

# Contents

|                                                        |             |
|--------------------------------------------------------|-------------|
| <b>Preface</b>                                         | <b>xiii</b> |
| <b>Acknowledgements</b>                                | <b>xv</b>   |
| <b>1 Introduction</b>                                  | <b>1</b>    |
| 1.1 Supervised learning . . . . .                      | 1           |
| 1.2 Unsupervised learning . . . . .                    | 7           |
| 1.3 Reinforcement learning . . . . .                   | 11          |
| 1.4 Ethics . . . . .                                   | 12          |
| 1.5 Structure of book . . . . .                        | 15          |
| 1.6 Other books . . . . .                              | 15          |
| 1.7 How to read this book . . . . .                    | 16          |
| <b>2 Supervised learning</b>                           | <b>17</b>   |
| 2.1 Supervised learning overview . . . . .             | 17          |
| 2.2 Linear regression example . . . . .                | 18          |
| 2.3 Summary . . . . .                                  | 22          |
| <b>3 Shallow neural networks</b>                       | <b>25</b>   |
| 3.1 Neural network example . . . . .                   | 25          |
| 3.2 Universal approximation theorem . . . . .          | 29          |
| 3.3 Multivariate inputs and outputs . . . . .          | 30          |
| 3.4 Shallow neural networks: general case . . . . .    | 33          |
| 3.5 Terminology . . . . .                              | 35          |
| 3.6 Summary . . . . .                                  | 36          |
| <b>4 Deep neural networks</b>                          | <b>41</b>   |
| 4.1 Composing neural networks . . . . .                | 41          |
| 4.2 From composing networks to deep networks . . . . . | 43          |
| 4.3 Deep neural networks . . . . .                     | 45          |
| 4.4 Matrix notation . . . . .                          | 48          |
| 4.5 Shallow vs. deep neural networks . . . . .         | 49          |
| 4.6 Summary . . . . .                                  | 52          |

|           |                                                  |            |
|-----------|--------------------------------------------------|------------|
| <b>5</b>  | <b>Loss functions</b>                            | <b>56</b>  |
| 5.1       | Maximum likelihood . . . . .                     | 56         |
| 5.2       | Recipe for constructing loss functions . . . . . | 60         |
| 5.3       | Example 1: univariate regression . . . . .       | 61         |
| 5.4       | Example 2: binary classification . . . . .       | 64         |
| 5.5       | Example 3: multiclass classification . . . . .   | 67         |
| 5.6       | Multiple outputs . . . . .                       | 69         |
| 5.7       | Cross-entropy loss . . . . .                     | 71         |
| 5.8       | Summary . . . . .                                | 72         |
| <b>6</b>  | <b>Fitting models</b>                            | <b>77</b>  |
| 6.1       | Gradient descent . . . . .                       | 77         |
| 6.2       | Stochastic gradient descent . . . . .            | 83         |
| 6.3       | Momentum . . . . .                               | 86         |
| 6.4       | Adam . . . . .                                   | 88         |
| 6.5       | Training algorithm hyperparameters . . . . .     | 91         |
| 6.6       | Summary . . . . .                                | 91         |
| <b>7</b>  | <b>Gradients and initialization</b>              | <b>96</b>  |
| 7.1       | Problem definitions . . . . .                    | 96         |
| 7.2       | Computing derivatives . . . . .                  | 97         |
| 7.3       | Toy example . . . . .                            | 100        |
| 7.4       | Backpropagation algorithm . . . . .              | 103        |
| 7.5       | Parameter initialization . . . . .               | 107        |
| 7.6       | Example training code . . . . .                  | 111        |
| 7.7       | Summary . . . . .                                | 111        |
| <b>8</b>  | <b>Measuring performance</b>                     | <b>118</b> |
| 8.1       | Training a simple model . . . . .                | 118        |
| 8.2       | Sources of error . . . . .                       | 120        |
| 8.3       | Reducing error . . . . .                         | 124        |
| 8.4       | Double descent . . . . .                         | 127        |
| 8.5       | Choosing hyperparameters . . . . .               | 132        |
| 8.6       | Summary . . . . .                                | 133        |
| <b>9</b>  | <b>Regularization</b>                            | <b>138</b> |
| 9.1       | Explicit regularization . . . . .                | 138        |
| 9.2       | Implicit regularization . . . . .                | 141        |
| 9.3       | Heuristics to improve performance . . . . .      | 144        |
| 9.4       | Summary . . . . .                                | 154        |
| <b>10</b> | <b>Convolutional networks</b>                    | <b>161</b> |
| 10.1      | Invariance and equivariance . . . . .            | 161        |
| 10.2      | Convolutional networks for 1D inputs . . . . .   | 163        |
| 10.3      | Convolutional networks for 2D inputs . . . . .   | 170        |

|           |                                                                  |            |
|-----------|------------------------------------------------------------------|------------|
| 10.4      | Downsampling and upsampling . . . . .                            | 171        |
| 10.5      | Applications . . . . .                                           | 174        |
| 10.6      | Summary . . . . .                                                | 179        |
| <b>11</b> | <b>Residual networks</b>                                         | <b>186</b> |
| 11.1      | Sequential processing . . . . .                                  | 186        |
| 11.2      | Residual connections and residual blocks . . . . .               | 189        |
| 11.3      | Exploding gradients in residual networks . . . . .               | 192        |
| 11.4      | Batch normalization . . . . .                                    | 192        |
| 11.5      | Common residual architectures . . . . .                          | 195        |
| 11.6      | Why do nets with residual connections perform so well? . . . . . | 199        |
| 11.7      | Summary . . . . .                                                | 199        |
| <b>12</b> | <b>Transformers</b>                                              | <b>207</b> |
| 12.1      | Processing text data . . . . .                                   | 207        |
| 12.2      | Dot-product self-attention . . . . .                             | 208        |
| 12.3      | Extensions to dot-product self-attention . . . . .               | 213        |
| 12.4      | Transformer layers . . . . .                                     | 215        |
| 12.5      | Transformers for natural language processing . . . . .           | 216        |
| 12.6      | Encoder model example: BERT . . . . .                            | 219        |
| 12.7      | Decoder model example: GPT3 . . . . .                            | 222        |
| 12.8      | Encoder-decoder model example: machine translation . . . . .     | 226        |
| 12.9      | Transformers for long sequences . . . . .                        | 227        |
| 12.10     | Transformers for images . . . . .                                | 228        |
| 12.11     | Summary . . . . .                                                | 232        |
| <b>13</b> | <b>Graph neural networks</b>                                     | <b>240</b> |
| 13.1      | What is a graph? . . . . .                                       | 240        |
| 13.2      | Graph representation . . . . .                                   | 243        |
| 13.3      | Graph neural networks, tasks, and loss functions . . . . .       | 245        |
| 13.4      | Graph convolutional networks . . . . .                           | 248        |
| 13.5      | Example: graph classification . . . . .                          | 251        |
| 13.6      | Inductive vs. transductive models . . . . .                      | 252        |
| 13.7      | Example: node classification . . . . .                           | 253        |
| 13.8      | Layers for graph convolutional networks . . . . .                | 256        |
| 13.9      | Edge graphs . . . . .                                            | 260        |
| 13.10     | Summary . . . . .                                                | 261        |
| <b>14</b> | <b>Unsupervised learning</b>                                     | <b>268</b> |
| 14.1      | Taxonomy of unsupervised learning models . . . . .               | 268        |
| 14.2      | What makes a good generative model? . . . . .                    | 269        |
| 14.3      | Quantifying performance . . . . .                                | 271        |
| 14.4      | Summary . . . . .                                                | 273        |
| <b>15</b> | <b>Generative Adversarial Networks</b>                           | <b>275</b> |

|           |                                                                         |            |
|-----------|-------------------------------------------------------------------------|------------|
| 15.1      | Discrimination as a signal . . . . .                                    | 275        |
| 15.2      | Improving stability . . . . .                                           | 280        |
| 15.3      | Progressive growing, minibatch discrimination, and truncation . . . . . | 286        |
| 15.4      | Conditional generation . . . . .                                        | 288        |
| 15.5      | Image translation . . . . .                                             | 290        |
| 15.6      | StyleGAN . . . . .                                                      | 295        |
| 15.7      | Summary . . . . .                                                       | 297        |
| <b>16</b> | <b>Normalizing flows</b>                                                | <b>303</b> |
| 16.1      | 1D example . . . . .                                                    | 303        |
| 16.2      | General case . . . . .                                                  | 306        |
| 16.3      | Invertible network layers . . . . .                                     | 308        |
| 16.4      | Multi-scale flows . . . . .                                             | 316        |
| 16.5      | Applications . . . . .                                                  | 317        |
| 16.6      | Summary . . . . .                                                       | 320        |
| <b>17</b> | <b>Variational autoencoders</b>                                         | <b>326</b> |
| 17.1      | Latent variable models . . . . .                                        | 326        |
| 17.2      | Nonlinear latent variable model . . . . .                               | 327        |
| 17.3      | Training . . . . .                                                      | 330        |
| 17.4      | ELBO properties . . . . .                                               | 333        |
| 17.5      | Variational approximation . . . . .                                     | 335        |
| 17.6      | The variational autoencoder . . . . .                                   | 335        |
| 17.7      | The reparameterization trick . . . . .                                  | 338        |
| 17.8      | Applications . . . . .                                                  | 339        |
| 17.9      | Summary . . . . .                                                       | 342        |
| <b>18</b> | <b>Diffusion models</b>                                                 | <b>348</b> |
| 18.1      | Overview . . . . .                                                      | 348        |
| 18.2      | Encoder (forward process) . . . . .                                     | 349        |
| 18.3      | Decoder model (reverse process) . . . . .                               | 355        |
| 18.4      | Training . . . . .                                                      | 356        |
| 18.5      | Reparameterization of loss function . . . . .                           | 360        |
| 18.6      | Implementation . . . . .                                                | 362        |
| 18.7      | Summary . . . . .                                                       | 367        |
| <b>19</b> | <b>Reinforcement learning</b>                                           | <b>373</b> |
| 19.1      | Markov decision processes, returns, and policies . . . . .              | 373        |
| 19.2      | Expected return . . . . .                                               | 377        |
| 19.3      | Tabular reinforcement learning . . . . .                                | 381        |
| 19.4      | Fitted Q-learning . . . . .                                             | 385        |
| 19.5      | Policy gradient methods . . . . .                                       | 388        |
| 19.6      | Actor-critic methods . . . . .                                          | 393        |
| 19.7      | Offline reinforcement learning . . . . .                                | 394        |
| 19.8      | Summary . . . . .                                                       | 395        |

|                                                                       |            |
|-----------------------------------------------------------------------|------------|
| <b>20 Why does deep learning work?</b>                                | <b>401</b> |
| 20.1 The case against deep learning . . . . .                         | 401        |
| 20.2 Factors that influence fitting performance . . . . .             | 402        |
| 20.3 Properties of loss functions . . . . .                           | 406        |
| 20.4 Factors that determine generalization . . . . .                  | 410        |
| 20.5 Do we need so many parameters? . . . . .                         | 414        |
| 20.6 Do networks have to be deep? . . . . .                           | 417        |
| 20.7 Summary . . . . .                                                | 418        |
| <b>21 Deep learning and ethics</b>                                    | <b>420</b> |
| 21.1 Value alignment . . . . .                                        | 420        |
| 21.2 Intentional misuse . . . . .                                     | 426        |
| 21.3 Other social, ethical, and professional issues . . . . .         | 428        |
| 21.4 Case study . . . . .                                             | 430        |
| 21.5 The value-free ideal of science . . . . .                        | 431        |
| 21.6 Responsible AI research as a collective action problem . . . . . | 432        |
| 21.7 Ways forward . . . . .                                           | 433        |
| 21.8 Summary . . . . .                                                | 434        |
| <b>A Notation</b>                                                     | <b>436</b> |
| <b>B Mathematics</b>                                                  | <b>439</b> |
| B.1 Functions . . . . .                                               | 439        |
| B.2 Binomial coefficients . . . . .                                   | 441        |
| B.3 Vector, matrices, and tensors . . . . .                           | 442        |
| B.4 Special types of matrix . . . . .                                 | 445        |
| B.5 Matrix calculus . . . . .                                         | 447        |
| <b>C Probability</b>                                                  | <b>448</b> |
| C.1 Random variables and probability distributions . . . . .          | 448        |
| C.2 Expectation . . . . .                                             | 452        |
| C.3 Normal probability distribution . . . . .                         | 456        |
| C.4 Sampling . . . . .                                                | 459        |
| C.5 Distances between probability distributions . . . . .             | 459        |
| <b>Bibliography</b>                                                   | <b>462</b> |
| <b>Index</b>                                                          | <b>513</b> |