

Inhalt

Einleitung	13
-------------------------	-----------

Teil I: Mehr als normales Python: Jupyter

1 Der Einstieg in IPython und Jupyter	21
Die IPython-Shell starten	21
Das Jupyter Notebook starten	22
Hilfe und Dokumentation in IPython	22
Tastaturkürzel in der IPython-Shell	27
2 Erweiterte interaktive Features	31
Magische Befehle in IPython	31
Verlauf der Ein- und Ausgabe	33
IPython und Shell-Befehle	36
3 Debugging und Profiling	41
Fehler und Debugging	41
Profiling und Timing von Code	45
Weitere IPython-Ressourcen	50

Teil II: Einführung in NumPy

4 Die Datentypen in Python	55
Python-Integer sind mehr als nur ganzzahlige Werte	56
Python-Listen sind mehr als nur einfache Listen	57
Arrays feststehenden Typs in Python	59
Arrays anhand von Listen erzeugen	59
Neue Arrays erzeugen	60
NumPys Standarddatentypen	61
5 Grundlagen von NumPy-Arrays	63
Attribute von NumPy-Arrays	63
Indizierung von Arrays: Zugriff auf einzelne Elemente	64
Slicing: Teilmengen eines Arrays auswählen	65
Arrays umformen	68
Arrays verketteten und aufteilen	69

6	Berechnungen mit NumPy-Arrays: universelle Funktionen	71
	Langsame Schleifen	71
	Kurz vorgestellt: UFuncs	72
	NumPys UFuncs im Detail	73
	UFunc-Features für Fortgeschrittene	77
	UFuncs: mehr erfahren	79
7	Aggregationen: Minimum, Maximum und alles dazwischen	81
	Summieren der Werte eines Arrays	81
	Minimum und Maximum	82
	Beispiel: Durchschnittliche Größe der US-Präsidenten	84
8	Berechnungen mit Arrays: Broadcasting	87
	Kurz vorgestellt: Broadcasting	87
	Für das Broadcasting geltende Regeln	89
	Broadcasting in der Praxis	91
9	Vergleiche, Maskierungen und boolesche Logik	95
	Beispiel: Regentage zählen	95
	Vergleichsoperatoren als UFuncs	96
	Boolesche Arrays verwenden	98
	Boolesche Arrays als Maskierungen	100
	Verwendung der Schlüsselwörter »and« bzw. »or« und der Operatoren & bzw.	101
10	Fancy Indexing	103
	Fancy Indexing im Detail	103
	Kombinierte Indizierung	105
	Beispiel: Auswahl zufälliger Punkte	105
	Werte per Fancy Indexing modifizieren	107
	Beispiel: Daten gruppieren	108
11	Arrays sortieren	111
	Schnelle Sortierung in NumPy: np.sort und np.argsort	112
	Nach Zeilen und Spalten sortieren	112
	Teilsortierungen: Partitionierung	113
	Beispiel: k nächste Nachbarn	113
12	Strukturierte Daten: NumPys strukturierte Arrays	117
	Strukturierte Arrays erzeugen	118
	Erweiterte zusammengesetzte Typen	119
	Record-Arrays: strukturierte Arrays mit Pfiff	120
	Weiter mit Pandas	120

Teil III: Datenbearbeitung mit Pandas

13 Kurz vorgestellt: Pandas-Objekte	125
Das Pandas-Series-Objekt	125
Das Pandas-DataFrame-Objekt	128
Das Pandas-Index-Objekt	131
14 Daten indizieren und auswählen	133
Series-Daten auswählen	133
DataFrame-Daten auswählen	136
15 Mit Pandas-Daten arbeiten	141
UFuncs: Indexerhaltung	141
UFuncs: Indexanpassung	142
UFuncs: Operationen mit DataFrame und Series	144
16 Handhabung fehlender Daten	147
Kompromisse beim Umgang mit fehlenden Daten	147
Fehlende Daten in Pandas	148
Pandas nullfähige Datentypen	151
Mit Nullwerten arbeiten	152
17 Hierarchische Indizierung	157
Mehrfach indizierte Series	157
Methoden zum Erzeugen eines MultiIndex	161
Indizierung und Slicing eines MultiIndex	163
Multi-Indizes umordnen	166
18 Datenmengen kombinieren: concat und append	171
Verkettung von NumPy-Arrays	172
Einfache Verkettungen mit pd.concat	172
19 Datenmengen kombinieren: merge und join	177
Relationale Algebra	177
Join-Kategorien	178
Angabe der zu verknüpfenden Spalten	180
Mengenarithmetik bei Joins	183
Konflikte bei Spaltennamen: das Schlüsselwort suffixes	184
Beispiel: Daten von US-Bundesstaaten	185
20 Aggregation und Gruppierung	191
Planetendaten	191
Einfache Aggregationen in Pandas	192
GroupBy: Aufteilen, Anwenden und Kombinieren	194

21	Pivot-Tabellen	203
	Gründe für Pivot-Tabellen	203
	Pivot-Tabellen von Hand erstellen	204
	Die Syntax von Pivot-Tabellen	204
	Beispiel: Geburtenraten	207
22	Vektorisierte String-Operationen	213
	Kurz vorgestellt: String-Operationen in Pandas	213
	Liste der Pandas-String-Methoden	214
	Beispiel: Rezeptdatenbank	218
23	Zeitreihen verwenden	223
	Kalenderdaten und Zeiten in Python	223
	Zeitreihen in Pandas: Indizierung durch Zeitangaben	227
	Datenstrukturen für Zeitreihen in Pandas	227
	Gleichförmige Sequenzen: <code>pd.date_range</code>	228
	Häufigkeiten und Abstände	229
	Resampling, zeitliches Verschieben und geglättete Statistik	231
	Beispiel: Visualisierung von Fahrradzählungen in Seattle	236
24	Leistungstarkes Pandas: <code>eval</code> und <code>query</code>	243
	Der Zweck von <code>query</code> und <code>eval</code> : zusammengesetzte Ausdrücke	243
	Effiziente Operationen mit <code>pandas.eval</code>	244
	<code>DataFrame.eval</code> für spaltenweise Operationen	246
	Die <code>DataFrame.query</code> -Methode	248
	Performance: Wann <code>eval</code> und <code>query</code> verwendet werden sollten	248
	Weitere Ressourcen	249

Teil IV: Visualisierung mit Matplotlib

25	Allgemeine Tipps zu Matplotlib	253
	Matplotlib importieren	253
	Stil einstellen	253
	<code>show</code> oder kein <code>show</code> ? – Anzeige von Diagrammen	253
26	Einfache Liniendiagramme	261
	Anpassen des Diagramms: Linienfarben und -stile	264
	Anpassen des Diagramms: Begrenzungen	266
	Diagramme beschriften	268
	Stolpersteine in Matplotlib	270

27 Einfache Streudiagramme	271
Streudiagramme mit plt.plot erstellen	271
Streudiagramme mit plt.scatter erstellen	273
plot kontra scatter: eine Anmerkung zur Effizienz	276
Visualisierung von Messunsicherheiten	276
28 Dichtediagramme und Konturdiagramme	281
Visualisierung einer dreidimensionalen Funktion	281
Histogramme, Binnings und Dichte	285
Zweidimensionale Histogramme und Binnings	287
29 Anpassen der Legende	291
Legendenelemente festlegen	293
Legenden mit Punktgrößen	295
Mehrere Legenden	296
30 Anpassen von Farbskalen	299
Farbskala anpassen	300
Beispiel: Handgeschriebene Ziffern	304
31 Untergeordnete Diagramme	307
plt.axes: untergeordnete Diagramme von Hand erstellen	307
plt.subplot: untergeordnete Diagramme in einem Raster anordnen	309
plt.subplots: das gesamte Raster gleichzeitig ändern	310
plt.GridSpec: kompliziertere Anordnungen	312
32 Text und Beschriftungen	315
Beispiel: Auswirkungen von Feiertagen auf die Geburtenzahlen in den USA	315
Transformationen und Textposition	317
Pfeile und Beschriftungen	319
33 Achsenmarkierungen anpassen	323
Vorrangige und nachrangige Achsenmarkierungen	323
Markierungen oder Beschriftungen verbergen	325
Anzahl der Achsenmarkierungen verringern oder erhöhen	326
Formatierung der Achsenmarkierungen	328
Zusammenfassung der Formatter- und Locator-Klassen	330
34 Matplotlib anpassen: Konfigurationen und Stylesheets	333
Diagramme von Hand anpassen	333
Voreinstellungen ändern: rcParams	335

35	Dreidimensionale Diagramme in Matplotlib	341
	Dreidimensionale Punkte und Linien	342
	Dreidimensionale Konturdiagramme	343
	Drahtgitter- und Oberflächendiagramme	344
	Triangulation von Oberflächen	346
	Beispiel: Visualisierung eines Möbiusbands	348
36	Visualisierung mit Seaborn	351
	Seaborn-Diagramme	352
	Kategoriale Diagramme	356
	Beispiel: Ergebnisse eines Marathonlaufs	359
	Weiterführende Ressourcen	365
	Weitere Grafikbibliotheken für Python	366

Teil V: Machine Learning

37	Was ist Machine Learning?	369
	Kategorien des Machine Learning	369
	Qualitative Beispiele für Machine-Learning-Anwendungen	370
	Zusammenfassung	379
38	Kurz vorgestellt: Scikit-Learn	381
	Datenrepräsentierung in Scikit-Learn	381
	Die Estimator-API	384
	Anwendung: Handgeschriebene Ziffern untersuchen	392
	Zusammenfassung	397
39	Hyperparameter und Modellvalidierung	399
	Überlegungen zum Thema Modellvalidierung	399
	Auswahl des besten Modells	403
	Lernkurven	410
	Validierung in der Praxis: Rastersuche	414
	Zusammenfassung	415
40	Feature Engineering	417
	Kategoriale Features	417
	Texte als Features	418
	Bilder als Features	420
	Abgeleitete Features	420
	Vervollständigung fehlender Daten	422
	Feature-Pipelines	423

41 Ausführlich: Naive Bayes-Klassifikation	425
Bayes-Klassifikation	425
Gaußsche naive Bayes-Klassifikation	426
Multinomiale naive Bayes-Klassifikation	429
Einsatzgebiete für naive Bayes-Klassifikation	432
42 Ausführlich: Lineare Regression	435
Einfache lineare Regression	435
Regression der Basisfunktion	437
Regularisierung	441
Beispiel: Vorhersage des Fahrradverkehrs	445
43 Ausführlich: Support Vector Machines	451
Gründe für Support Vector Machines	451
Support Vector Machines: Maximierung des Randbereichs	453
Beispiel: Gesichtserkennung	461
Zusammenfassung	465
44 Ausführlich: Entscheidungsbäume und Random Forests	467
Gründe für Random Forests: Entscheidungsbäume	467
Estimator-Ensembles: Random Forests	471
Random-Forest-Regression	473
Beispiel: Random Forest zur Klassifikation handgeschriebener Ziffern	475
Zusammenfassung	477
45 Ausführlich: Hauptkomponentenanalyse	479
Hauptkomponentenanalyse: ein Überblick	479
Hauptkomponentenanalyse als Rauschfilter	487
Beispiel: Eigengesichter	489
Zusammenfassung	492
46 Ausführlich: Manifold Learning	493
Manifold Learning: »HELLO«	494
Multidimensionale Skalierung	495
Nichtlineare Mannigfaltigkeiten: lokal lineare Einbettung	500
Überlegungen zum Thema Manifold-Methoden	502
Beispiel: Isomap und Gesichter	503
Beispiel: Visualisierung der Strukturen in Zifferndaten	507
47 Ausführlich: k-Means-Clustering	511
Kurz vorgestellt: der k-Means-Algorithmus	511
Expectation-Maximization	513
Beispiele	518

48 Ausführlich: Gaußsche Mixture-Modelle	525
Gründe für GMM: Schwächen von k-Means	525
EM-Verallgemeinerung: gaußsche Mixture-Modelle	528
Wahl des Kovarianztyps	532
GMM als Dichteschätzung	532
Beispiel: GMM zum Erzeugen neuer Daten verwenden	536
49 Ausführlich: Kerndichteschätzung	539
Gründe für Kerndichteschätzung: Histogramme	539
Kerndichteschätzung in der Praxis	543
Auswahl der Bandbreite durch Kreuzvalidierung	544
Beispiel: Nicht ganz so naive Bayes-Klassifikation	545
50 Anwendung: Eine Gesichtserkennungspipeline	551
HOG-Features	552
HOG in Aktion: eine einfache Gesichtserkennung	553
Vorbehalte und Verbesserungen	557
Weitere Machine-Learning-Ressourcen	559
Index	561