhos com Sublinhado e Saídas Anteriores	16
Suprimindo a Saída	17
Comandos Mágicos Relacionados	17
Comandos IPython e Shell	18
Breve Introdução ao Shell	18
Comandos Shell no IPython	19
Passando Valores no Shell	20
Comandos Mágicos Relacionados ao Shell	20
CAPÍTULO 3: Depuração e Perfil	23
Erros e Depuração	23
Controlando as Exceções: %xmode	23
Depuração: Quando a Leitura dos Rastreamentos Não É Suficiente	25
Perfilamento e Cronometragem de Código	27
Cronometrando Trechos de Código: %timeit e %time	28

Perfilando Scripts Completos: %prun	29
Perfilamento Linha a Linha com %lprun	30
Perfilando o Uso de Memória: %memit e %mprun	31
Mais Recursos para o IPython	33
Referências na Web [conteúdos em inglês]	33
Livros [alguns sem publicação no Brasil]	33

PARTE II Introdução ao NumPy

CAPÍTULO 4: Entendendo os Tipos de Dados em Python	37
Um Inteiro em Python É Mais do Que Um Inteiro	38
Uma Lista em Python É Mais Do Que uma Lista	39
Arrays do Tipo Fixo em Python	41
Criando Arrays a partir de Listas Python	41
Criando Arrays do Zero	42
Tipos Padrão de Dados do NumPy	43
CAPÍTULO 5: O Básico sobre Arrays NumPy	45
Atributos de um Array NumPy	45
Indexação de Arrays: Acessando Elementos Individualmente	46
Fatiamento de Arrays: Acessando os Subarrays	47
Subarrays Unidimensionais	47
Subarrays Multidimensionais	48
Subarrays como Views	49
Criando Cópias de Arrays	49
Remodelação de Arrays	50
Concatenação e Separação de Arrays	51
Concatenação de Arrays	51
Separação de Arrays	52
CAPÍTULO 6: Computação com Arrays NumPy: Funções Universais	53
A Lentidão dos Loops	53
Apresentando as Ufuncs	54
Explorando as Ufuncs do NumPy	55
Aritmética dos Arrays	55
Valor Absoluto	56
Funções Trigonométricas	57
Expoentes e Logaritmos	57
Ufuncs Especializadas	58
Recursos Avançados das Ufuncs	59
Especificando a Saída	59
Agregações	59
Produtos Diádicos	60
Ufuncs: Aprendendo Ainda Mais	60
CAPÍTULO 7: Agregações: min, max e Tudo Mais	61
Somando os Valores em um Array	61
Mínimo e Máximo	62
Agregações Multidimensionais	62
Outras Funções Agregadas	63
Exemplo: Qual a Altura Média dos Presidentes dos Estados Unidos?	64
CAPÍTULO 8: Cálculo em Arrays: Broadcasting	67
Apresentando o Broadcasting	67
Regras do Broadcasting	69
Exemplo de Broadcasting 1	69
Exemplo de Broadcasting 2	70

Exemplo de Broadcasting 3	70
Broadcasting na Prática	71
Centralizando um Array	71
Plotando uma Função Bidimensional	72
CAPÍTULO 9: Comparações, Máscaras e Lógica Booleana.....	73
Exemplo: Contando Dias Chuvosos	73
Operadores de Comparação como Ufuncs	74
Trabalhando com Arrays Booleanos	75
Contando Entradas	76
Operadores Booleanos	77
Arrays Booleanos como Máscaras	78
Usando as Palavras-chave and/or Versus os Operadores &/	79
CAPÍTULO 10: Fancy Indexing	81
Explorando a Fancy Indexing/Indexação Avançada	81
Indexação Combinada	82
Exemplo: Selecionando Pontos Aleatórios	83
Modificando Valores com Fancy Indexing/Indexação Avançada	85
Exemplo: Agrupando Dados	86
CAPÍTULO 11: Ordenação de Arrays.....	89
Ordenação rápida no NumPy: np.sort e np.argsort	89
Ordenando em Linhas ou Colunas	90
Ordenações Parciais: Particionando	91
Exemplo: K-ésimos Vizinhos mais Próximos	91
CAPÍTULO 12: Dados Estruturados: Arrays Estruturados do NumPy	95
Explorando a Criação dos Arrays Estruturados	96
Tipos Compostos Mais Avançados	97
Arrays de Registro: Arrays Estruturados com um Ajuste	98
Ao Pandas	98
PARTE III Manipulação de Dados com Pandas	
CAPÍTULO 13: Introdução aos Objetos Pandas.....	101
Objeto Series do Pandas	101
Series como um Array NumPy Generalizado	102
Series como um Dicionário Especializado	103
Criando Objetos Series	103
Objeto DataFrame do Pandas	104
DataFrame como um Array NumPy Generalizado	104
DataFrame como um Dicionário Especializado	105
Criando Objetos DataFrame	106
Objeto Index do Pandas	107
Index como Array Imutável	108
Index como Conjunto Ordenado	108
CAPÍTULO 14: Indexação e Seleção de Dados	109
Seleção de Dados em Series	109
Series como Dicionário	109
Series como Array Unidimensional	110
Indexadores: loc e iloc	111
Seleção de Dados nos DataFrames	112
DataFrame como Dicionário	112
DataFrame como Array Bidimensional	114
Outras Convenções de Indexação	115

CAPÍTULO 15: Operando com Dados no Pandas	117
Ufuncs: Preservação do Índice	117
Ufuncs: Alinhamento do Índice	118
Alinhamento do Índice em Series	118
Alinhamento do Índice em DataFrames	119
Ufuncs: Operações Entre DataFrames e Series	120
CAPÍTULO 16: Lidando com Dados Ausentes	123
Compensações nas Convenções de Dados Ausentes	123
Dados Ausentes no Pandas	124
None como um Valor Sentinela	124
NaN: Dados Numéricos Ausentes	125
NaN e None no Pandas	126
Dtypes Null do Pandas	127
Operando nos Valores Null	127
Detectando Valores Null	128
Descartando os Valores Null	128
Preenchendo Valores Null	129
CAPÍTULO 17: Indexação Hierárquica	131
Objeto Series Indexado e Múltiplo	131
Modo Errado	131
Modo Certo: MultiIndex do Pandas	132
MultiIndex Como uma Dimensão Extra	133
Métodos de Criação de MultiIndex	134
Construtores MultiIndex Explícitos	135
Nomes do Nível de MultiIndex	136
MultiIndex para Colunas	136
Indexando e Fatiando um MultiIndex	137
Objeto Series Indexado e Múltiplo	137
DataFrames Indexados e Múltiplos	139
Reorganizando os Índices Múltiplos	140
Índices Ordenados e Não Ordenados	140
Montando e Desmontando os Índices	141
Definindo e Redefinindo o Índice	142
CAPÍTULO 18: Combinando Conjuntos de Dados: concat e append	143
Lembrete: Concatenação dos Arrays NumPy	144
Concatenação Simples com pd.concat	144
Índices Duplicados	145
Concatenação com Joins	147
Método append	148
CAPÍTULO 19: Combinando Conjuntos de Dados: merge e join	149
Álgebra Relacional	149
Categorias de Junções	150
Junções Um para Um	150
Junções Muitos para Um	150
Junções Muitos para Muitos	151
Especificação da Chave de Mesclagem	152
Palavra-chave on	152
Palavras-chave left_on e right_on	152
Palavras-chave left_index e right_index	153
Especificando o Conjunto Aritmético para as Junções	154
Sobrepondo os Nomes da Coluna: Palavra-chave suffixes	156
Exemplo: Dados dos Estados nos EUA	157

CAPÍTULO 20: Agregação e Agrupamento.....	161
Dados dos Planetas	161
Agregação Simples no Pandas	162
groupby: Dividir, Aplicar, Combinar	164
Dividir, Aplicar, Combinar	164
Objeto GroupBy	166
Agregar, Filtrar, Transformar, Aplicar	167
Especificando a Chave da Divisão	170
Exemplo de Agrupamento	171
CAPÍTULO 21: Tabelas Dinâmicas.....	173
Motivando as Tabelas Dinâmicas	173
Tabelas Dinâmicas Feitas à Mão	173
Sintaxe da Tabela Dinâmica	174
Tabelas Dinâmicas com Vários Níveis	174
Opções Adicionais da Tabela Dinâmica	175
Exemplo: Dados de Natalidade	176
CAPÍTULO 22: Operações de String Vetorizadas	181
Introdução das Operações de String do Pandas	181
Tabelas de Métodos de String do Pandas	182
Métodos Similares aos Métodos de String do Python	182
Métodos Usando Expressões Regulares	183
Métodos Diversos	184
Exemplo: Banco de Dados de Receitas	186
Recomendação Simples de Receitas	188
Indo Além com Receitas	189
CAPÍTULO 23: Trabalhando com Séries Temporais	191
Datas e Horas no Python	191
Datas e Horas Nativas do Python: datetime e dateutil	192
Arrays Tipados de Horas: datetime64 do NumPy	192
Datas e Horas no Pandas: O Melhor dos Dois Mundos	194
Série Temporal do Pandas: Indexando por Tempo	195
Estruturas de Dados da Série Temporal do Pandas	196
Sequências Regulares: pd.date_range	196
Frequências e Diferenças de Data	198
Nova Amostragem, Deslocamento e Janelas	199
Amostragem e Conversão das Frequências	200
Mudanças do Tempo	202
Janelas Contínuas	203
Exemplo: Visualizando a Contagem de Bicicletas em Seattle	204
Visualizando os Dados	206
Aprofundando os Dados	208
CAPÍTULO 24: Pandas de Alta Performance: eval e query	211
Motivando query e eval: Expressões Compostas	211
pandas.eval para Operações Eficientes	212
DataFrame.eval para Operações por Coluna	214
Atribuição em DataFrame.eval	215
Variáveis Locais em DataFrame.eval	215
Método DataFrame.query	216
Performance: Quando Usar Estas Funções	216
Mais Recursos	217

PARTE IV Visualização com Matplotlib

CAPÍTULO 25: Dicas Gerais do Matplotlib	221
Importando o Matplotlib	221
Definindo Estilos	221
Usando show ou Não? Como Exibir Suas Plotagens	221
Plotando a partir de um Script	221
Plotando a partir de um Shell IPython	222
Plotando a partir de um Jupyter Notebook	222
Salvando Imagens no Arquivo	223
Duas Interfaces pelo Preço de Uma	225
CAPÍTULO 26: Gráficos com Linhas Simples.....	227
Ajustando a Plotagem: Cores e Estilos de Linha	229
Ajudando a Plotagem: Limites dos Eixos	231
Rotulando as Plotagens	234
Pegadinhas do Matplotlib	235
CAPÍTULO 27: Gráficos de Dispersão Simples	237
Gráficos de Dispersão com plt.plot	237
Gráficos de Dispersão com plt.scatter	240
plot Versus scatter: Uma Observação sobre Eficiência	242
Visualizando as Incertezas	243
Barras de Erros Básicas	243
Erros Contínuos	245
CAPÍTULO 28: Gráficos de Densidade e Contorno	247
Visualizando uma Função Tridimensional	247
Histogramas, Agrupamentos e Densidade	251
Histogramas e Agrupamentos Bidimensionais	253
plt.hist2d: Histograma Bidimensional	254
plt.hexbin: Agrupamentos Hexagonais	255
Estimativa de Densidade por Kernel	255
CAPÍTULO 29: Customizando as Legendas do Gráfico.....	257
Escolhendo os Elementos da Legenda	259
Legenda para o Tamanho dos Pontos	260
Múltiplas Legendas	261
CAPÍTULO 30: Customizando as Barras de Cores	263
Customizando as Barras de Cores	264
Escolhendo o Mapa de Cores	265
Limites e Extensões das Cores	267
Barras de Cores Separadas	268
Exemplo: Dígitos Feitos à Mão	268
CAPÍTULO 31: Múltiplos Gráficos Secundários.....	271
plt.axes: Gráficos Secundários Feitos à Mão	271
plt.subplot: Grades Simples dos Gráficos Secundários	273
plt.subplots: A Grade Inteira de Uma Só Vez	274
plt.GridSpec: Organizações Mais Complexas	275
CAPÍTULO 32: Texto e Anotação	279
Exemplo: Efeitos dos Feriados nos Nascimentos nos EUA	279
Transformações e Posição do Texto	281
Setas e Anotação	283
CAPÍTULO 33: Customizando os Ticks	287
Marcas Maiores e Menores	287
Ocultando Marcas ou Rótulos	288

Reduzindo ou Aumentando o Número de Marcas	290
Formatos Elegantes das Marcas	291
Resumo de Formataadores e Localizadores	293
CAPÍTULO 34: Customizando o Matplotlib: Configurações e Folhas de Estilo	295
Customização do Gráfico à Mão	295
Mudando os Padrões: rcParams	297
Folhas de Estilo	298
Estilo Default	299
Estilo FiveThirtyEight	299
Estilo ggplot	300
Métodos Bayesianos para o Estilo Hackers	300
Estilo com Fundo Escuro	301
Estilo em Tons de Cinza	301
Estilo Seaborn	302
CAPÍTULO 35: Gráfico Tridimensional no Matplotlib	303
Pontos e Linhas Tridimensionais	304
Gráficos de Contorno Tridimensional	304
Gráficos Wireframes e de Superfície	306
Triangulações da Superfície	308
Exemplo: Visualizando uma fita de Möbius	310
CAPÍTULO 36: Visualização com Seaborn	313
Explorando os Gráficos Seaborn	314
Histogramas, KDE e Densidades	314
Gráficos em Pares	316
Histogramas Facetados	317
Gráficos com Categorias	318
Distribuições Conjuntas	318
Gráficos de Barras	320
Exemplo: Explorando os Tempos de Término da Maratona	321
Mais Recursos	329
Outras Bibliotecas de Visualização do Python	330
PARTE V Aprendizado de Máquina	
CAPÍTULO 37: O que É Aprendizado de Máquina?	333
Categorias de Aprendizado de Máquina	333
Exemplos Qualitativos das Aplicações do Aprendizado de Máquina	334
Classificação: Prevendo Rótulos Separados	334
Regressão: Prevendo Rótulos Contínuos	336
Clusterização: Inferindo Rótulos em Dados Não Rotulados	339
Redução da Dimensionalidade: Inferindo a Estrutura dos Dados Sem Rótulo	340
Resumo	342
CAPÍTULO 38: Apresentando o Scikit-Learn	343
Representação dos Dados no Scikit-Learn	343
Matriz de Features	344
Array de Destino	344
API Estimator	346
Noções Básicas da API	347
Exemplo de Aprendizado Supervisionado: Regressão Linear Simples	347
Exemplo de Aprendizado Supervisionado: Classificação Iris	350
Exemplo de Aprendizado Não Supervisionado: Dimensionalidade de Iris	351
Exemplo de Aprendizado Não Supervisionado: Clusterização Iris	352
Aplicação: Explorando Dígitos Feitos à Mão	353

Carregando e Visualizando os Dados dos Dígitos	354
Exemplo de Aprendizado Não Supervisionado: Redução da Dimensionalidade	355
Classificação nos Dígitos	356
Resumo	358
CAPÍTULO 39: Hiperparâmetros e Validação do Modelo	359
Considerando a Validação do Modelo	359
Validação do Modelo do Jeito Errado	359
Validação do Modelo do Jeito Certo: Conjuntos de Validação	360
Validação do Modelo com Validação Cruzada	361
Escolhendo o Melhor Modelo	363
Equilíbrio entre Viés e Variância	363
Curvas de Validação no Scikit-Learn	366
Curvas de Aprendizado	369
Validação na Prática: Pesquisa em Grade	373
Resumo	374
CAPÍTULO 40: Feature Engineering	375
Features Categóricas	375
Features de Texto	376
Features da Imagem	378
Features Derivadas	378
Imputação de Dados Ausentes	381
Feature Pipelines	382
CAPÍTULO 41: Aprofundando: Classificação Naive Bayes	383
Classificação Bayesiana	383
Gaussian Naive Bayes	384
Multinomial Naive Bayes	386
Exemplo: Classificando Texto	387
Quando Usar Naive Bayes	390
CAPÍTULO 42: Aprofundando: Regressão Linear	391
Regressão Linear Simples	391
Regressão de Função Base	393
Funções de Base Polinomial	394
Funções de Base Gaussianas	395
Regularização	397
Regressão Ridge (Regularização L_2)	398
Regressão Lasso (Regularização L_1)	399
Exemplo: Prevendo o Tráfego de Bicicletas	400
CAPÍTULO 43: Aprofundando: Máquinas de Vetores de Suporte	407
Motivando as Máquinas de Vetores de Suporte	407
Máquinas de Vetores de Suporte: Maximizando a Margem	409
Ajustando uma Máquina de Vetores de Suporte	410
Além dos Limites Lineares: SVM do Kernel	413
Ajustando o SVM: Suavizando as Margens	416
Exemplo: Reconhecimento Facial	417
Resumo	422
CAPÍTULO 44: Aprofundando: Árvores de Decisão e Florestas Aleatórias	423
Motivando as Florestas Aleatórias: Árvores de Decisão	423
Criando uma Árvore de Decisão	424
Árvores de Decisão e Sobreajuste	426
Conjuntos de Estimadores: Florestas Aleatórias	427
Regressão da Floresta Aleatória	429
Exemplo: Floresta Aleatória para Classificar Dígitos	430
Resumo	432

CAPÍTULO 45: Aprofundando: Análise de Componentes Principais	433
Introduzindo a Análise de Componentes Principais	433
PCA como Redução de Dimensionalidade	436
PCA para Visualização: Dígitos Feitos à Mão	437
O que Significam os Componentes?	439
Escolhendo o Número de Componentes	440
PCA como Filtragem de Ruído	441
Exemplo: Eigenfaces	443
Resumo	446
CAPÍTULO 46: Aprofundando: Manifold Learning	447
Manifold Learning: “HELLO”	447
Escalonamento Multidimensional	449
MDS como Manifold Learning	451
Mapeamentos Não Lineares: Onde o MDS Falha	453
Variedades Não Lineares: Mapeamento Linear Local	454
Considerações sobre os Manifold Methods	456
Exemplo: Isomap nos Rostos	457
Exemplo: Visualizando a Estrutura nos Dígitos	460
CAPÍTULO 47: Aprofundando: Clusterização k-Means.....	463
Apresentando o k-Means	463
Expectativa-Maximização	465
Exemplos	470
Exemplo 1: k-Means nos Dígitos	470
Exemplo 2: k-Means para Compressão de Cores	473
CAPÍTULO 48: Aprofundando: Modelos de Mistura Gaussiana	477
Motivando as Misturas Gaussianas: Pontos Fracos do k-Means	477
Generalizando E-M: Modelos de Mistura Gaussiana	480
Escolhendo o Tipo de Covariância	484
Modelos de Mistura Gaussiana como Estimativa de Densidade	484
Exemplo: GMMs para Gerar Novos Dados	488
CAPÍTULO 49: Aprofundando: Estimativa de Densidade por Kernel	491
Motivando a Estimativa de Densidade por Kernel: Histogramas	491
Estimativa de Densidade por Kernel na Prática	496
Selecionando a Largura de Banda com a Validação Cruzada	497
Exemplo: Naive Bayes, Mas Nem Tão Ingênuo	498
Anatomia de um Estimador Personalizado	500
Usando Nosso Estimador Personalizado	501
CAPÍTULO 50: Aplicação: Pipeline de Detecção Facial.....	505
Features HOG	505
HOG em Ação: Detector Facial Simples	506
1. Obtenha um Conjunto de Amostras de Treinamento Positivas	507
2. Obtenha um Conjunto de Amostras de Treinamento Negativas	507
3. Combine os Conjuntos e Extraia as Features HOG	508
4. Treine uma Máquina de Vetores de Suporte	509
5. Encontre os Rostos na Nova Imagem	510
Advertências e Melhorias	511
Mais Features do Aprendizado de Máquina	513
Índice	515
Sobre o Autor.....	521
Colofão.....	523

AMOSTRA

O Que É Data Science?

Este é um livro sobre fazer data science com Python, o que nos traz imediatamente a seguinte questão: o que é *data science*? Essa é uma definição surpreendentemente difícil de estabelecer, sobretudo ao considerarmos a abrangência do termo nos dias atuais. Os críticos mais incansáveis já desdenharam do termo como um rótulo supérfluo (afinal, qual ciência não gira em torno de dados?) ou como um mero jargão, que só existe para preencher o currículo e chamar a atenção de recrutadores excessivamente cautelosos.

Na minha opinião, nessas críticas falta uma consideração importante. Data science, apesar da primeira impressão como apenas a “moda da vez”, talvez seja o melhor rótulo à disposição para o conjunto interdisciplinar de habilidades que tem se tornado cada vez mais importante em muitas aplicações, tanto no âmbito do setor quanto no acadêmico. A chave aqui é justamente a *interdisciplinaridade*: para mim, a melhor definição de data science foi muito bem ilustrada pelo Diagrama de Venn do Data Science criado por Drew Conway e publicado pela primeira vez em seu blog, em setembro de 2010 (Figura P-1).

Embora alguns títulos das interseções possam parecer um pouco irônicos, o diagrama captura a essência do que as pessoas pensam ao ouvir falar em “data science”: um assunto interdisciplinar em essência. Data science é composto por três áreas distintas que se sobrepõem: as habilidades de um *estatístico* que sabe como modelar e resumir os conjuntos de dados (que estão se tornando cada vez maiores); as habilidades de um *cientista da computação* que pode desenvolver e usar algoritmos para armazenar, processar e visualizar esses dados de forma eficiente; e o *conhecimento do domínio* — o que geralmente pensamos como o treinamento “clássico” em determinado assunto — necessário tanto para formular as perguntas certas quanto para colocar as respostas em um contexto adequado.

Com isso em mente, peço que você não pense em data science como um novo domínio do conhecimento a ser compreendido, mas um novo conjunto de habilidades que você pode aplicar dentro da sua área de conhecimento atual. Criando um relatório sobre os resultados das últimas eleições, realizando previsões a respeito das devoluções do estoque, otimizando cliques em anúncios online, identificando micro-organismos em fotos de microscópios, buscando novas