

Indice

| | |
|---|-------------|
| Presentazione dell'edizione Italiana | XIII |
| Prefazione alla seconda edizione | XV |
| Prefazione alla prima edizione | XIX |
| 1 Introduzione | 1 |
| 1.1 L'ascesa inarrestabile dei dati | 1 |
| 1.2 Cosa significa data science ? | 1 |
| 1.3 Un scenario aziendale plausibile: DataSciencecenter | 2 |
| 1.3.1 Trovare le relazioni fondamentali | 3 |
| 1.3.2 Data scientist che potresti conoscere | 6 |
| 1.3.3 Stipendi ed esperienza | 8 |
| 1.3.4 Account a pagamento | 10 |
| 1.3.5 Argomenti di interesse | 11 |
| 1.3.6 Domani | 12 |
| 2 Un corso accelerato di Python | 13 |
| 2.1 Lo Zen di Python | 13 |
| 2.2 Ottenere Python | 14 |
| 2.3 Ambienti virtuali | 14 |
| 2.4 Formattazione degli spazi | 15 |
| 2.5 Moduli | 16 |
| 2.6 Funzioni | 17 |
| 2.7 Stringhe | 18 |
| 2.8 Eccezioni | 19 |
| 2.9 Liste | 19 |
| 2.10 Tuple | 21 |
| 2.11 Dizionari | 21 |
| 2.11.1 defaultdict | 22 |
| 2.12 Contatori | 23 |
| 2.13 Set (insiemi) | 24 |
| 2.14 Flusso di controllo | 24 |
| 2.15 Valori di verità | 25 |

| VI | Indice |
|---|-----------|
| 2.16 Ordinamento | 26 |
| 2.17 Comprehension di liste | 27 |
| 2.18 Test automatizzati e assert | 27 |
| 2.19 Programmazione a oggetti | 28 |
| 2.20 Iterabili e generatori | 30 |
| 2.21 Numeri casuali | 32 |
| 2.22 Espressioni regolari | 33 |
| 2.23 Programmazione funzionale | 33 |
| 2.24 zip e spaccettamento degli argomenti | 33 |
| 2.25 args e kwargs | 34 |
| 2.26 Annotazioni di tipo | 35 |
| 2.26.1 Come scrivere annotazioni di tipo | 37 |
| 2.27 Benvenuti a DataSciencester! | 39 |
| 2.28 Per approfondire | 39 |
| 3 Visualizzazione dei dati | 41 |
| 3.1 matplotlib | 41 |
| 3.2 Grafici a barre | 43 |
| 3.3 Grafici a linee | 46 |
| 3.4 Grafici a dispersione | 47 |
| 3.5 Per approfondire | 48 |
| 4 Algebra lineare | 51 |
| 4.1 Vettori | 51 |
| 4.2 Matrici | 55 |
| 4.3 Per approfondire | 57 |
| 5 Statistica | 59 |
| 5.1 Descrivere un singolo insieme di dati | 59 |
| 5.1.1 Indici di tendenza centrale | 61 |
| 5.1.2 Dispersione | 62 |
| 5.2 Correlazione | 64 |
| 5.3 Il paradosso di Simpson | 67 |
| 5.4 Qualche altra precisazione sulla correlazione | 68 |
| 5.5 Correlazione e causalità | 68 |
| 5.6 Per approfondire | 69 |
| 6 Probabilità | 71 |
| 6.1 Dipendenza e indipendenza | 71 |
| 6.2 Probabilità condizionata | 72 |
| 6.3 Teorema di Bayes | 73 |
| 6.4 Variabili aleatorie | 74 |
| 6.5 Distribuzioni continue | 75 |
| 6.6 La distribuzione normale | 76 |
| 6.7 Il teorema del limite centrale | 79 |

| Indice | VII |
|---|------------|
| 6.8 Per approfondire | 80 |
| 7 Ipotesi e inferenza | 83 |
| 7.1 Test di ipotesi statistiche | 83 |
| 7.2 Esempio: lancio di una moneta | 83 |
| 7.3 p-values | 86 |
| 7.4 Intervalli di confidenza | 87 |
| 7.5 p-Hacking | 88 |
| 7.6 Esempio: eseguire un test A/B | 89 |
| 7.7 Inferenza bayesiana | 91 |
| 7.8 Per approfondire | 93 |
| 8 Discesa del gradiente | 95 |
| 8.1 L'idea alla base della discesa del gradiente | 95 |
| 8.2 Stimare il gradiente | 96 |
| 8.3 Usare il gradiente | 99 |
| 8.4 Scegliere le dimensioni giuste del passo | 99 |
| 8.5 Usare la discesa del gradiente per apprendere modelli | 100 |
| 8.6 Minibatch e discesa stocastica del gradiente | 101 |
| 8.7 Per approfondire | 103 |
| 9 Come si ottengono i dati | 105 |
| 9.1 stdin e stdout | 105 |
| 9.2 Leggere file | 107 |
| 9.2.1 Gli elementi fondamentali dei file di testo | 107 |
| 9.2.2 File delimitati | 108 |
| 9.3 Web scraping | 110 |
| 9.3.1 HTML e la sua analisi sintattica | 110 |
| 9.3.2 Esempio: tenere sott'occhio il Congresso | 112 |
| 9.4 Usare le API | 114 |
| 9.4.1 JSON e XML | 114 |
| 9.4.2 Usare un'API non autenticata | 115 |
| 9.4.3 Trovare le API | 116 |
| 9.5 Esempio: usare le API di Twitter | 117 |
| 9.5.1 Ottenere le credenziali | 117 |
| 9.6 Per approfondire | 121 |
| 10 Lavorare con i dati | 123 |
| 10.1 Esplorare i dati | 123 |
| 10.1.1 Esplorare dati monodimensionali | 123 |
| 10.1.2 Due dimensioni | 125 |
| 10.1.3 Molte dimensioni | 126 |
| 10.2 Usare NamedTuple | 128 |
| 10.3 Dataclass | 129 |
| 10.4 Pulizia e trasformazione | 130 |

| VIII | Indice |
|---|------------|
| 10.5 Manipolare i dati | 132 |
| 10.6 Rescaling | 135 |
| 10.7 Per inciso: tqdm | 136 |
| 10.8 Riduzione della dimensionalità | 137 |
| 10.9 Per approfondire | 142 |
| 11 Apprendimento automatico (machine learning) | 145 |
| 11.1 Modellazione | 145 |
| 11.2 Che cos'è l'apprendimento automatico? | 146 |
| 11.3 Overfitting e underfitting | 146 |
| 11.4 Correttezza | 149 |
| 11.5 Il compromesso fra bias e varianza | 152 |
| 11.6 Estrazione e selezione di feature | 152 |
| 11.7 Per approfondire | 154 |
| 12 L'algoritmo k-nearest neighbors | 155 |
| 12.1 Il modello | 155 |
| 12.2 Esempio: il dataset Iris | 157 |
| 12.3 La maledizione della dimensionalità | 160 |
| 12.4 Per approfondire | 163 |
| 13 Naïve Bayes | 165 |
| 13.1 Un filtro per lo spam davvero stupido | 165 |
| 13.2 Un filtro per lo spam più sofisticato | 166 |
| 13.3 Implementazione | 167 |
| 13.4 Test del modello | 169 |
| 13.5 Uso del modello | 170 |
| 13.6 Per approfondire | 173 |
| 14 Regressione lineare semplice | 175 |
| 14.1 Il modello | 175 |
| 14.2 Uso della discesa del gradiente | 178 |
| 14.3 Stima della massima verosimiglianza | 179 |
| 14.4 Per approfondire | 180 |
| 15 Regressione multipla | 181 |
| 15.1 Il modello | 181 |
| 15.2 Ulteriori assunzioni del modello dei minimi quadrati | 182 |
| 15.3 Adattamento del modello | 183 |
| 15.4 Interpretazione del modello | 185 |
| 15.5 Bontà dell'adattamento | 185 |
| 15.6 Digressione: il bootstrap | 186 |
| 15.7 Errori standard dei coefficienti di regressione | 187 |
| 15.8 Regolarizzazione | 189 |
| 15.9 Per approfondire | 191 |

| Indice | IX | |
|-----------|---|------------|
| 16 | Regressione logistica | 193 |
| 16.1 | Il problema | 193 |
| 16.2 | La funzione logistica | 195 |
| 16.3 | Applicazione del modello | 197 |
| 16.4 | Bontà di adattamento | 198 |
| 16.5 | Support vector machines | 200 |
| 16.6 | Per approfondire | 202 |
| 17 | Alberi di decisione | 203 |
| 17.1 | Che cos'è un albero di decisione? | 203 |
| 17.2 | Entropia | 205 |
| 17.3 | L'entropia di una partizione | 207 |
| 17.4 | Creazione di un albero decisionale | 208 |
| 17.5 | Mettere insieme il tutto | 210 |
| 17.6 | Random Forest | 213 |
| 17.7 | Per approfondire | 213 |
| 18 | Reti neurali | 215 |
| 18.1 | Perceptron | 215 |
| 18.2 | Reti neurali feed-forward | 217 |
| 18.3 | Retropropagazione | 219 |
| 18.4 | Esempio: Fizz Buzz | 222 |
| 18.5 | Per approfondire | 225 |
| 19 | Deep learning: apprendimento profondo | 227 |
| 19.1 | Il tensore | 227 |
| 19.2 | L'astrazione Layer | 229 |
| 19.3 | Lo strato lineare | 231 |
| 19.4 | Reti neurali come successione di strati | 234 |
| 19.5 | Loss e ottimizzazione | 235 |
| 19.6 | Esempio: una rivisitazione di XOR | 237 |
| 19.7 | Altre funzioni di attivazione | 238 |
| 19.8 | Esempio: una rivisitazione di FizzBuzz | 239 |
| 19.9 | Softmax e cross-entropy | 240 |
| 19.10 | Dropout | 242 |
| 19.11 | Esempio: MNIST | 243 |
| 19.12 | Salvataggio e caricamento di modelli | 247 |
| 19.13 | Per approfondire | 248 |
| 20 | Clustering | 251 |
| 20.1 | L'idea | 251 |
| 20.2 | Il modello | 252 |
| 20.3 | Esempio: meetup | 254 |
| 20.4 | Scelta di k | 255 |
| 20.5 | Esempio: clusterizzazione di colori | 257 |

| X | Indice |
|---|------------|
| 20.6 Clusterizzazione gerarchica ascendente (bottom-up) | 259 |
| 20.7 Per approfondire | 264 |
| 21 Elaborazione del linguaggio naturale | 265 |
| 21.1 Nuvole di parole | 265 |
| 21.2 Modelli linguistici a n-gram | 267 |
| 21.3 Grammatiche | 269 |
| 21.4 Una digressione: il campionamento di Gibbs | 271 |
| 21.5 Topic modeling | 273 |
| 21.6 Vettori di parole | 278 |
| 21.7 Reti neurali ricorrenti | 286 |
| 21.8 Esempio: Uso di una RNN a livello di carattere | 288 |
| 21.9 Per approfondire | 291 |
| 22 Analisi delle reti | 293 |
| 22.1 Betweenness centrality: i nodi importanti | 293 |
| 22.2 Eigenvector centrality | 298 |
| 22.2.1 Moltiplicazione di matrici | 298 |
| 22.2.2 Centralità | 300 |
| 22.3 Grafi orientati e PageRank | 301 |
| 22.4 Per approfondire | 304 |
| 23 Sistemi di recommendation | 305 |
| 23.1 Recommendation da esperti | 305 |
| 23.2 Raccomandare ciò che è popolare | 306 |
| 23.3 Collaborative filtering basato sugli utenti | 307 |
| 23.4 Collaborative filtering basato sugli oggetti | 309 |
| 23.5 Fattorizzazione di matrici | 311 |
| 23.6 Per approfondire | 316 |
| 24 Database e SQL | 317 |
| 24.1 CREATE TABLE e INSERT | 317 |
| 24.2 UPDATE | 320 |
| 24.3 DELETE | 321 |
| 24.4 SELECT | 321 |
| 24.5 GROUP BY | 323 |
| 24.6 ORDER BY | 326 |
| 24.7 JOIN | 326 |
| 24.8 Sottoquery | 329 |
| 24.9 Indici | 329 |
| 24.10 Ottimizzazione delle query | 330 |
| 24.11 NoSQL | 331 |
| 24.12 Per approfondire | 331 |

| | |
|--|------------|
| Indice | XI |
| 25 MapReduce | 333 |
| 25.1 Esempio: conteggio di parole | 333 |
| 25.2 Perché MapReduce ? | 335 |
| 25.3 MapReduce più in generale | 336 |
| 25.4 Esempio: analisi degli aggiornamenti di stato | 337 |
| 25.5 Esempio: moltiplicazione di matrici | 339 |
| 25.6 Divagazione: combinatori | 340 |
| 25.7 Per approfondire | 341 |
| 26 Etica dei dati | 343 |
| 26.1 Che cos'è l'etica dei dati? | 343 |
| 26.2 No, davvero, che cos'è l'etica dei dati? | 343 |
| 26.3 Devo preoccuparmi dell'etica dei dati? | 344 |
| 26.4 Costruire cattivi prodotti con i dati | 345 |
| 26.5 Compromessi fra accuratezza ed equità | 345 |
| 26.6 Collaborazione | 347 |
| 26.7 Interpretabilità | 347 |
| 26.8 Raccomandazioni | 348 |
| 26.9 Dati distorti | 349 |
| 26.10 Protezione dei dati | 349 |
| 26.11 In breve | 350 |
| 26.12 Per approfondire | 350 |
| 27 Andate avanti e praticate la scienza dei dati | 351 |
| 27.1 IPython | 351 |
| 27.2 Matematica | 351 |
| 27.3 Non da zero | 352 |
| 27.3.1 NumPy | 352 |
| 27.3.2 pandas | 352 |
| 27.3.3 scikit-learn | 352 |
| 27.3.4 Visualizzazione | 353 |
| 27.3.5 R | 353 |
| 27.3.6 Apprendimento profondo | 353 |
| 27.4 Trovare dati | 354 |
| 27.5 Fate scienza dei dati | 354 |
| 27.5.1 Hacker News | 354 |
| 27.5.2 Camion dei pompieri | 355 |
| 27.5.3 T-shirt | 355 |
| 27.5.4 Tweet su un globo | 355 |
| 27.5.5 E voi? | 356 |